Modeling Textual Entailment with Role-Semantic Information

Aljoscha Burchardt

Saarbrücken, 2008

Dissertation zur Erlangung des akademischen Grades eines
Doktors der Philosophie der Philosophischen Fakultäten
der Universität des Saarlandes

"In creating a system which accepts text, answers questions, or enters into a dialogue, we have not created a theory of semantics, we have created another class of objects for which such a theory is needed."

Terry Winograd (Winograd, 1978)

Die Dekanin:	Prof. Dr. Susanne Kleinert
Berichterstatter/innen:	Prof. Dr. Manfred Pinkal
	Prof. Dr. Anette Frank
Tag der letzten Prüfungsleistung:	

Abstract

In this thesis, we present a novel approach for modeling textual entailment using lexical-semantic information on the level of predicate-argument structure. To this end, we adopt information provided by the Berkeley FrameNet repository and embed it into an implemented end-to-end system. The two main goals of this thesis are the following: (i) to provide an analysis of the potential contribution of frame semantic information to the recognition textual entailment and (ii) to present a robust system architecture that can serve as basis for future experiments, research, and improvement.

Our work was carried out in the context of the textual entailment initiative, which since 2005 has set the stage for the broad investigation of inference in natural-language processing tasks, including empirical evaluation of its coverage and reliability. In short, textual entailment describes inferential relations between (entailing) texts and (entailed) hypotheses as interpreted by typical language users. This pre-theoretic notion captures a natural range of inferences as compared to logical entailment, which has traditionally been used within theoretical approaches to natural language semantics.

Various methods for modeling textual entailment have been proposed in the literature, ranging from shallow techniques like lexical overlap to shallow syntactic parsing and the exploitation of WordNet relations. Recently, there has been a move towards more structured meaning representations. In particular, the level of predicate-argument structure has gained much attention, which seems to be a natural and straightforward choice. Predicate-argument structure allows annotating sentences or texts with nuclear meaning representations ("who did what to whom"), which are of obvious relevance for this task. For example, it can account for paraphrases like "Ghosts scare John" vs. "John is scared by ghosts".

In this thesis, we present an approach to textual entailment that is centered around the analysis of predicate-argument structure. It combines LFG grammatical analysis, predicate-argument structure in the FrameNet paradigm, and taxonomic information from WordNet into tripartite graph structures. By way of a declarative graph matching algorithm, the "structural and semantic" similarity of hypotheses and texts is computed and the result is represented as feature vectors. A supervised machine learning architecture trained on entailment corpora is used to check textual entailment for new text/hypothesis pairs. The approach is implemented in the SALSA RTE system, which successfully participated in the second and third RTE challenge.

While system performance is on a par with that of comparable systems, the intuitively expected strong positive effect of using FrameNet information has not yet been confirmed. In order to evaluate different system components and to assess the potential contribution of FrameNet information for checking textual entailment, we conducted a number of experiments. For example, with the help of a gold-standard corpus, we

experimentally analyzed different factors that can limit the applicability of frame semantics in checking textual entailment, ranging from issues related to resource coverage to knowledge modeling problems.

Ausführliche Zusammenfassung

Die vorliegende Arbeit beschäftigt sich mit einem neuen Ansatz zur Modellierung natürlichsprachlicher Folgerungsbeziehungen ("Textual Entailment") mithilfe lexikalisch-semantischer Information auf Ebene der Prädikat-Argument-Struktur. Zu diesem Zwecke verwenden wir Information aus der Berkeley FrameNet-Datenbank und betten diese in ein von uns implementiertes "end-to-end" System ein. Die zwei Hauptbeiträge dieser Dissertation sind (i) die Untersuchung des potentiellen Beitrages von *Framesemantik* zur Modellierung von Textual Entailment und (ii) die Präsentation einer robusten Systemarchitektur als Basis für zukünftige Untersuchungen, Experimente und Weiterentwicklungen.

Eine Motivation dieser Arbeit die Erkenntnis, dass derzeitige, "flache" Verfahren des automatischen Informationszugriffes, wie stichwortbasierte Suche im WWW, den menschlichen Benutzer nicht optimal beim Zugriff auf die vorhandene Flut von natürlichsprachlicher Information unterstützen. Wir zeigen, dass Textual Entailment ein geeignete Grundlage ist, in einer Vielzahl von Anwendungen intelligentere, semantische Verfahren zu implementieren, die typische Fehler flacher Verfahren vermeiden.

Textual Entailment ist ein unlängst eingeführtes, prä-theoretisches Konzept, das – kurz gesagt – gerade die Arten von Folgerungsbeziehungen zwischen Sätzen (dem "entailenden" *Text* und der "entailten" *Hypothese*) beschreibt, die typische Sprecher für gewöhnlich herstellen. Es versteht sich als Erweiterung des logisches Entailmentbegriffes, welcher über viele Jahre der vorherrschende Folgerungsbegriff in Ansätzen zur Formalisierung natürlichsprachlicher Bedeutung war, dabei jedoch nur einen kleinen, eher uninteressanten Teil dessen, was Menschen zu folgern in der Lage sind, abbildet. Die seit 2005 jährlich stattfindenden "Recognizing Textual Entailment" (RTE)-Wettbewerbe bieten die Möglichkeit, formale Modelle von Textual Entailment auf Grundlage von Korpusdaten zu entwickeln und zu evaluieren.

In der Literatur werden verschiedenste Verfahren zur Modellierung von Textual Entailment vorgeschlagen, die zum Beispiel von Maßen der Wort-Übereinstimmung, syntaktischem Parsing und WordNet Relationen Gebrauch machen. Hierbei geht der Trend in letzter Zeit hin zu stärker strukturierter, semantischer Information und weg von irrelevanten Oberflächenmerkmalen. Die Ebene der Prädikat-Argument-Struktur hat dabei eine gewisse Aufmerksamkeit erlangt, da sie eine natürliche und folgerichtige Wahl zu sein scheint. Prädikat-Argument-Strukturen beschränken sich auf die Modellierung von Kernbedeutungen ("wer tut wem was") und können zum Beispiel Paraphrasen wie "Peter hat Angst vor Gespenstern" und "Gespenster ängstigen Peter" erklären. Verschiedene Studien haben gezeigt, dass Variationen auf dieser Ebene einen nennswerten Teil der Inferenzen in den vorhandenen RTE-Korpora ausmachen.

Ein zentraler Beitrag dieser Arbeit ist das von uns entworfene und implementierte

SALSA RTE-System, welches erfolgreich am zweiten und dritten RTE-Wettbewerb teilgenommen hat. Es ist das erste System, welches als semantische Hauptinformation Beschreibungen der Prädikat-Argument-Struktur von Text und Hypothese verwendet. Im System werden grammatische Informationen einer LFG-Grammatik, Prädikat-Argument-Struktur im framesemantischen Paradigma und taxonomische Information aus WordNet in dreigeteilte Graphstrukturen zusammengeführt. Wir folgen dabei der LFG-Projektions-Architektur, indem die Information der einzelnen Analysebenen getrennt repräsentiert und durch Projektionen verlinkt wird. Nach verschiedenen Schritten, bei denen die Bedeuntungsinformation weiter verdichtet und normalisiert wird, werden die Graphen von Hypothese und Text unter verschiedenen Gesichtspunten miteinander verglichen und das Ergebnis als Merkmalsvektoren repräsentiert. Für den Vergleich der Analysen von Hypothese und Text haben wir ein Graph-Matching-Verfahren konzipiert und implementiert, welches die "strukturelle und semantische" Überlappung beider auf eine deklarative Art und Weise beschreibt. Die Merkmalsvektoren dienen einer maschinellen Lernarchitektur als Eingabe, die auf RTE-Korpora trainiert wird, um Textual Entailment auf unbekannten Korpora zu bestimmen. Die gesamte Systemarchitektur ist offen angelegt und für Erweiterungen vorbereitet. Wir illustrieren zum Beispiel, wie Hintergrundwissen aus der SUMO-Ontologie integriert werden kann und zeigen, wie Negation und Modalität approximativ behandelt werden können.

Forschung im Bereich Frame-Semantik hat sich in den letzten Jahren vorwiegend mit der automatischen Annotation von Text beschäftigt. Ein Resultat dieser Forschung ist Shalmaneser, der semantische Parser, den wir in unserer Arbeit verwenden. In Anwendungs-Szenarien ist Frame-Semantik nur in geringem Umfang eingesetzt worden. In dieser Arbeit zeigen wir zunächst anhand einer Handannotation eines RTE-Korpus', dass die Abdeckung von FrameNet auf den RTE-Korpora gut ist. 92% der relevanten Prädikate werden durch vorhandene Frames beschrieben. Um mit automatischen System bestmögliche Abdeckung zu erreichen, haben wir unter Ausnutzung der sehr guten Abdeckung von WordNet das Detour System entwickelt, welches Lücken im FrameNet-Lexikon ausgleicht und in Kombination mit Shalmaneser im SALSA RTE-System Einsatz findet.

In einer ausführlichen Evaluation untersuchen wir die Performanz des SALSA RTE-Systems, mit besonderem Hinblick auf die Frage, was der derzeitige und potentielle Beitrag von Frame-Semantik zum Erkennen von Textual Entailment ist. Ein Resultat ist, dass die automatische frame-semantische Analyse einen manuell erzeugten Gold Standard noch nicht gut annähert. Darüberhinaus zeigen wir, dass selbst auf der manuellen Frame-Annotation von Text und Hypothese der einfache Vergleich von übereinstimmenden Frames, Rollen und Rollenfüllern keinen substantiellen Beitrag zur Entailmententscheidung liefert. Als entscheidendes Manko zeigt sich hier das Fehlen von Ähnlichkeitsmaßen für Frames und Frame-Annotationen. Für die Erzeugung solcher Maße ist die FrameNet-Hierarchie, welche verschiedene Relationen zwischen Frames kodiert, unerlässlich. Bis dato gibt es jedoch keine automatischen Methoden, die diese geeignet interpretieren. Im Ausblick skizzieren wir Möglichkeiten hierfür, wie auch für den erweiterten Einsatz der Hierarchie zur Analyse längerer Texte, was für RTE von zunehmend größerem Interesse ist.

Acknowledgments

I first want to thank my "Doktorvater" Manfred Pinkal, who put trust in me over many years. Working in his outstanding department[1], I was lucky to learn many interesting things that can hardly be documented in a scientific work like this.

Many thanks to Anette Frank, with whom I developed main contributions of thesis.

I also thank all other people that supported me in the preparation of this thesis in one way or another: Cristoph Clodo, Alexander Koller, Ursula Kröner, Sebastian Pado, Marco Pennacchiotti, Nils Reiter, Caroline Sporleder, Stefan Thater, Stephan Walter, and Michael Wirth.

Thanks to the people that helped me survive my Saarbrücken exile: Walt, Albert[2], Lutz for all@Wellness, my nice neighbors Magda and Michal.

Thanks to my parents and family and also Anne's family for all their love and support.

My deepest thanks to Anne, who never complained seeing me only from the back on Sundays, sitting at my desk.

[1] Part of this work has be funded by the German Science Fundation DFG (Titel PI 154/9-3, "SALSA II: Semantisches Lexikon").

[2] `www.hifigalerie.de`

Contents

List of Figures

List of Tables

Part I.

Introduction and Background

1. Introduction

With the rapid growth of intranets and the World Wide Web, more and more text documents are available (only) in electronic form. Information access has become a central challenge in this context. Due to the vast number of documents, computational models of information access using natural language processing techniques are indispensable. In order to satisfy the needs of human users, intelligent, meaning-based processing deserves special attention – a view that is supported by the vision of a "Semantic Web". Yet, current technologies such as search engines or prototypical question answering systems typically consider semantic information only in limited ways.

In theoretical and computational semantics, truth-conditional logic formalisms have been the standard framework for modeling natural language meaning over the last decades. Reasoning for natural language processing has been conceived as being more or less equivalent to logical reasoning (e.g. Blackburn, Bos, Kohlhase, and de Nivelle, 2001; Monz and de Rijke, 1999). For example, inferential relations between natural language sentences have been modeled in terms of logical entailment. Still, so far we have hardly seen any robust and broad-coverage semantic analysis system that provides "deep" semantic representations within any major computational semantics framework.

At the same time, large-scale lexical semantic resources such as WordNet (Fellbaum, 1998) have been developed and put to use for *approximate* semantic modeling in prototype applications. However, existing approaches tend to develop semantic frameworks that fit their very particular needs (e.g. a surface-near representation in Harabagiu, Moldovan, Pasca, Mihalcea, Surdeanu, Bunescu, Girju, Rus, and Morarescu (2000) or a statistical measure in Mohit and Narayanan (2003)). This makes it difficult to evaluate and compare the performance of different approaches and systems, in particular if they implement different tasks.

More recently, the seminal paper Monz and de Rijke (2001) and the *Recognizing Textual Entailment* (RTE) initiative (Dagan, Glickman, and Magnini, 2006) have set the stage for the broad investigation of inference in natural-language processing. The new framework for textual inference lays emphasis on realistic scenarios and empirical evaluation of coverage and reliability. The notion *textual* entailment is used to stress its foundation in natural language, in contrast to *logical* entailment. Textual entailment captures the intuitive concept of entailment typical natural language users have. RTE is a promising candidate for a semantic framework that is useful for a wide range of natural language processing tasks. Therefore, the development of computational models has become an active field of research.

In this thesis, we argue that *predicate-argument structure*, describing natural language semantics on a medium level of complexity, is well-suited for modeling textual entailment on real text. We present the SALSA RTE system, an end-to-end system for

modeling textual entailment, based on frame-semantic information provided by Berkeley FrameNet.

This chapter is structured as follows. We first shortly review current information access tasks and point to problems related to missing semantic analysis (Section 1.1). Section 1.2 introduces textual entailment as a realistic semantic inference scenario that can be used in many application. Section 1.3 is concerned with the question how to model textual entailment, demonstrating the appeal of predicate-argument structure. Finally, in Section 1.4, we summarize the contributions of this thesis and present its structure in Section 1.5.

1.1. Background: Issues of Automatic Information Access

By automatic information access, we refer to a variety of different tasks having in common that a human interacts with a machine to access information from a collection of natural language documents stored in electronic form. Scenarios include key-word based search, where the user wants to access, e.g., all documents containing the word *dianetics* or information extraction, where the user wants, e.g., a table filled with information about earthquakes such as time, place, and magnitude, gathered from all relevant documents.

It is in the nature of the task that the user is interested in the *meaning* conveyed by the documents. However, as automatic access of natural language meaning is a difficult task, current information access systems often only roughly *approximate* it. The most prominent systems on the market are search engines like Google or Yahoo implementing document retrieval, a special case of information retrieval (IR). Within these systems, minimal linguistic processing is performed, mostly to identify content words like nouns and verbs ("terms") within indexed documents. The frequencies of the terms occurring in documents are then taken as clues to their meaning. In a standard IR set-up (Baeza-Yates and Ribeiro-Neto, 2000), terms within queries are compared to terms in the indexed documents and relevance of the documents is assessed on the basis of statistical frequency measures. In fact, this method has established itself as "best practice" and is used within a number different applications.

These "shallow" approaches work well if the documents to be queried exhibit a certain level of redundancy. Redundancy here means that a variety of potential user query terms occurs either within one document or across equivalent documents. If this is the case, different user queries can be matched with a suitable document. The Wikipedia entry about the former German chancellor Gerhard Schröder can be accessed with all queries given below as the respective search terms all occur on the page.

(1.1) Ex-Bundeskanzler Schröder

(1.2) Bundeskanzler Schröder

(1.3) Kanzler Deutschland Schröder

However, such approximative approaches do not scale – as soon as the documents to be found are rare (higher recall is needed), or there are documents with different meaning that contain the same terms (higher precision is needed), these approaches perform badly or fail. For example, if one searches the World Wide Web for information about the business deal where BMW took over Rover, using the query (1.4), more than half of the ten best hits returned by the state-of-the-art Yahoo search engine report on another, more recent business deal where Rover was sold again. While (1.5) is a snippet from a relevant document, (1.6) is from an unwanted document.

(1.4) BMW bought Rover.

(1.5) German motor company **BMW** had **bought Rover** in 1994 and the series follows the sometimes fraught relationship between the two.

(1.6) Ford **bought** Land **Rover** from **BMW** in 2000, and have held onto the rights to buy the Rover name, a right they are now exercising.

A related problem occurs in the answer retrieval component of question answering systems (cf. Voorhees, 1999). Such systems typically use standard IR techniques to retrieve answer candidates from a given document collection (e.g., Harabagiu et al., 2000). If a user asks a question like (1.7), such a system uses a query pattern like (1.8) in order to derive relevant documents. With this pattern, both source documents (1.9) and (1.11) can be retrieved. Yet, only (1.11) contains the right information. Depending on which of the source documents is chosen, either the correct answer (1.12) or the false answer (1.10) is derived by the system.

(1.7) How many inhabitants has Slovenia?

(1.8) Slovenia * `?NUMBER` * Inhabitants

(1.9) The capital of **Slovenia** is Ljubljana, with **270,000 inhabitants**.

(1.10) Slovenia has 270,000 inhabitants.

(1.11) **Slovenia**, with its **1.95 million inhabitants**, will be the fifth smallest state to join the EU, reports Le Monde (France).

(1.12) Slovenia has 1.95 million inhabitants.

1.2. Textual Entailment – A Framework for Modeling Natural Language Inference

Above, we have indicated that current "shallow" information access techniques do not meet the requirement to be "intelligent", suffering from the limited expressivity of term-based IR. For example, systems systematically return wrong results in cases, where query

terms merely co-occur in documents as they cannot determine the *relations* between the underlying concepts.

Checking the system results and the respective original sentences from the document collection for *semantic* plausibility would make it possible to sort out false hits. Intuitively, it must be ensured that the system results somehow follow from the evidence contained in the retrieved documents. A semantic notion which models such a "follows from" relation is that of *entailment*. Applied to the given examples, (1.9) does not entail (1.10) while (1.11) entails (1.12). Likewise if we consider sentence (1.4), we see that it is entailed by (1.5) but not by (1.6). Checking entailment can be used to distinguish between plausible and implausible system results not only in these examples. Crouch, Condoravdi, de Paiva, Stolle, and Bobrow (2003) argue that being able to model entailment relations between sentences is a minimal, necessary criterion of (automatic) language understanding.

The question is, how such an entailment relation can be modeled. Traditionally, logic has been used as a formal model of natural language semantics. Therefore, entailment between natural language sentences has been understood as being largely equivalent to logical entailment (see, e.g. Blackburn et al., 2001). However, this logic-based approach has never been implemented in realistic scenarios. First of all, designing a logic-based processing architectures meets with a number of severe difficulties and secondly, the strict notion of logical entailment has shown to cover only a part of what humans judge as inferable (we will elaborate this in Chapter 2).

Starting from the observation that the logical approach has led to a narrow view on inference in natural language semantics, Monz and de Rijke (2001) in their seminal paper propose (i) to develop heuristic, approximative computational semantic methods on the basis of available resources and (ii) to evaluate these methods in an empirical, corpus-based fashion. The authors exemplify both ideas with a simple heuristic method for entailment checking using basically a statistical measure known from IR (*idf – inverse document frequency*). The proposal of Monz and de Rijke (2001) has more recently been take up by the PASCAL recognizing textual entailment (RTE) initiative (e.g. Dagan et al., 2006), which introduced the notion of *textual entailment* and provided an informal definition hereof.

The general idea behind the pre-theoretic notion of textual entailment is to model "prototypical entailment" as performed by humans. The RTE initiative also launched a series of challenges, where the task is to compute entailment judgments for sentence pairs provided by corpora. For example, (1.13)-(1.14) is a so-called *text-hypothesis* pair from such a corpus.[1]

(1.13) Everest summiter David Hiddleston has passed away in an avalanche of Mt. Tasman.

(1.14) A person died in an avalanche. (TRUE)

[1] Throughout this thesis, examples marked *TRUE* or *FALSE* are from the corpora of the Recognizing Textual Entailment Challenge (Dagan et al., 2006; Bar-Haim, Dagan, Dolan, Ferro, Giampiccolo, Magnini, and Szpektor, 2006; Giampiccolo, Magnini, Dagan, and Dolan, 2007a) if not indicated otherwise.

Textual entailment holds between an entailing text (T) and an entailed hypothesis (H) if the meaning of the hypothesis follows from the meaning of the text as interpreted by a typical language user. In (1.13)-(1.14), textual entailment holds, as indicated by the label *TRUE*. In this case, the hypothesis also follows from the text in a logical sense. However, in many cases, where an inference is plausible rather than logically justified, textual entailment deviates from logical entailment. Consider example (1.15)-(1.16).

(1.15) The anti-terrorist court found two men guilty of murdering Shapour Bakhtiar and his secretary Sorush Katibeh, who were found with their throats cut in August 1991.

(1.16) Shapour Bakhtiar died in 1991. (TRUE)

Most humans would probably accept the "default" inference that (1.16) follows from (1.15). This inference is not strict in a logical sense. It is defeasible as it may also be the case that Bakhtiar died before 1991 and that his dead body was found only in 1991. Therefore, logical entailment does not hold, while textual entailment does. In Chapter 2, we will discuss the notion of textual entailment at length. Overall, if applied to real data, the textual entailment scenario is more natural and realistic than the traditional, logical approach. At the same time, the informal definition of textual entailment is a challenge for the development of computational models, as can be seen in the comparably low accuracy achieved by the majority of existing systems.

1.3. Modeling Textual Entailment with Predicate-Argument Structure

Various methods for modeling textual entailment have been proposed in the literature, ranging from shallow techniques like measuring lexical overlap, to syntactic parsing and the exploitation of WordNet relations (see Bar-Haim et al., 2006, for an overview). Recently, there has been a move towards more structured meaning representations, abstracting away from semantically irrelevant surface. In particular, the level of predicate-argument structure, which seems to be a natural and straightforward choice, has gained much attention (e.g. Bobrow, Crouch, King, Condoravdi, Karttunen, Nairn, de Paiva, and Zaenen, 2007; Delmonte, Bristot, Piccolino Boniforti, and Tonelli, 2007). Predicate-argument structure allows annotating a sentence or text with nuclear meaning representations ("who did what to whom"). It is not concerned with problems of "deep" semantic analysis such as modality, negation, or scope ambiguity.

The notion of predicate-argument structure refers to *semantic* arguments on the level of meaning (representations). Semantic approaches dealing with predicate-argument structure are concerned with the question of (i) how to describe the semantic valency of predicates and (ii) how to map syntactic arguments onto semantic arguments. A typical notion used in this context are *semantic (thematic) roles*. The most well-known roles are probably the traditional, generic roles like `agent` or `patient`, which are annotated in the example sentences below.

(1.17) [BMW]$_{\text{AGENT}}$ has bought [Rover]$_{\text{PATIENT}}$.

(1.18) [BMW]$_{\text{PATIENT}}$ was bought by [Rover]$_{\text{AGENT}}$.

While the meaning of both sentences is quite dissimilar, the surface strings are almost identical. Also, the syntactic analyses are relatively similar – both have in common that *BMW* is the subject, differing in the syntactic position assigned to *Rover*. A comparison of the semantic roles reveals that the meaning of these sentences is in fact different. Agent and patient roles are assigned to different entities across the sentences. The underlying reason is that (1.18) is a passive sentence, where syntactic and thematic roles are assigned "cross-over", e.g., the syntactic subject becomes the semantic patient.

This simple example was made up only for illustration. Current approaches to predicate-argument structure typically use more elaborate types of markup. The most prominent resources are PropBank (Palmer, Gildea, and Kingsbury, 2005) and FrameNet (Baker, Fillmore, and Lowe, 1998). PropBank information allows to map alternative syntactic representations of the same lexical expression to one and the same predicate-argument structure. It has predominantly been used for for the study of role labeling methods (e.g. Gildea and Hockenmaier, 2003; Pradhan, Ward, Hacioglu, Martin, and Jurafsky, 2005). As thematic roles are lemma-specific in PropBank, it is not clear to what extent PropBank can help to recognize entailment in pairs such as (1.13)-(1.14). FrameNet, on the other hand, abstracts over individual lemmas and groups words evoking the same situation type into so-called *frames*. Lemmas belonging to the same frame share thematic roles, which supports the recognition of entailment for sentence pairs containing verbs which belong to the same frame.

We propose to apply frame semantic information to the task of recognizing textual entailment. Several studies (e.g.,Litkowski (2006); Bar-Haim, Szpektor, and Glickman (2005); Clark, Harrison, Thompson, Murray, Hobbs, and Fellbaum (2007)) indicate that the level of granularity offered by FrameNet is relevant for modeling many phenomena which occur in the current textual entailment corpora. For example, Bar-Haim et al. (2005) show that 31% of the RTE-2 positive dataset involves paraphrase at the predicate level. These numbers are comparable to those obtained in the RTE-2 *ARTE* annotation (Garoufi, 2007), which demonstrates that at least 20% of the positive examples in the RTE-2 challenge test set can be treated by inferences at the frame level (such as nominalizations and argument variations). In this thesis, we will demonstrate that frame semantic analysis –in combination with other resources– provides a relevant contribution to the task at hand.

As illustration, Figure 1.1 and Figure 1.2 show frame semantic analyses of (1.13) and (1.14), respectively. The figures –like similar figures in this thesis– were generated with the SALTO tool (Burchardt, Erk, Frank, Kowalski, Pado, and Pinkal, 2006b), which was mainly designed for manual frame annotation, but can also be used as a viewer for automatically generated annotations (see Chapter 3 for details). In Figure 1.1 and Figure 1.2, both the verb *die* and the phrasal verb *pass away* evoke the frame DEATH. The thematic roles point to constituents of the respective sentences. For example, the PROTAGONIST role of Figure 1.2 points to the filler *A person* and the CAUSE role points to

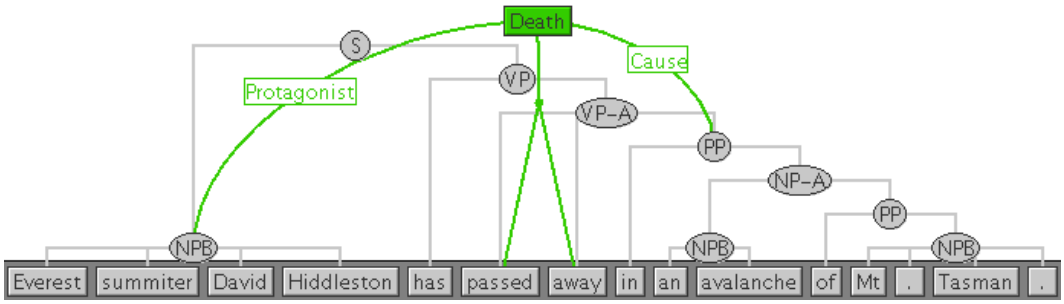


Figure 1.1.: Frame semantic analysis of text (2.4).

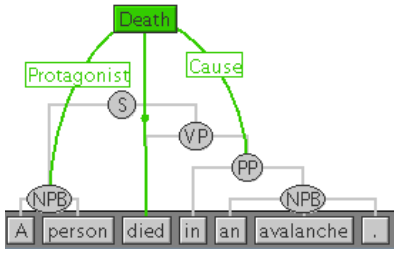


Figure 1.2.: Frame semantic analysis of hypothesis (1.14).

an avalanche. This type of semantic normalization not only provides access to relevant parts of the sentences but also identifies the parts which should correspond for entailment to hold. In this case, the fillers of both roles are compatible.[2] In non-entailed cases like (1.19)-(1.20), the frame semantic annotation can reveal the difference in meaning. Here, *BMW* appears in two incompatible roles, as BUYER in one case and as SELLER in the other case.

(1.19) [Ford]$_{\text{BUYER}}$ bought [Land Rover]$_{\text{GOODS}}$ from [BMW]$_{\text{SELLER}}$ [in 2000]$_{\text{TIME}}$ ....
(frame: COMMERCE_BUY)

(1.20) [BMW]$_{\text{BUYER}}$ bought [Rover]$_{\text{GOODS}}$.
(frame: COMMERCE_BUY)

FrameNet can also contribute to RTE by suggesting not so straightforward, inferential relations via the frame *hierarchy* defined by FrameNet. As an example, consider the pair (1.21)-(1.22) from the RTE-3 development corpus, where entailment can only be established by linking *being arrested* and *being detained*.

[2] The analysis of the compatibility of, e.g., the fillers of the CAUSE roles (*an avalanche* vs. *an avalanche of Mt. Tasman*) goes beyond the analysis of predicate argument structure and has to be achieved by other means, e.g., string comparison as a very first approximation.

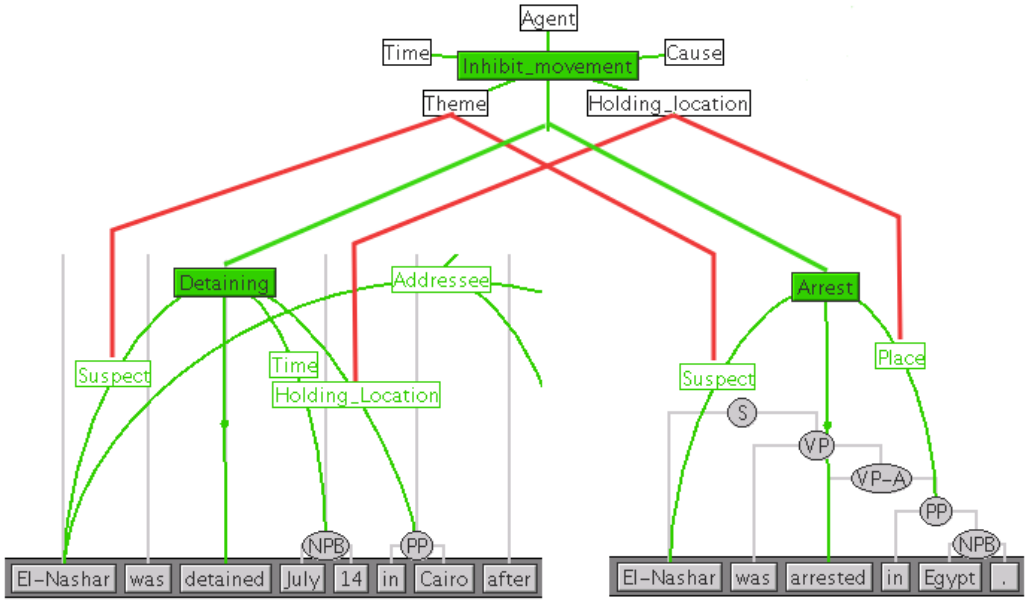


Figure 1.3.: Making use of frame relations.

(1.21) T: El-Nashar was detained July 14 in Cairo. Britain notified Egyptian authorities that it suspected he may have had links to some of the attackers.

(1.22) H: El-Nashar was arrested in Egypt. (TRUE)

As can be seen in Figure 1.3 on the bottom, the main verbs of both sentences evoke different frames (DETAINING vs. ARREST). Also, the roles are slightly different (HOLDING_LOCATION vs. PLACE). Yet, both frames are defined to inherit from a common ancestor, INHIBIT_MOVEMENT. This makes it is possible to come up with a uniform analysis of both sentences.[3]

While the above examples illustrate the intuitive appeal of using FrameNet frames to model textual entailment, frames so far have only rarely been used for this task or for comparable natural language processing tasks. Exceptions are Narayanan and Harabagiu (2004), who use frames as additional, optional feature in a question answering system if available and Fliedner (2007); Frank, Krieger, Xu, Uszkoreit, Crysmann, Jörg, and Schäfer (2006), who both use it for question answering in limited domains. A main reason of why FrameNet is not used more often is probably the incompleteness of the Berkeley FrameNet database (see Section 3.3.4 for details). One of the aims of this thesis is to explore FrameNet's interoperability with other resources for alleviating this coverage issue.

[3] Again, the information that *Cairo* and *Egypt* are in fact compatible has to be provided by other sources.

1.4. Contributions

In this thesis, we demonstrate that frame semantic information can successfully be integrated into an end-to-end system for checking textual entailment. To this end, we present the SALSA RTE system, which is intended as a basis for future research on both frame-based inference systems and lexical-semantic approaches to textual entailment. The system achieved competitive results in RTE challenges. Within several experiments, we thoroughly evaluated the system's performance and pinpoint prerequisites needed for further achievements. Our main contributions are:

(i) An architecture for checking textual entailment based on LFG grammatical and frame semantic analysis, implemented in the SALSA RTE system.

(ii) Interfaces between FrameNet and WordNet ("Detour system"), as well as between FrameNet and the SUMO ontology.

(iii) The design of several experiments for evaluation of frame-based RTE systems, including the FATE corpus.

(iv) A worked-out study of how to use frame semantics for discourse analysis.

Part of this thesis has been collaborative work and some parts of it have been published before elsewhere: (Burchardt and Frank, 2006; Burchardt, Reiter, Thater, and Frank, 2007, (i)), (Burchardt, Erk, and Frank, 2005a; Reiter, 2007, (ii)), (Burchardt and Pennacchiotti, to appear; Burchardt, Pennacchiotti, Thater, and Pinkal, submitted, (iii)), and (Burchardt, Frank, and Pinkal, 2005b, (iv)).

1.5. Structure of this Thesis

In Chapter 2, we will elaborate on the concept of textual entailment, which is a pre-theoretic notion so far. We present its main characteristics and set it in relation to logical entailment. We also present the Recognizing Textual Entailment (RTE) Challenge, the respective corpus data, and give an overview of existing approaches to the task of modeling textual entailment. In a small survey, we illustrate linguistic phenomena occurring in the RTE data and report quantitative studies on the nature of the information that effects entailment in the RTE corpora. We show that lexical-semantics level of predicate-argument structure is especially interesting for modeling textual entailment although existing systems integrate this kind of knowledge only in very limited ways.

Chapter 3 gives an overview of the state of the art in natural language analysis to an extent we consider relevant for modeling textual entailment. Our primary focus will be on the lexical-semantic level of predicate-argument structure. We start from grammatical analysis and proceed to lexical semantic approaches like WordNet, PropBank, and FrameNet. At the end of the chapter, we introduce knowledge ontologies like SUMO, which provide information that goes beyond classical natural language (semantic) analysis and can be helpful in certain cases to confirm or reject entailment.

Two problems impede the usability of components we chose as candidates for our model of textual entailment – a severe problem is the limited coverage of FrameNet and another issue is the absence of suitable natural language interfaces for accessing the information contained in knowledge ontologies like SUMO. Chapter 4 will be concerned with these problems. In order to use FrameNet as an interface to SUMO, we will report a method for a semi-automatic linking of both resources, which we apply, however, only for evaluation purposes. An important contribution that will be used in the SALSA RTE system is the combination of WordNet and FrameNet we implemented in the Detour system. Making use of the good coverage of WordNet, this helps alleviate FrameNet's coverage problem.

In Chapter 5, we present our main contribution – a frame-based approach to textual entailment and its implementation in the SALSA RTE system. We will detail all three stages of the architecture – linguistic analysis, semantic refinement, and matching of text and hypothesis ("entailment reasoning"). The *linguistic analysis* integrates LFG grammatical, frame semantic, and ontological information. For each of text and hypothesis, we construct a tripartite graph structure, which is held together by two so-called *projections*. The analyses are further refined, e.g., by integration of knowledge from different levels of analysis. Based on these linguistic analyses, in the *entailment reasoning* stage, first the "structural and semantic overlap" of text and hypothesis is computed in a robust, declarative way and the result is represented in a structure we call *match graph*. In a second step, feature vectors are derived and the entailment decision is made by a state-of-the-art statistical classifier trained on entailment corpora.

A detailed evaluation of the SALSA RTE system is presented in Chapter 6. After presenting its results in RTE challenges, we inspect system behavior and report various experiments for quantitative and qualitative evaluation we performed. One experiment compares the performance of the linguistically informed system against a shallow baseline in different conditions. In another experiment, we assess the potential for modeling entailment of the system and of frame semantic analysis in general on the basis of a manually constructed gold standard corpus

One result of this thesis is that it will be necessary for further achievements to access information coded in FrameNet's hierarchy, which cannot be interpreted automatically so far. In Chapter 7, we will elaborate problems of and prospects for the usage of the FrameNet hierarchy for natural language processing. To guide future research, we will illustrate issues of the current structure in detail and describe how to arrive at dense frame semantic analyses of multi-sentence fragments with a worked out case study. Chapter 8 will conclude our thesis.

2. Textual Entailment

This chapter is concerned with the concept of *textual entailment*, which we have introduced above as a versatile inference task useful for many natural language processing applications. For a start, we will shortly review the classic notion of logical entailment and point to problems that occur when it is used as model for natural language inference (Section 2.1). In Section 2.2, we will introduce the notion of textual entailment. We discuss its characteristics and clarify its relation to logical entailment. Section 2.3 presents the Recognizing Textual Entailment (RTE) Challenge and describes the RTE corpus data, which was compiled from "real-world" sources. In Section 2.4, we will survey linguistic phenomena occurring in the datasets, followed by a brief discussion of the data. In Section 2.5, we will report system results of the third RTE challenge and present selected existing approaches to the task of modeling textual entailment in detail. Section 2.6 shortly summarizes this chapter.

2.1. Logical Entailment

Traditionally, logic has been used as a framework to model natural language semantics. Correspondingly, entailment between natural language sentences has been understood as more or less equivalent to logical entailment. Logical entailment is a relation between sets of logical formulae, defined in model-theoretic semantics as follows:

> If Δ is a set of sentences, we say that Δ logically entails a sentence ϕ ($\Delta \models \phi$) if and only if every model of Δ is also a model of ϕ.

In short, the models determine the *truth conditions* for the logical sentences involved. Different models represent different (possible) states of affairs. Therefore, the definition above can be paraphrased that for $\Delta \models \phi$ to hold, the truth of all sentences in Δ has to correlate with the truth of ϕ under any possible circumstances – it must be impossible to construct a model that makes all sentences in Δ true while ϕ is false.

The general architecture of the logical approach to checking entailment between natural language sentences is displayed in Figure 2.1. First, the natural language sentences under consideration are translated into logical representations, typically in the style of Montague (1973). It is then checked whether logical entailment between the logical representations holds or not. This can be done, e.g., with theorem proving techniques. The result is assumed to carry over to the original question whether the sentences stand in an entailment relation. A prototype implementation of such an architecture is described in Blackburn et al. (2001). This logical approach to natural language semantics has a number of advantages that result from properties of logical formalisms. A central point

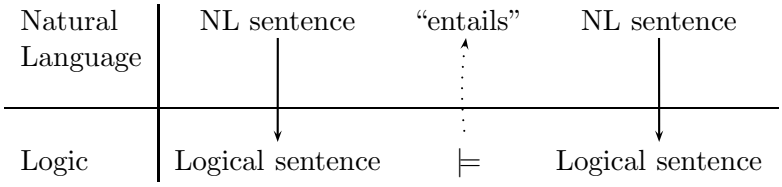


Figure 2.1.: Logical approach to natural language entailment.

is the possibility of providing a formal, model-theoretic interpretation of logical formulae (Tarski, 1983). This makes it possible to capture (natural language) meaning in a mathematically sound way.

However, the entailment relations that can be established in a naive and most straightforward way typically do not correspond to the types of entailment observed in real datasets. One such straightforward example is the entailment between (2.1) and (2.2), as shown in (2.3).

(2.1) John drinks and (John) drives.

(2.2) John drinks.

(2.3) $\{drink(john) \wedge drive(john)\} \models drink(john)$

A more realistic sentence pair is (1.13) and (1.14) (repeated below as (2.4) and (2.5)). Predicate logic translations of both sentences are found in (2.6) and (2.7), respectively. In order to prove entailment in a neat way, we chose a *Davidsonian* translation, using event variables for treating modifiers (cf. Davidson, 1967). Tense is ignored.

(2.4) Everest summiter David Hiddleston has passed away in an avalanche of Mt. Tasman.

(2.5) A person died in an avalanche.

(2.6) $\exists x, e(everest\_summiter(dh) \wedge pass\_away(e, dh) \wedge in(e, x) \wedge avalanche(x) \wedge of(x, mt))$

(2.7) $\exists x, y, e(person(y) \wedge die(e, y) \wedge in(e, x) \wedge avalanche'(x))$

For proving logical entailment between (2.6) and (2.7), additional knowledge is needed. First, a so-called meaning postulate is needed to relate *pass away* and *die*. Second, the fact that everest summiters are persons has to be made explicit (alternatively, one could add the fact that David Hiddleston is a person $person'(dh)$). The knowledge is provided by (2.8) and (2.9), respectively.

(2.8) $\forall e, x(pass\_away'(e, x) \rightarrow die'(e, x))$

(2.9) $\forall y(everest\_summiter(y) \rightarrow person(y))$

These axioms effectively constrain the models under consideration to those that obey this knowledge. The final entailment statement that can logically be proved is this:

$$\{ (2.6) \wedge (2.8) \wedge (2.9) \} \models (2.7)$$

Issues of the Logical Approach

The logic-based approach to natural language semantics has never established for theoretical as well as practical reasons. As Blackburn et al. (2001) already point out, it does not easily scale. First, the translation of real-world sentences into logic is difficult because of issues such as ambiguity or vagueness (e.g. Pinkal, 1995). Second, even for decidable fragments of predicate logic, computation is expensive and requires huge amounts of additional knowledge. The type of additional knowledge that can be needed ranges from linguistic knowledge, e.g., about word meaning, to non-linguistic background knowledge, sometimes called *world knowledge*. This dependency on additional knowledge is problematic as it is an open question how the knowledge can (automatically) be acquired. Moreover, even if vast amounts of such axioms would be available, efficient processing would probably be impossible.

A fundamental problem is that logic inference often exhibits a too high level of precision and strictness as compared to human judgments (cf. Bos and Markert, 2006). It possible to model elementary inferences on this precise level. But many many pragmatic aspects that play a role in "everyday inference" cannot be accounted for. Inferences which are *plausible* but not logically stringent cannot be modeled in a straightforward way. Consider example (1.15)-(1.16) again (repeated below as (2.10)-(2.11)).

(2.10) The anti-terrorist court found two men guilty of murdering Shapour Bakhtiar and his secretary Sorush Katibeh, who were found with their throats cut in August 1991.

(2.11) Shapour Bakhtiar died in 1991.

While humans easily make the plausible "default" inference that (2.11) follows from (2.10), it is not logically entailed. Therefore, this inference cannot be treated in a standard logic setup. Different adjustments to standard logic such as default logic or non-monotonic logic have been proposed to capture defeasible notions of entailment. However, these formalisms are typically very complex, not implemented, and also suffer from the other issues of logical approaches mentioned above.

All in all, there is a conflict between the well-understood, precise theoretical concept of logical entailment on the one hand and an empirical notion of entailment that is based on language user's intuition on the other hand. It has never been shown how the conflict can be resolved.

2.2. Textual Entailment

Starting from observations about the practical intractability of logical approaches and the narrow view on inference they represent, Monz and de Rijke (2001) in their programmatic paper develop the idea of an empirical, corpus-based approach to computational semantics. The authors present an implementation and evaluation of a rather simple method of approximating entailment using word overlap measures. Although Monz and de Rijke (2001) admit that their approach might compute topic overlap rather than proper entailment, they set the stage for a broad investigation of inference in natural language information access tasks. The idea of a new, empirical approach to entailment got a wider recognition when the *Recognizing Textual Entailment* (RTE) initiative (Dagan et al., 2006) introduced the notion of *textual entailment* and set up the ongoing series of RTE challenges.

Below, we introduce the notion of textual entailment (Section 2.2.1). Section 2.2.2 characterizes textual entailment in detail and in Section 2.2.3, we set it in relation to logical entailment.

2.2.1. The Notion of Textual Entailment

Textual entailment is a pre-theoretical notion, described by Dagan et al. (2006) as a directional relationship between a pair of natural language texts – an entailing *text* (T) and an entailed *hypothesis* (H). Entailment[1] holds if the meaning of the hypothesis can be inferred from the meaning of the text as interpreted by a typical language user. The authors state that this somewhat informal definition relates to the common practice of evaluating, e.g., question answering systems, by having human judges rate system results.

This scenario circumvents the discrepancy between observable human behavior and the inference mechanism in the logic-based approach by defining entailment directly on the textual level. Capturing an intuitive notion of natural language entailment was not the only motivation of the RTE initiative. Another motivation was the observation that existing approaches tend to develop and use (approximative) semantic frameworks that fit their very particular needs (e.g., what is called logical form in Harabagiu et al. (2000) or a layered semantic representation in Crouch, Condoravdi, Stolle, King, Everett, and Bobrow (2002)). One disadvantage of this parallel development is that it is difficult to evaluate and compare the performance of related approaches and systems. Dagan et al. (2006) state that "different applications need similar models of linguistic variability" and propose textual entailment as "task that captures major semantic inference needs across applications".

2.2.2. Characteristics of Textual Entailment

Although textual entailment lacks a formal definition, some further specifications of the notion are provided by the RTE initiative, as we will detail below.

[1] From here on, by *entailment* we mean textual entailment if not otherwise indicated.

Probability and Pragmatics

One characteristic property of textual entailment is described in the annotation guidelines for the annotators of the RTE corpora (Dagan et al., 2006):

> "cases in which inference is very probable (but not completely certain) are still judged at *True*."

The inclusion of highly probable, prototypical inferences is intuitively appealing and at the same time challenging for modeling textual entailment. This observation is supported by a short annotation experiment on a randomly chosen subset of 10 pairs from the RTE-1 corpus, which we performed during a reading circle at our department. A central result was that it is relatively easy to decide *whether* textual entailment holds while it often remained controversial *why* this is the case. In particular, it seems difficult to tell whether an inference is strict or just plausible. One of the examples discussed in this experiment is (2.12)-(2.13).

(2.12) Researchers at the Harvard School of Public Health **say** that people who drink coffee **may** be doing a lot more than keeping themselves awake - this kind of consumption **apparently** also **can** help reduce the risk of diseases

(2.13) Coffee drinking has health benefits. (TRUE)

Most participants agreed that textual entailment holds although the respective statement in the text is embedded in multiple modality contexts, marked in boldface above. Still, human judges, who are familiar with journalists writing and who accept the Harvard School of Public Health as authority, consider the modal expressions a matter of a style of cautious formulation. Interestingly, after an extensive discussion of the example, some of the participants in our experiment tended to revise their initial acceptance of this pair as entailed. Another example of such a "probabilistic" entailment pair is (2.14)-(2.15).

(2.14) As a real native Detroiter, I want to remind everyone that Madonna is from Bay City, Mich., a nice place in the thumb of the state's lower peninsula.

(2.15) Madonna was born in Bay City, Mich. (TRUE)

This example clearly shows a pragmatic aspect of textual entailment. In general *be from X* does not entail *being born in X*. However, in order to make sense of the given sentence pair, humans interpret *Bay City* as Madonna's place of birth in (2.14). At the same time, we observe converse examples like (2.16)-(2.17).

(2.16) Iraq's representative to the United Nations, Nizar Hamdoun, announced today, Sunday, that thousands of people **were killed** or injured during the four days of air bombardment against Iraq.

(2.17) Nizar HAMDOON, Iraqi ambassador to the United Nations, announced that thousands of people **could be killed** or wounded due to the aerial bombardment of Iraq. (FALSE)

Here, the hypothesis is in fact logically entailed by the text (if we ignore the report context) –`kill` implies `possibly kill`– but pragmatic principles seem to block entailment.

Text Must Explain Hypothesis

Another feature of textual entailment is that it only holds if the statement in the text licenses the statement in the hypothesis, or as the annotation guidelines in Dagan et al. (2006) put it:

> "In principle, the hypothesis must be fully entailed by the text. Judgment would be *False* if the hypothesis includes parts that cannot be inferred from the text."

This constrains entailment pairs to be semantically highly related. Moreover, there must be a "proof" of the hypothesis on the basis of (mainly) the assumptions given in the text. This constraint explains, why (2.18)-(2.19) does not count as entailed although both statements are true in isolation.

(2.18) The European Union is an economic heavyweight, but it is not a monolith. It works for the good of its own members but it also takes into account the global good.

(2.19) European Union expands its membership. (FALSE)

Admissible Background Knowledge

The quotation above does not explicitly comment on whether additional knowledge may be used to establish entailment. It therefore remains a little unclear how "cannot be inferred from the text" has to be interpreted as it is common practice to make use of some sort of background knowledge for related inference tasks. In their discussion, Dagan et al. (2006) make the impact of background knowledge more explicit in the following statement.

> "Furthermore, the criteria defining what constitutes acceptable background knowledge may be hypothesis dependent. For example, it is inappropriate to assume as background knowledge that the national language of Yemen is Arabic when judging example 1586, since this is exactly the hypothesis in question. On the other hand, such background knowledge might be assumed when examining the entailment "Grew up in Yemen" "Speaks Arabic"."

The example mentioned is displayed below.

(2.20) The Republic of Yemen is an Arab, Islamic and independent sovereign state whose integrity is inviolable, and no part of which may be ceded.

(2.21) The national language of Yemen is Arabic. (TRUE)

Actually, we find the given argument difficult to interpret as it is not clear what "assume as background knowledge" precisely means. However, as long as there is no (formal) theory of textual entailment, which makes it possible to somehow mark or delimit what knowledge is needed for a given inference, discussion of this and related matters is relatively arbitrary if not pointless.

2.2.3. Logical vs. Textual Entailment

On a very abstract level, the two phenomena logical entailment and textual entailment capture the same relation between two "statements" A and B. A (textually) entails B if B "follows from" A. In the case of textual entailment, A and B are natural language sentences and entailment is an *empirical* phenomenon that relies on human judgment. Logical entailment defines a relation between logical sentences within a formal system based on the concept of *truth* and truth conditions.

In Section 2.1, we have argued that logical entailment does not lend itself for capturing a considerable part of natural language inference. While some inferences can be enabled by the inclusion of additional knowledge, defeasible inferences cannot be modeled at all. In fact, the reduction of natural language inference to a truth functional problem sets the focus on the less challenging part of the phenomenon. The gap between the empirical phenomenon and the reach of the formal model is immense.

Textual entailment fills this gap. It sets aside any formal model and captures all types of inferences a typical language user would make "by definition". This makes it possible to study the phenomenon of natural language inference in its entirety and moreover to perform empirical evaluation.

At the same time, the lack of a formal definition of textual entailment poses new challenges for theoretical study of natural language semantics as well as for the design of actual systems. We think that the design of "natural" textual entailment examples (and corpora) that support system development is one of the critical issues. In principle, the two-sentence scheme allows for coding a wide range of inferential relations from simple paraphrases to intricate algebra story problems. While the main objective in the design of positive examples is to represent inferences humans are typically capable of, finding criteria for the design of good negative examples is more difficult. Possible options include random sentence pairs, typical false conclusions humans make, or unintended system results. The latter option has been pursued in the design of existing RTE corpora.

2.3. The Recognizing Textual Entailment Task

The notion of textual has been introduced together with the ongoing series of PASCAL Recognizing Textual Entailment (RTE) Challenges.[2] The first RTE challenge took place in 2005 and since then, there have been annual follow-up challenges (Dagan et al., 2006; Bar-Haim et al., 2006; Giampiccolo et al., 2007a). Below, we present the RTE task (Section 2.3.1) and the respective corpora (Section 2.3.2).

2.3.1. The Task

The scheme of the RTE challenge has not changed much over time. Participants are provided with development corpora containing 800 sentence pairs with annotation of the entailment value (TRUE/FALSE) and the subtask – in the current RTE-3 corpus (Giampiccolo et al., 2007a) question answering (QA, e.g., Moldovan, Harabagiu, Paşca, Mihalcea, Goodrum, Gîrji, and Rus, 1999), information retrieval (IR, e.g., Baeza-Yates and Ribeiro-Neto, 2000), information extraction (IE, e.g. Sundheim, 1991), and summarization (SUM, e.g., Mani, 2001) in equal shares.

The task is to predict correct entailment judgments on a test corpus of the same size as the development corpus. Accuracy is measured as the percentage of correctly classified pairs. True and false entailments are balanced in the corpora, therefore the baseline of guessing is 50% accuracy. Optionally, confidence measures can be handed in as well, which are then taken into account in a second measure, average precision (see Bar-Haim et al., 2006, for details). We postpone an overview of participating systems and detailed results until Section 2.5.

2.3.2. The Textual Entailment Corpora

The corpora provided for the RTE challenges are compiled using "real-world" sources with the aim of representing a certain range of phenomena as described, e.g., by Bar-Haim et al. (2006):

> "Our main focus in creating the RTE-2 dataset was to provide more "realistic" text-hypothesis examples, based mostly on outputs of actual systems. [...] The examples represent different levels of entailment reasoning, such as lexical, syntactic, morphological and logical."

In general, pairs where entailment holds correspond to success cases of existing systems, and non-entailed pairs to cases of failure. Bar-Haim et al. (2006) describe how the text-hypothesis pairs for the four different marked subtasks were compiled by human annotators. In brief, the procedures were the following:

Information Extraction Relations to be extracted from two existing IE tasks such as "X works for Y" were taken as templates for hypotheses. Relevant news articles were collected as texts. Given these texts, actual IE systems then generated positive

[2] www.pascal-network.org/Challenges/RTE/

and negative hypotheses by instantiating "X" and "Y". Additional entailment pairs were compiled manually, Bar-Haim et al. (2006) give as example *An interpreter works for Afghanistan* as non-entailed hypothesis generated for the text *An Afghan interpreter, employed by the United States, was also wounded.* Finally, more pairs were generated manually for relations from domains such as sports, entertainment, or science.

Information Retrieval Hypotheses –formulated in propositional form– were adapted and simplified from existing IR datasets. Entailing text were taken from the datasets, non-entailing texts were retrieved using search engines.

Question Answering Questions from existing datasets and answer texts generated by QA systems were used. The answer texts serve as texts. The questions are transformed into hypotheses by re-formulating them as affirmative sentence and "plugging in" an *answer term* which has been extracted from the answer text. For example, from the question *How many inhabitants does Slovenia have?* and the answer text *In other words, with its 2 million inhabitants, Slovenia has only 5.5 thousand professional soldiers*, the hypothesis *Slovenia has 2 million inhabitants* is generated.

Summarization Texts and Hypothesis are drawn from news document clusters reporting on the same event and the respective output of summarization systems. Annotators were instructed to choose one sentence generated by a system per pair if possible. Hypotheses for both entailed and non-entailed pairs were simplified until fully entailed by the chosen text.

The generated text-hypothesis pairs were judged at least by two of the challenge organizers. The average agreement is reported as 89.2%. Bar-Haim et al. (2006) name as main cases of disagreement cases where i) the text gives approximate numbers and the hypothesis exact ones, ii) the text states a proposition in a report context and the hypothesis the proposition itself, and iii) the hypothesis "makes a slightly stronger statement" than the text. 18.2% of the initial pairs were removed due to disagreement. In a subsequent stage, the organizers again removed 25.5% of the sentence pairs which were judged as too controversial, difficult or redundant (similar to other pairs). Text correction was limited to a minimum in order to provide a realistic sample.

Table 2.1 displays corpus examples from all subtasks – a positive and a negative entailment pair per task. From inspection of these and other examples, we could, however, not derive substantially different characteristics of the pairs across the subtasks. This might be taken as further evidence for the similarity of the needs of different applications. Still, the results of existing RTE systems show varying performance on different subtasks. We will come back to this in Chapter 6.

ID	Text	Hypothesis	Value	Task
1	Claude Chabrol (born June 24, 1930) is a French movie director and has become well known in the 40 years since his first film, Le Beau Serge, for his chilling tales of murder, including Le Boucher.	Le Beau Serge was directed by Chabrol.	YES	IE
5	The Communist Party USA was a small Maoist political party which was founded in 1965 by members of the Communist Party around Michael Laski who took the side of China in the Sino-Soviet split.	Michael Laski was an opponent of China.	IE	
201	Berlin has a new landmark. Among the cranes which still dominate the skyline of Europe's newest capital now stands a chancellery, where the head of government Gerhard Schroeder will live and the German cabinet will hold its regular meetings.	New buildings have been erected in Berlin.	YES	IR
202	The Reichstag building in Berlin was constructed to house the Reichstag, the original parliament of the German Empire. It was opened in 1894 and housed the Reichstag until 1933.	New buildings have been erected in Berlin.	NO	IR
401	It marked the first official visit of Iran's President Mahmoud Ahmadinejad to Saudi Arabia as he spoke with its leader, King Abdullah. Both leaders have expressed concern over sectarian tensions in Iraq, fearing they could spread through the Middle East.	The President of Iran is Mahmoud Ahmadinejad.	YES	QA
403	King Abdullah and Ahmadinejad were believed to be focused on finding ways to end the political standoff in Lebanon between Hezbollah, backed by Iran, and the government of Fouad Siniora, supported by the United States.	The President of Iran is Mahmoud Ahmadinejad.	NO	QA
603	The most common technology in minimills, thin steel slab casting was developed by SMS and put into use by NUCOR.	SMS developed a new steel casting process.	YES	SUM
601	NUCOR has pioneered a giant mini-mill in which steel is poured into continuous casting machines.	Nucor has pioneered the first mini-mill.	NO	SUM

Table 2.1.: Textual entailment examples (from RTE-3 corpus).

2.4. Linguistic Properties of RTE Corpora

The corpus examples in Table 2.1 indicate the variety of types of inferences that potentially relate text and hypothesis. For example, in pair 401, we observe a systematic linguistic variation between an appositive formulation and a copula construction (*X, Y of Z* vs. *X is the Y of Z*). In contrast, in pair 201, considerable background knowledge is needed to infer from cues such as *crane* or ***now*** *stands a chancellery* that new buildings in fact have been erected.

Textual entailment is an inference task on free text. In principle, a considerable number of linguistic phenomena can be encoded into the text/hypothesis scheme. Therefore, it is difficult to a priori delimit the levels of linguistic analysis that can be involved in the entailment decisions. From a practical perspective, a driving question is which levels of analysis, techniques, and resources are best suited to treat a broad range of examples in an effective way.

Before we discuss existing approaches to textual entailment in the next section, we will have a closer look at the data. In Section 2.4.1, we will provide a survey of typical levels of analysis and phenomena we observed in the data. Subsequently, we will report results of more exhaustive studies (Section 2.4.2). In Section 2.4.3, we will refer to a discussion in the community concerning the appropriateness of the RTE corpus data and finally discuss some issues we encountered.

2.4.1. Survey of Linguistic Levels and Phenomena

As has been noted frequently, one of the most indicative features for textual entailment as implemented by the RTE corpora is string identity. This means, in many positive pairs, the decisive part of the hypothesis occurs literally in the text while this is not the case in negative pairs. Therefore, the *lexical baseline* measuring string overlap on content words is already fairly high, at more than 60% on the current corpora. As we will see, lexical overlap indeed is a good measure for entailment on the given data. Yet, the precision and also the recall that can be reached with a surface overlap feature is limited. Starting from surface phenomena, we will now illustrate a broad range of linguistic phenomena occurring in the RTE dataset.

In this context, we will extend the notion of entailment slightly. While entailment is originally defined on the sentence level, we assume that it is also possible to identify "entailment relations" between smaller parts of text and hypothesis like predications or concepts. We might, e.g., say that *car* entails *vehicle*.

Surface Overlap

As has been shown by Garoufi (2007) and others, in positive entailment cases the hypothesis is often a substring of the text. Example (2.22)-(2.23) exhibits 100% surface overlap, i.e., all words of the hypothesis occur in the text.

(2.22) There are currently eleven (**11**) **official** languages (**EU languages**) of the European Union in number.

(2.23) There are **11 official EU languages**. (TRUE)

As the sentence pairs were designed to be "realistic", they include surface differences like bracketing, spelling variants, e.g., of dates or numbers, abbreviations or even typos. This can be a challenge for "deeper" approaches as minor surface differences can lead to unwanted consequences on deeper levels of analysis, e.g., different parses or even parse errors can lead to divergent or missing semantic representations.

Syntax and Grammar

A relatively frequent pattern in the RTE data is syntactic or grammatical variation like passivization or appositive vs. copula constructions as illustrated by (2.24)-(2.25).

(2.24) The **Arabic-language television network Al-Jazeera** reports it has received a statement and a videotape from militants.

(2.25) **Al-Jazeera is an Arabic-language television network**. (TRUE)

The true entailment example above again involves a high degree of lexical overlap. But it can also be the case that measuring overlap leads to wrong predictions. One such difficult example is the negative pair (2.26)-(2.27), which exhibits a high degree of overlap.

(2.26) **Oscar-winning actor Nicolas Cage**'s new son and Superman have sth. in common [...].

(2.27) **Nicolas Cage's new son was awarded an Oscar**. (FALSE)

A fine-grained syntactic analysis is required to detect that the subjects of both predications (*Oscar-winning actor* and *was awarded an Oscar*) are different – at least in the most probable reading of (2.26). This would correctly indicate the absence of an entailment relation. The identity of the predications has to be established by a lexical semantic analysis.

Lexical Semantics and Paraphrases

Semantic phenomena in the textual entailment data often reside within the area of lexical semantics. Lexical variations on the word level such as synonymy or hypernymy as in (2.28)-(2.29) occur frequently. Likewise, one can observe variation on the phrase level as in (2.30)-(2.31).

(2.28) A Union Pacific freight train **hit** five people.

(2.29) A Union Pacific freight train **struck** five people. (TRUE)

(2.30) Satomi Mitarai **died of blood loss**.

(2.31) Satomi Mitarai **bled to death**. (TRUE)

Another type of example is (2.32)-(2.33). In this particular case, the multi-word expression *give up one's right to X* is synonymous to *abdicate* in (2.33) as *X* meets the additional condition of being an official position. And furthermore, it has to be known (or assume) that Edward VIII in fact *was* King and not only designated for the throne.

(2.32) [O]n December 10th 1936 King Edward VIII **gave up his right to the British throne**.

(2.33) King Edward VIII **abdicated** on the 10th of December, 1936. (TRUE)

What is special about this type of example is that expressions in text and hypothesis mutually disambiguate each other. Still, this kind of reasoning goes for the most part beyond what we consider feasible with the available linguistic and knowledge resources.

Modality

The organizers of the RTE challenge try to avoid "relatively delicate logical issues" (Dagan et al., 2006) in the construction of the entailment corpora. Some phenomena like modality, which are typically discussed in the realm of logic, are nevertheless found relatively frequently in the data (cf. Pennacchiotti, 2007). Proper treatment of examples like (2.34)-(2.35) requires checking modality expressions for compatibility.

(2.34) U.S. Secretary of State Condoleezza Rice said Thursday that North Korea **should** return to nuclear disarmament talks [...].

(2.35) North Korea says it **will** rejoin nuclear talks. (FALSE)

In this example, entailment can be rejected because of a modality mismatch – *shall* does not entail *will*.

Reasoning and Background Knowledge

The last class of examples we want to present requires the use of background knowledge and reasoning capabilities, which go beyond the linguistic phenomena we sketched above. In the case of (2.36)-(2.37), geographical knowledge is needed to relate *Chernobyl* with the *USSR* and the *outside the ex-USSR* with *Great Britain*. Contamination has to be related to *repercussions* and it has to be inferred that the text is in fact talking about the Chernobyl *disaster*, which is mentioned only in the hypothesis.

(2.36) Busby countered, telling The Iconoclast, the point is that **material from Chernobyl which is 1,800 miles to the east of Great Britain traveled to Great Britain and contaminated Wales, Scotland, and various parts of the United Kingdom**. [...]

(2.37) **The Chernobyl disaster had repercussions outside the ex-USSR.** (TRUE)

Temporal reasoning is also necessary for pairs like (2.38)-(2.39), although it is supposed not to be by the creators of the corpora. In this example, it is critical to infer from the fact that lifting the ban *again* would lead to extinction that having the ban is efficient for preventing extinction.

(2.38) Definitely do not **lift the ban on ivory**, it will **drive the species to near extinction again**. Elephant numbers should be controlled but what has ivory got to do with it? It is a barbaric trade and should be banned permanently.

(2.39) The **ban on ivory** trade **has been effective in protecting the elephant from extinction**. (TRUE)

It should be clear from the discussion of examples above, that it is not possible to isolate a single source of information as "solution" to the problem of detecting textual entailment. As has been argued (e.g. Litkowski, 2006), *integrated* approaches are needed. Existing approaches typically integrate different knowledge sources and layers of analysis.

2.4.2. Quantitative Studies of Linguistic Factors in Textual Entailment Corpora

Recent annotation studies have been concerned with the question of what levels of analysis and phenomena determine the entailment decisions in the RTE datasets. Bar-Haim et al. (2005) compare the purely lexical level against the lexical level plus syntactic information and report that the latter outperforms the former. The authors also argue that the paraphrase level is of special interest and suggest experiments on the lexical-semantic level as options for future research.

A thorough classification and annotation of linguistic phenomena involved in textual entailment decisions is provided by Garoufi (2007). The author presents a manual annotation of the subset of positive entailment pairs from the RTE-2 test corpus. She shows that phenomena on the level of lexicon, syntax, and discourse are frequently found and that for many examples, some sort of reasoning is necessary. Figure 2.2 from Garoufi (2007) shows the distribution of phenomena that have been observed in the positive pairs of the RTE-2 test corpus. Without going into detail, the chart shows that many high-valued features are from the area of *grammar* and *lexical semantics* (Nominals[3], Genitive, Apposition), as well as concerning co-reference. Another prevalent feature is the *reasoning* feature, which subsumes a variety of inferences that require knowledge which goes beyond what is captured by the linguistic features considered.

[3]The feature *Nominal* describes relations between NPs.

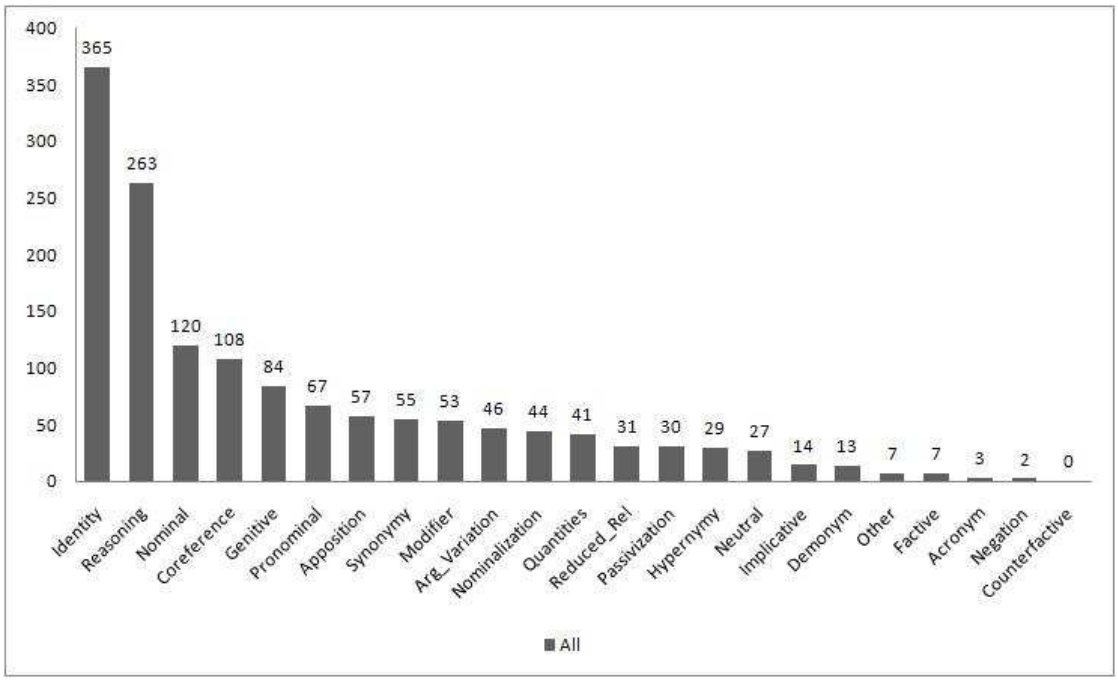


Figure 2.2.: Entailment feature distribution on the positive subset of the RTE-2 test set (from Garoufi, 2007).

2.4.3. Discussion of the Data

The RTE task and corpora have been criticized, mostly for the weak definition of textual entailment and for problematic examples encountered in the corpora. A more fundamental discussion took place between (Zaenen, Karttunen, and Crouch, 2005; Crouch, Karttunen, and Zaenen, 2006) and Manning (Manning, 2006). Zaenen et al. (2005) propose to include (and annotate accordingly) more examples of "classical" phenomena like implicatures or examples involving quantifier monotonicity – phenomena that had been studied intensively in the logic-based approach to natural language semantics. At the same time, they criticize the unclear situation concerning the use of background knowledge ("world knowledge") in the RTE task and suggest to define more precisely which knowledge is allowed or needed. Manning (2006) in contrast defends the design of the RTE task and points out that the ideas of Zaenen et al. (2005) are not realistic and would lead to a much less natural and useful resource. He also argues that trying to further pinpoint the question of what precisely is world knowledge and where it can be used is doomed to failure. In his article, he also gives the following interpretation of textual entailment:

> "One way of thinking about whether an hypothesis follows from a text is: if a

> person asked you for a piece of text that establishes a certain hypothesis, then would showing them the given text satisfy the person. From an operational perspective, thus seems just what we want."

While we agree with Manning's view in general, we think that this operational definition of textual entailment can be used to justify entailment pairs we consider debatable. One such example is (2.40)-(2.41) from the RTE corpus:

(2.40) Albert Sabin developed an oral, attenuated (live) vaccine, which, with Salk's discovery, brought polio under control.

(2.41) Polio is under control in the world. (TRUE)

The example is from the IR section and is judged as true textual entailment. However, in general, the fact that a disease was brought under control at some time in the past does not entail that the disease is still globally under control. Still, if we imagine a situation like the one Manning described – someone asks a medical expert to confirm that polio is under control and the expert cites (2.40) without any further comment, one is inclined to believe that this in fact confirms the statement. But in this situation, decisive factors are the trustworthiness of the expert and the fact that he does not mention recent cases of Polio in addition the information provided by the text.

We prefer the narrower, original definition of textual entailment. As a consequence, we would have either included this as a negative example or maybe deleted it during corpus revision.

2.5. Related Approaches to Textual Entailment

Textual entailment has a pragmatic character and it is defined only informally. Therefore, it cannot be computed directly. Textual entailment is typically *approximated* instead. While first attempts have been made to come up with precise solutions to selected sub-problems (e.g. MacCartney and Manning, 2007; Bobrow et al., 2007), the most widely used approach is to measure *similarity* between text and hypothesis (cf. Pennacchiotti, 2007) in a supervised learning scenario. In Section 2.5.1, we will present the typical system architecture which is instantiated by most current approaches to textual entailment. An overview of participating systems and results of the current, third RTE challenge will be given in Section 2.5.2. We will then present three of the best performing systems of the second RTE challenge (Bos and Markert, 2006; Tatu et al., 2006; Hickl et al., 2006a) in detail (Section 2.5.3). In Section 2.6, we draw conclusions from the discussion of related works.

2.5.1. A General RTE Architecture

The typical architecture instantiated by most approaches and systems for recognizing textual entailment is the pipeline architecture displayed in Figure 2.3 (cf. Bar-Haim et al., 2006). In a preprocessing and analysis step, text and hypothesis are linguistically

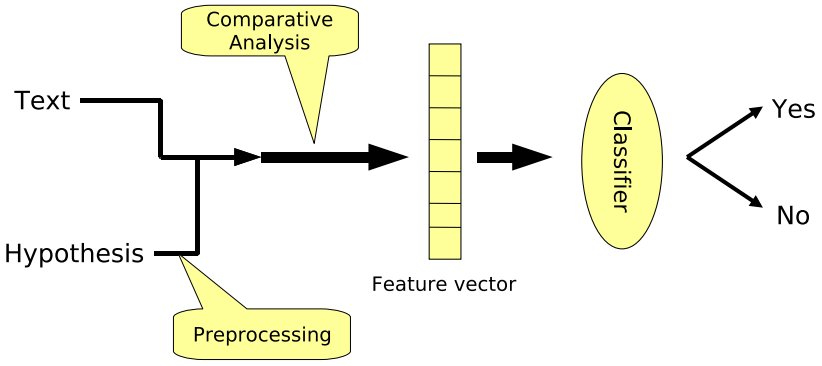


Figure 2.3.: General RTE architecture (from Bar-Haim et al., 2007).

analyzed. The respective levels of analysis range from tokenization (e.g., Adams, Nicolae, Nicolae, and Harabagiu, 2007) over dependency parsing (e.g., Herrera, Penas, Rodrigo, and Verdejo, 2006) to logical analysis (e.g., Bos and Markert, 2006). In a second step, text and hypothesis are compared. Again, a variety of techniques is used, e.g., lexical alignment (e.g., Hickl et al., 2006a) or transformations on tree kernels (e.g., Zanzotto, Pennacchiotti, and Moschitti, 2007). Overall, the most common comparative measures and techniques used are surface-based – lexical overlap (unigram, N-gram, subsequence), lexical substitution (WordNet, statistical), syntactic matching/transformations, as well as the study of lexical-syntactic variations ("paraphrases"). Sometimes, explicit techniques for detecting *mismatch* are also used. (cf. Vanderwende and Dolan, 2005). As concerns "deeper" semantic computation, logical inference is used by some approaches (see below).

The typical result of the comparative stage is a feature vector representing similarity measures of text and hypothesis. Usually, the entailment decision is made by a machine learning component, trained on the RTE training corpora.

While first research have touched the impact of additional material in the hypothesis (e.g. MacCartney and Manning, 2007; Bobrow et al., 2007), it is difficult to make generalizations about how it influences the textual entailment decision.

2.5.2. Results of the Third RTE Challenge

Table 2.3 from (Giampiccolo et al., 2007b) gives an overview of the systems that participated in the third RTE challenge, the latest finished competition by the time of writing. The table and also provides a broad classification of techniques and resources used by the respective systems. Although we will not comment it in detail, the table illustrates the variety of components used by different systems and it also shows the variance in system performance.

23 international groups participated in the challenge – 12 from Europe, 10 from the

	Average accuracy	Best result
SUM	67.9	84.5
IR	60.8	74.5
QA	58.2	70.5
IE	52.2	73.0

Table 2.2.: RTE-3 per task analysis (from Bar-Haim et al., 2007).

USA and Canada, 1 from Australia. The SALSA RTE system, listed under *Burchardt* in the table, was one of two German contributions. Overall, systems' accuracy ranges between 49% and 80%, averaging at 60% (median 59%). This result indicates the difficulty of the task, seen the baseline of 50% for guessing (and about 60% for measuring lexical overlap). What is notable is that most systems perform in a range between 59% and 66% while two systems designed by a private company score considerably better, namely at 71% and 80% (Tatu and Moldovan, 2007; Hickl and Bensley, 2007). This shows that there is room for improvement concerning the resources available in the research community.

A per-task analysis of system results is given in Table 2.2. The average results differ considerably. However, it is still an open question whether this is due to idiosyncrasies in the corpora or whether this indicates, e.g., relative difficulty of the subtasks. We will come back to this in Chapter 6

All in all, the results of the third challenge were slightly better than in the previous challenges. One novelty in this challenge was the inclusion of comparably longer texts. While the first two challenges presented single-sentence pairs, now 10% of the texts consisted of more than one sentence and it is under discussion whether to move towards a general, more realistic multi-sentence setting in the future. As concerns system performance on the longer texts, Giampiccolo et al. (2007a) report a slight drop in accuracy as compared to single-sentence texts (average accuracy 58.7% vs. 61.9%).

2.5.3. Selected Systems

An in-depth discussion of all currently existing approaches to textual entailment is beyond the scope of this thesis. Below, we present three of the best performing systems of the second RTE challenge (Bos and Markert, 2006; Tatu et al., 2006; Hickl et al., 2006a), two of which also participated in the third challenge (Hickl and Bensley, 2007; Tatu and Moldovan, 2007, see Table 2.3). Inspection of these systems is also relevant as they show a variety of methods and techniques that have been used for the task at hand.

Bos and Markert (2006)

The authors present a hybrid system consisting of a "shallow" component based on word overlap and a "deep" component based on semantic analysis. We will focus on the deep

First Author	Accuracy	Average precision	System Components								
			Lexical Relation, WordNet	n-gram\word similarity	Syntactic Matching\Aligning	Semantic Role Labeling\ Framenet\Probank, Verbnet	Logical Inference	Corpus/ Web-based Statistics, LSA	ML Classification	Anaphora resolution	Entailment Corpora – DIRT Background Knowledge
Adams	0.6700		X	X				X	X		
Bar-Haim	0.6112	0.6118	X		X			X		X	X
	0.5837	0.6093	X		X			X		X	
Baral	0.4963	0.5364	X				X				X
Blake	0.6050	0.5897	X		X				X		
	0.6587	0.6096	X		X				X		
Bobrow	0.5112	0.5720	X			X	X				
	0.5150	0.5807	X			X	X				
Burchardt	0.6250		X		X	X					
	0.6262										
Burek	0.5500			X				X			
	0.5500	0.5514									
Chambers	0.6050	0.6341	X		X		X		X	X	
	0.6362	0.6527	X		X		X		X	X	
Clark	0.5088	0.4961	X				X				X
	0.4725	0.4961	X				X				X
Delmonte	0.5875	0.5830	X		X	X	X			X	
Ferrandez	0.6563		X	X	X						
	0.6375										
Ferrés	0.6062		X	X					X		
	0.6150		X	X					X		
Harmling	0.5600	0.5813	X		X				X		
	0.5775	0.5952	X		X				X		
Hickl	0.8000	0.8815	X	X			X		X	X	X
Iftene	0.6913		X		X						X
	0.6913		X		X						X
Li	0.6400		X	X					X		
	0.6488										
Litkowski	0.6125										
Malakasiotis	0.6175	0.6808		X					X		
Marsi	0.5913				X						X
Montejo-Ràez	0.5888		X	X	X				X		
	0.6038		X	X	X				X		
Rodrigo	0.6238		X	X	X				X		
	0.6312		X	X	X				X		
Roth	0.6262		X	X							X
	0.5975				X					X	
Settembre	0.6100	0.6195	X	X					X		
	0.6262	0.6274	X	X					X		
Tatu	0.7225	0.6942	X				X			X	X
	0.7175	0.6797	X				X			X	
Wang	0.6650				X				X		
	0.6687										
Zanzotto	0.6675	0.6674	X		X				X		
	0.6575	0.6732	X		X				X		

Table 2.3.: RTE-3 results and system components (from Giampiccolo et al., 2007b).

component. Using the CCG parser (Bos, Clark, Steedman, Curran, and Hockenmaier, 2004), input sentences are transferred into Discourse Representation Structures (DRSs), the semantic representations used in the DRT framework (Kamp and Reyle, 1993). Figure 2.4 shows an analysis of the sentence pair (2.42)-(2.43) (from Bos and Markert, 2006).

(2.42) Mr. Fitzgerald revealed he was one of several top officials who told Mr. Libby in June 2003 that **Valerie Plame, wife of the former ambassador Joseph Wilson,** worked for the CIA.

(2.43) **Valerie Plame is married to Joseph Wilson**.

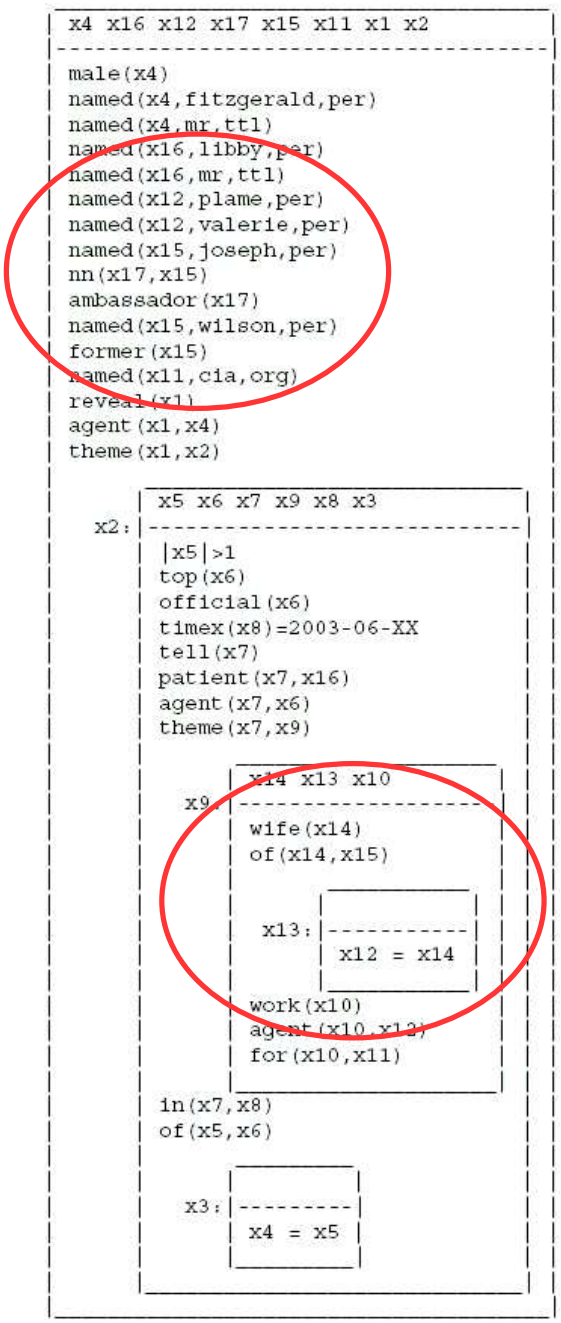
Although going into the details of semantic representation and DRT is not possible here, we want to point out, that these representations deviate from standard DRT. For example, named entities are treated in a uniform way with the `named` predicate and a so-called neo-Davidsonian representation (going back to Davidson, 1967) is used for coding argument structure by way of semantic roles like `agent` or `theme`.

For checking whether entailment holds, Bos and Markert (2006) translate the DRSs into logical formulae and generate what they call "relevant background knowledge" on the basis of WordNet (see Section 3.3.1) and 115 hand-coded axioms. Then they use both theorem proving and model building for generating different entailment measures. This includes checking, e.g., whether direct logical entailment holds between text and hypothesis or whether the background knowledge is consistent with text and hypothesis. One problem of any direct check of entailment is the strictness of the logical setting – it is not possible to have a notion of, e.g., *almost entailed*. To this end, the authors use model building. They compare the domain size of the text (including background knowledge) with the domain size of text and hypothesis. The idea behind is that an entailed hypothesis should not be very informative with respect to the text. Thus the more the domain size grows when adding the hypothesis, the less likely entailment holds.

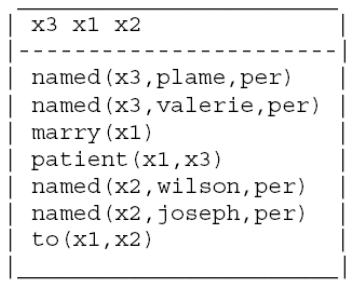
The different resulting measures for theorem proving and model building are then input to a machine learning component. The results reported by Bos and Markert (2006) indicate that the domain size measure is the only informed, semantic feature selected by the machine learner. Overall system accuracy was at 61% at the second RTE challenge.

Example and discussion. The analysis of (2.42)-(2.43) in Figure 2.4 can be used to illustrate some disadvantages of this logical approach. First of all, especially the structure representing the information contained in the text is relatively complex. In contrast, the decisive information is provided in the text by a relatively simple apposition (printed in boldface in (2.42)). In the DRS representation, this information is embedded within different substructures (encircled in the figure). Moreover, the analysis of the apposition, i.e., the identification of *Valerie Plame* and *wife* as referring to the same referent is achieved in a complex way – the discourse referent `x12` refers to *Valerie Plame*, `x15` stands for Joseph Wilson and `x14` represents the the wife of the latter. `x12` is correctly

DRS T:



DRS H:



Axiom:

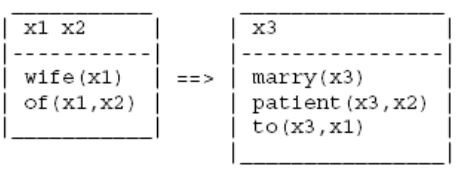


Figure 2.4.: DRSs of (2.42)-(2.43) and axiom used for establishing entailment (from Bos and Markert, 2006).

set equal to `x14` within an extra DRS. Yet, this is not a case of reference resolution, but proper semantic construction.

Figure 2.4 also shows one of the axioms needed to establish the entailment relation here. The axiom is intended to encode the information that `wife(X,Y)` entails `married_to(Y,X)`. Yet, it is provided in an idiosyncratic manner. In the right side of the axiom, it is stated that there is a "marry-event" (`x3`) which is in a "to" relation to the wife of its patient (`to(x3, x1), patient(x3,x2), of(x1,x2)`). Bos and Markert (2006) themselves state the problem that large-scale coverage cannot be reached by manual generation of such axioms. While the authors speculate that the given axioms might be taken for bootstrapping, we doubt that this kind of information can fully automatically be acquired, e.g., from given resources or corpora.

Bos and Markert (2006) work within the "classical" logic paradigm of natural language semantics. Yet, the authors themselves point out that it does not perform better than their lexical baseline system computing word overlap. In fact, both systems succeed in modeling by and large the same sentence pairs. This is supported by the true positive examples the authors presented at their RTE workshop presentation, including (2.44)-(2.45) below. In this example, all material of the hypothesis is literally contained in the text.

(2.44) On Friday evening, **a car bomb exploded outside a** Shiite **mosque** in Iskandariyah, 30 miles south of the capital.

(2.45) **A bomb exploded outside a mosque**. (TRUE)

Overall, our impressions is that this direct translation of natural language into logic translates one highly complex system into another complex system without making appropriate generalizations for the given task. Apart from the problem of generating appropriate representations, the limited flexibility of the logical approach, i.e., the binary decision (logically) entailed/non-entailed, introduces problem that have to be tackled.

Cogex (Tatu et al., 2006)

The authors, members of LCC corporation, participated with their COGEX system in the second RTE challenge. This system is based on the resolution prover COGEX (Moldovan, Clark, Harabagiu, and Maiorano, 2003), which was designed for processing data in a special logic-like format called *logic form* (Moldovan and Rus, 2001). This format was designed as intermediate representation level between natural language surface and standard "deep" semantic representation formats. In the RTE system, two variants ("layers") of logic form are generated, one coding constituent-based information and one coding dependency information. The running example in Tatu et al. (2006) is (2.46) below. The two logic forms on the respective layers provided in Tatu et al. (2006) are displayed in Figure 2.5.

(2.46) Gilda Flores was kidnapped on the 13th of January 1990.

Constituency. $Gilda\_NN(x1)$ & $Flores\_NN(x2)$ & $nn\_NNC(x3, x1, x2)$ & $\_human\_NE(x3)$ & $kidnap\_VB(e1, x9, x3)$ & $on\_IN(e1, x8)$& $13th\_NN(x4)$ & $of\_NN(x5)$ & $January\_NN(x6)$ & $1990\_NN(x7)$ & $nn\_NNC(x8, x4, x5, x6, x7)$ & $\_date\_NE(x8)$& $THM\_SR(x3, e1)$ & $TMP\_SR(x8, e1)$ & $time\_TMP(BeginFN(x1), 1990, 1, 13, 0, 0, 0)$ & $time\_TMP(EndFN(x1), 1990, 1, 13, 23, 59, 59)$ & $during\_TMP(e1, x8)$

Dependency. $Gilda\_Flores\_NN(x2)$ & $\_human\_NE(x2)$ & $kidnap\_VB(e1, x4, x2)$ & $on\_IN(e1, x3)$ & $13th\_NN(x3)$&$of\_IN(x3, x1)$ & $January\_1990\_NN(x1)$

Figure 2.5.: Two different logic forms (from Tatu et al. (2006)).

These logic forms contain syntactic and semantic information. For example, $kidnap\_VB(e1, x9, x3)$ indicates a ditransitive verb and $THM\_SR(x3, e1)$ states that $x3$ (representing *Gilda Flores*) fills the semantic role *theme* of the event denoted by the verb. Another feature of these representations is that time is modeled fairly explicitly – the special predicates like $BeginFN$ in fact are classes of the SUMO ontology (see Section 3.4.1), generated by hand-coded rules.

Apart from the information contained in these logical forms, the COGEX system also generates additional axioms at the subsequent proof stage. The main types of axioms are sketched below. Based on the words occurring in text and hypothesis, so-called *lexical chains* are generated using eXtendedWordNet (Moldovan and Novischi, 2002). These include lexical relations such as synonymy or hyponymy, or relations between Named Entities and the respective adjectives like *Nicaraguan* and *Nicaragua*. By way of rewrite rules, *linguistic knowledge* is used, e.g., to decompose compound nouns or provide normalized representations of coordinations. The last type of axioms codes semantic and *background knowledge*, e.g., that the Pope is the head of the Roman Catholic Church. The authors state that 310 hand-coded axioms were re-use from previous projects and 73 have been manually added for the task of checking entailment.

Moreover, the system implements a so-called *semantic calculus* to increase the connectivity of the semantic analyses. The respective 82 axioms model, e.g., spatial knowledge or transitivity of kinship/causation/isa relations, or implement temporal reasoning.

The COGEX prover used in the subsequent proof step is capable of generating proofs of different "strictness". If a strict proof is not found, constraints on arguments can be relaxed and arguments from the hypothesis can be dropped completely if they cannot be proven from facts in the text. Most importantly, this relaxation proceeds in a weighted manner and the complete proofs is also *scored*. This proof score provides a measure of the difficulty of proving the hypothesis from the text.

In a parallel module, Tatu et al. (2006) also implement a lexical alignment algorithm, which computes the edit distance between text and hypothesis based on a cost function (deletion costs 0, insertion costs depend on the type of word, etc.).

Overall, the system achieved an accuracy of 74% at the RTE-2 challenge. The authors state that the constituent-based logic form is primarily responsible for the overall score. Yet, the combination of the constituent-based, the dependency-based, and the alignment based scores outperforms the individual results (no individual figures have been published).

Discussion. The system of Tatu et al. (2006) it contains a multitude of modules for specific tasks including a database containing a large number of axioms. Moreover, the "logical" prover used is tailored towards the task at hand. It is more flexible than usual theorem provers, offering the possibility to turn non-proofs into proofs by dynamically relaxing the goal and to score proofs. This complex system architecture can extract and process relevant information from the textual data with considerable success. At the same time, even this huge system designed with a lot of manual effort and using established technology performs less than 15% better than a simple lexical baseline. The question is, to what extent this approach scales and how much accuracy can be reached in the same manner (the system performed slightly worse at RTE-3, contrary to the general tendency). Unfortunately, the documentation is not verbose in parts such that the reader is largely left uninformed about how certain functionalities are implemented. One general question is whether comparable performance can be achieved without having access to the manpower, technology, and proprietary resources used here, e.g., if developing a system for a new language.

Groundhog (Hickl et al., 2006a)

The best performing system in the RTE-2 challenge was the Groundhog system presented by members of the LCC company (Hickl et al., 2006a). An overview of the system components is given in Figure 2.6. In the central part of the system, a lexical alignment is computed to identify those portions of text and hypothesis, which most probably belong together. The respective parts also serve as input to a module which derives "paraphrases" from the WWW in order to check whether there exist phrase-level alternations of one of text and hypothesis or both that match. A key feature of this approach is that a huge amount of extra training-data is generated and used at different steps. The final entailment classification is based on features extracted from the different sources. Below, we will sketch the main system components.

Linguistic analysis. The system performs an extensive linguistic analysis of text and hypothesis on different levels including a powerful named-entity recognizer distinguishing 150 categories (locations, sports events, persons, etc.) and tools for linking and resolving referential expressions such as anaphora, relative pronouns, or co-referent nouns. As in Tatu et al. (2006)'s system, time expressions including deixis are normalized and their order is computed. After syntactic parsing, a lexical semantic analysis in terms of predicate-argument structure based on PropBank (see Section 3.3.3) is computed by a statistically trained semantic role labeling system. Finally, heuristics are used to mark polarity and factivity of predicates embedded under respective markers. For example,

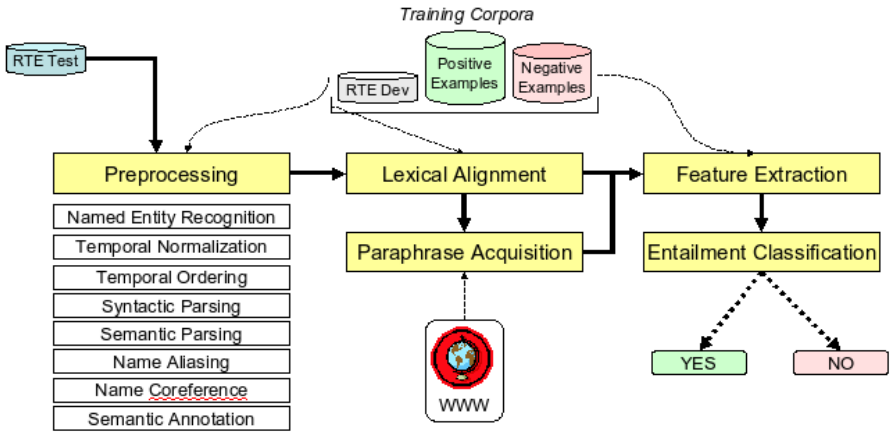


Figure 2.6.: Architecture of Groundhog system (from Hickl et al., 2006b).

in the case of negation or verbs like *refuse*, the embedded predicates are marked *false*. In the case of non-factive speech act verbs like *deny* or *claim* and verbs marking belief, intentions, etc., the embedded predicates are marked as *unresolved*, i.e., neither definitely true nor false. Also, constructions that are known to trigger conventional implicatures such as nominal appositive or non-restrictive relative clauses are analyzed by appending the respective implicature to the sentence. An example from Hickl et al. (2006b) is (2.47)-(2.48) below.

(2.47) Shia pilgrims converge on Karbala to mark the death of Hussein, the prophet Muhammad's grandson, 1300 years ago.

(2.48) Shia pilgrims converge on Karbala to mark the death of Hussein 1300 years ago **AND** Hussein is the prophet Muhammad's grandson.

Lexical alignment. In order to identify compatible material in text and hypothesis, the approach uses machine learning to compute lexical alignment based on (i) statistical features, (ii) lexico-semantic features such as WordNet distance, and (iii) string-based features such as Levenshtein distance. The alignment classifier is trained in two steps. First, 10.000 word pairs from texts and hypotheses of the RTE-2 development set are classified by human annotators as either positive or negative instances of lexical alignment. The resulting data is then generalized by training another classifier on 450.000 alignment pairs extracted from additional entailment corpora generated by the authors (see below).

Additional entailment corpora. The Hickl et al. (2006a) group generated more than 100.000 additional positive entailment pairs by grouping news headlines with the respective first sentences in the style of Burger and Ferro (2005). Additional filtering, e.g.,

Training Set	Accuracy
RTE-2 Dev.	65.25%
25% LCC Corpora + Dev.	67.00%
50% LCC Corpora + Dev.	72.25%
75% LCC Corpora + Dev.	74.38%
100% LCC Corpora + Dev.	75.38%

Table 2.4.: Impact of training data (from Hickl et al., 2006a).

removing those sentence pairs that do not share any NP resulted in an accuracy of 91.8% assessed on a random sample of 2500 pairs. Negative pairs were acquired in two ways. A first set was generated by selecting 100.000 pairs of subsequent sentence from documents that mention the same named entities; a second set was generated by gathering 20.000 sentence pairs that are linked by "contrastive" discourse connectives such as *although* or *in contrast*. The accuracy of these pairs was assessed to be at about 95%. The impact of this additional training material on the overall system performance is shown in Table 2.4 (from Hickl et al., 2006a) displaying performance without and with different percentages of the additional training data. Indeed, the amount of training data is an important factor in the overall system result.

Paraphrase acquisition. The component of the system, which is most successful on its own according to Hickl et al. (2006a) is the acquisition of paraphrases. Unfortunately, details of the component are not provided. The two lexical alignments with the highest confidence as well as "matching instances" from the large corpus described above are used to generate Google queries. The results are then somehow clustered. The general idea is to test whether there exist paraphrases of text or hypothesis within the same or within "near" clusters.

Entailment decision and results. The entailment classifier used by Hickl et al. (2006a) works on four different types of features, which are displayed in Table 2.5. Overall, the system was the most successful system in the RTE-2 challenge with an accuracy of 75%. Table 2.6 (from Hickl et al., 2006a) displays the performance of several combinations of the feature types described above. First of all, compared to the performance of concurrent systems, the three best features alone already perform very well. Yet, combination always results in considerable improvement. The semantic feature adds about 2% precision to the overall result (e.g., from 73.62% to 75.38%).

Discussion. Hickl et al. (2006a)'s system follows a corpus-based approach. As in Tatu et al. (2006)' system, text and hypothesis are extensively linguistically analyzed by various specific sub-modules. They are then compared on different comparably "shallow" levels, e.g., on the lexical level, on the level of predicate-argument structure, or on the

Features	Description
Semantic	The two so-called semantic features count cases of boolean truth-value mismatch as well as polarity mismatch based on the assignment of factivity and polarity, respectively.
Paraphrase	Features derived from the paraphrase clusters measuring, e.g., whether there is a paraphrase available for text or hypothesis and whether the paraphrases match.
Dependency	Several features check whether entities and their arguments in hypothesis and text are classified consistently according to the PropBank argument position. For example, one feature measures strict match while another conflates ARG1 and ARG2 (cf. Section 3.3.3).
Alignment	Alignment features include the length of the longest common substring of hypothesis and text or the number of chunks from the hypothesis that do not have a match in the text.

Table 2.5.: Features used by Hickl et al. (2006a)'s system.

		+Alignment	+Dependency	+Paraphrase
Semantic	58.00	66.25	71.25	75.38
Paraphrase	65.88	69.13	73.62	
Dependency	62.50	68.00		
Alignment	65.25			

Table 2.6.: Performance of different feature combinations (from Hickl et al., 2006a).

paraphrase level. Logical phenomena are approximated only in a limited way (e.g., polarity). A crucial factor this system relies on is the availability of training data. In fact, it is trained on 200.000 sentence pairs rather than the 800 pairs provided by the organizer's of the RTE challenge. We think that the dynamic usage of Google and the creation of huge corpora makes it difficult to compare the results of this system with that of other systems based on the RTE data and other, stable resources. At the same time, the question is what is the upper bound of precision that can be reached in such a statistical manner (at RTE-3, the system performed 5% better as on RTE-2, in-line with the general tendency).

2.6. Summary of this Chapter

In this chapter, we have shed light on different aspects of textual entailment. We have set it in relation to logical entailment and have provided an overview of the linguistic phenomena that can be observed in the currently available entailment corpora.

Textual entailment presents itself as a pre-theoretic notion, defined on pairs of texts.

While this makes it possible to generate corpora with realistic inference problems, the question of how to model this intuitively clear concept best in terms of natural language processing is largely open. We have presented an architecture that has established as best practice, where text and hypothesis are first linguistically analyzed on various layers. In a subsequent processing step, their similarity is computed and the final entailment decision is made by a statistical classifier. We have discussed three different approaches in more detail, a logic-based approach (Bos and Markert, 2006), one approach based on an intermediate level of semantic information (Tatu et al., 2006), and one largely corpus-based approach (Hickl et al., 2006a).

From the overview of systems presented in Table 2.3 as well as from the inspection of selected systems, it is evident that the task of detecting textual entailment requires access to different levels of analysis and detailed modeling of different issues rather than implementing one sophisticated "textual entailment algorithm". Bos and Markert (2006) combine a "shallow" component having high coverage with a semantic component, which exhibits high precision. The system includes knowledge supplied by WordNet and by manually generated rules. Tatu et al. (2006) and (Hickl et al., 2006a) both comprise a multitude of components dealing with different phenomena on different levels of analysis.

While logic-based approaches including resources like WordNet can be successful within their limits, approximative modeling on the level of paraphrase and alignment has proven to work fine if enough training data is available. Predicate-argument structure has been used in the context of RTE only in limited ways so far. Apart from Hickl et al. (2006a), semantic role labeling is used on a small scale by Delmonte et al. (2007); Bobrow et al. (2007), where abstract labels like *agent* or *patient* are applied.

In this thesis, we will pursue the idea of applying the elaborate semantic role concept of FrameNet to the task of modeling textual entailment. Like related approaches, we will combine it with information from other levels of linguistic analysis, in the first place grammatical analysis and lexical-semantic information from WordNet. We will present a graph matching algorithm that identifies related material in hypothesis and text. Our approach has been implemented in the SALSA RTE system, whose architecture instantiates the general scheme presented in Figure 2.3. The system participated in the second and third RTE challenge. It is listed in Table 2.3 as *Burchardt*.

3. Linguistic Analysis and Ontological Resources

The aim of this Chapter is twofold. First of all, it gives an overview of the state of the art in natural language analysis to an extent we consider relevant for modeling textual entailment. Secondly, as we want to implement our model, the respective resources are introduced.

Establishing textual entailment relations can depend on factors operating on very many levels of linguistic analysis. However, to arrive at a concise model, we will restrict ourselves to using mainly a combination of grammatical and lexical semantic analysis. Both are highly relevant for modeling textual entailment as implemented by the currently available data. Still, a remaining class of textual entailment cases rely on additional extra-linguistic knowledge, as provided by knowledge ontologies. We will discuss the integration of this kind of knowledge in Chapter 4.

This Chapter is structured as follows. Section 3.1 gives a brief overview on the different levels of linguistic analysis. Section 3.2 is about syntactic parsing and grammatical analysis. Section 3.3 deals with lexical semantic analysis, in particular with frame semantics as implemented by the Berkeley FrameNet project. Section 3.4 gives an overview of knowledge ontologies.

3.1. Levels of Linguistic Analysis

Three large areas of natural language analysis are *syntax, semantics* and *pragmatics*. Syntactic analysis studies linguistic entities (words, constituents, sentences) as formal objects and describes their structural properties. Semantic analysis studies the (potential) meaning of language entities while pragmatics deals with the communicative use of language. The levels are usually seen as being dependent on each other – semantic analysis depends on syntactic analysis while pragmatic analysis requires a semantic analysis.

Each area is concerned with particular phenomena and aspects of natural language and thus requires special kinds of knowledge and suitable processing techniques. Although we will not deal with all levels at the same depth, Table 3.1 gives an overview of these areas, central sub-areas, and characteristic phenomena.

With respect to processing, the sub-areas in Table 3.1 are roughly ordered in terms of rising difficulty. This is also mirrored in the declining availability of resources. For example, basic syntactic processing is supported by quite efficient and reliable tools. Lexical and sentence semantic processing is to some extent supported by resources and tools while pragmatic analysis today remains a predominantly theoretical issue.

		Phenomena
Syntax	Morphology and word level processing	Tokens, lemmata, parts of speech, abbreviations, measures, dates, named entities
	Constituent Structure	Constituents, subcategorization, agreement
	Grammatical Analysis	More abstract dependencies and functional relations between constituents
Semantics	Lexical	Concepts, concept relations (synonymy, hyponymy), predicate-argument structure
	Compositional (Sentence)	Truth, factivity, modality, negation, implication, quantification
	Discourse	Text coherence (anaphora, definite descriptions, nominal chains), rhetorical structure
Pragmatics	"Systematic"	Presuppositions, Implicatures
	"Spontaneous"	Reliability of source ("trustful reinterpretation"), performance effects

Table 3.1.: Areas of linguistic and semantic analysis.

In our model for textual entailment, we will focus on *lexical semantic* information complemented with *grammatical analysis*, which subsumes more shallow syntactic analysis. Although compositional semantics and discourse phenomena do not play a central role in the given corpus data, we will include an *approximative* treatment of the phenomena we identify based on information provided by the deep syntactic analysis. Still, the model we will present is flexible and open for extensions in this direction. Finally, *pragmatics* has a huge impact on textual entailment, which must not be neglected. In line with other current approaches to textual entailment, we will treat the pragmatic dimension implicitly by using a statistical component, which we train on annotated textual entailment examples.

Extra-Linguistic Knowledge. Arguably, an important factor impacting most levels of language analysis and processing is extra-linguistic knowledge like *ontological knowledge*. However, it is an open question how to clearly distinguish between linguistic knowledge, which is acquired as part of a speaker's competence as a speaker and knowledge, which is based on, e.g., past experiences or observation of the world. In practice, it is often intuitively plausible and reasonable to make a division at some point. For example, the fact that an apple is a fruit is probably linguistic knowledge and would thus be included in a natural language semantic resource. But the fact that apples are rich in vitamins and thus healthy is more likely to be included in a knowledge ontology. Yet, this division should not be confused with the division between strict and defeasible knowledge (the

former typically being easier to formalize in a logical framework). As modeling textual entailment requires linguistic as well as ontological knowledge to variable degrees, we will do without a precise boundary and concentrate on the question of how to access the required knowledge.

3.2. Syntactic Analysis

Syntactic analysis focuses on the structural properties of natural language, taking words and phrases, particularly sentences, as the object of study. On this level of analysis, first of all well-formedness conditions for the combination of linguistic entities are explored. Moreover, syntactic analysis is the basis for most kinds of semantic analysis, as implemented in a so-called *syntax-semantics interface*.

While semantic analyses (see Section 3.3) generalize over linguistic surface structure to provide meaning representations on a more abstract level, different levels of abstraction can be identified also within syntactic analysis, ranging from being relatively "surface-near" to being more semantically oriented. Throughout this thesis, we will focus on the "deeper" syntactic levels of grammatical analysis (Section 3.2.2). Still, we will also touch the "shallow" level of basic syntactic parsing (Section 3.2.1), as it forms the basis for the deeper analysis.

Syntactic processing for recognizing textual entailment. In the last decade, progress has been made in robust, corpus-based probabilistic parsing for English (e.g. Charniak, 2000; Collins, 1999). At the same time, deep grammatical frameworks such as LFG, HPSG, TAGs or CCG have considerably improved in coverage and robustness, benefiting from progress in statistics-based processing (Riezler, King, Kaplan, Crouch, Maxwell, and Johnson., 2002; Copestake and Flickinger, 2000; XTAG Research Group, 2001; Clark, Hockenmaier, and Steedman, 2002). Existing approaches to recognizing textual entailment mostly include some sort of syntactic analysis –even if they do not generate a semantic representation– because entailment relations can to some degree be modeled by comparing the syntactic structure of text and hypothesis (e.g. Zanzotto, Moschitti, Pennacchiotti, and Pazienza, 2006). As we will detail in subsequent chapters, grammatical analysis is suited for modeling a considerable number of entailment relations, especially in cases where linguistic alternation phenomena such as passivization or appositions are involved. We will also devise a frame-semantic projection from a grammatical analysis (see Frank and Erk, 2004), where the grammatical structures do not only serve as anchor for the semantic descriptions. Their coarse-grained, approximative semantic characterization of text and hypothesis will also serve as a fall-back for cases where the semantic analysis is missing or wrong, e.g., for problems of disambiguation or lacking coverage.

3.2.1. Basic Syntactic Analysis

A standard syntactic analysis comprises a morphological analysis and parsing. Morphological analysis studies the regularities within words, and provides, e.g., a classification

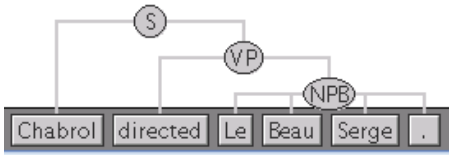


Figure 3.1.: Syntactic analysis of (3.1) with Collins parser.

of the *part of speech* and the *lemma* of a word. For example, *directed* in (3.1) is a transitive verb with the lemma *direct*.

(3.1) Chabrol directed "Le Beau Serge".

The typical result of the process of *parsing* a sentence is its *constituent (phrase) structure*, where the words are grouped into phrases such as noun or verb phrases according to the rules of a context-free grammar. Figure 3.1 shows an analysis in terms of a parse tree generated by the Collins parser (Collins, 1999) for (3.1) (only branching categories are shown). The sentence (S) is analyzed as consisting of the noun *Chabrol* and a verb phrase (VP), which in turn consists of the verb *directed* and the (basic) noun phrase (NPB) *Le Beau Serge*. The phrases are usually named after the so-called *head*, which is most important for the phrase in terms of syntactic properties. The head of the sentence is the main verb.

We will use syntactic analyses as provided by the Collins parser because the Shalmaneser system for automatic frame semantic analysis (see Section 3.3.4) is trained on these structures. In parallel, we will use the LFG parser of Riezler et al. (2002) as it also provides grammatical analyses (see Section 3.2.2).

Ambiguity

A main issue within syntactic analysis is ambiguity. Typically, more than one syntactic structure can be assigned to a given sentence. A standard example is the possibility to attach a modifier to the main verb or to a preceding noun. The problem is that in syntactically parallel cases, one or the other may be the preferred analysis. For example, (3.3) and (3.2) have a parallel structure, but the phrases headed by *with* modify different entities.

(3.2) John bought a car with his first 10.000 Dollars.

(3.3) John bought a car with a removable top.

In (3.2), the buying event is modified and in (3.3) the car. Therefore, the syntactic structures should be different, in (3.2), the noun phrase should span only *a car* and in (3.3), *a car with a removable top* should be one noun phrase. Figure 3.2 shows the intended parse results.

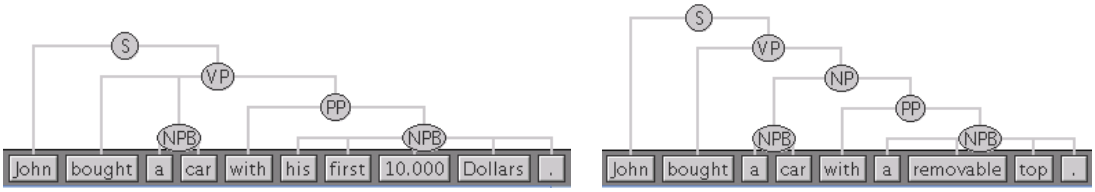


Figure 3.2.: Syntactic analysis of (3.2) and (3.3).

Two central problems relate to this type of ambiguity. First, the number of potential syntactic structures generated may become large for longer sentences since the number of alternatives for all ambiguities multiply. Second, the disambiguation requires accessing the sentence meaning, i.e., semantic processing. In order to keep syntactic and semantic processing apart, two options are (i) either not to resolve the syntactic ambiguity and to further process all possible analyses in parallel, hoping that subsequent semantic processing steps will facilitate the disambiguation, or (ii) to somehow select "the best" analysis and discard the alternatives. The first option is implemented in approaches using *underspecification*, the latter option is often realized by inducing stochastic models and deriving heuristics for automatic disambiguation. In the parsing architecture we use, both options are supported, as well as mixed forms, e.g., working on only the n best parses. We will provide more details below.

3.2.2. Grammatical Functions and LFG

In addition to the comparably shallow, surface-near syntactic analysis shown above, an analysis in terms of grammatical functions makes it possible to define a more meaning-oriented view on syntax and to link the syntactic entities to more abstract meaning representations. Examples of surface phenomena that are abstracted over on this level are word order or sentence mood. Within linguistics, grammar theory is a huge research area. While the observable phenomena are relatively clear, the existing theories and computational models differ. Two current major approaches are *Head-Driven Phrase Structure Grammar* (HPSG, Pollard and Sag, 1994) and *Lexical-functional Grammar* (LFG, Bresnan, 2001). Instead of adhering to a particular theory, we will however concentrate on the widely accepted notion of *grammatical functions* and use an existing LFG Grammar to generate the respective analyses.

Grammatical Functions

A common observation is that the different parts of a sentence have different functions in contributing to the overall meaning. This idea is captured in the notion of *grammatical functions*, the most familiar ones being *subject, predicate*, and *object*. These functions are defined on the sentence level. In (3.4), the predicate is *shave*, the subject is *John*,

and the object is *Bill*.

(3.4) John shaved Bill.

Towards semantic construction. Grammatical functions serve as important clues for assigning semantic representations to surface structure elements. To give a simplified example, a syntax-semantics interface could generate a predicate-logic semantic representation like $shave(john, bill)$ for (3.4) by taking the stem of the predicate as logical predicate and filling the argument slots with appropriate representations of the subject and the object. Yet, the interaction between surface form, grammar, and semantics can be much more complicated than in the simple example (3.4). Just to illustrate this point, consider (3.5).

(3.5) John shaved himself.

Assuming that the reflexive pronoun has been identified to refer to *John*, this sentence can be analyzed parallel to (3.4), resulting in the representation $shave(john, john)$. However, (3.6) and (3.7) have to receive different semantic analyses, although the grammatical structures are also identical.

(3.6) John is asking Mary whether to take along an umbrella.

(3.7) John is asking himself whether to take along an umbrella.

While both *Mary* and *himself* are the direct object of *ask*, only the former should fill a semantic argument slot. Here, a semantic disambiguation is needed to determine the intended reading of *ask* ($ask_1(john, mary, whether)$ vs. $ask_2(john, whether)$). While these examples suggest that there is more to a "real" semantic construction than a grammatical analysis, grammatical functions are a useful level of abstraction for the design of syntax-semantics interfaces. They are also useful to describe cases where the linguistic surface form structurally differs from the semantic argument structure. For example, in (3.8), *John* fills a semantic argument position of *buy* while being only the syntactic subject of *promise*.

(3.8) John promised to buy the tickets.

This phenomenon called (subject) control occurs with a number of verbs following the same pattern – the subject position of the embedded verb is "shared" with the subject of the matrix sentence. Such systematic non-local relations and a number of other linguistic variations can be modeled using grammatical functions. Therefore, they will play an important role in our approach to textual entailment.

Arguments and modifiers. A common distinction which can also be expressed in terms of grammatical functions (of mostly verbs) is that between *arguments* (complements), which are obligatory for forming a meaningful clause and *modifiers* (adjuncts), which provide more general, circumstantial information. One test for argumenthood is whether the respective constituent can be left out. For example, for the transitive verb *direct*, both the director and the piece of art are arguments. Leaving out either typically results in an ungrammatical sentence, as in (3.9) through (3.11). In certain contexts, however, arguments can also be left uninstantiated as in (3.12).

(3.9) Chabrol directed Le Beau Serge.

(3.10) * Directed Le Beau Serge.

(3.11) * Chabrol directed.

(3.12) Before he became producer, Chabrol had directed for 25 years.

For the task of recognizing textual entailment, both proper arguments and circumstantial information may be needed to approve or reject an entailment judgment. For example, (3.13) is not entailed by (3.12) because the time spans where Chabrol directed differ, irrespective of what he directed. Therefore, both arguments and complements have to be considered.

(3.13) Chabrol directed for 35 years.

Alternations

Often, (almost) identical meaning can be expressed with different sentences. For example, (3.14) through (3.16) all convey the information that John Smith is the head of the department in different syntactic constructions.

(3.14) John Smith is the head of the department. [He called me this morning.]

(3.15) John Smith, the head of the department, [called me this morning.]

(3.16) The head of the department, John Smith, [called me this morning.]

These examples belong to a class of regular alternations, which can be explained on the basis of a grammatical analysis alone. In other words, for these cases, it neither matters what the single words mean, nor what kind of semantic representations are used – the analyses have to be identical. For the given kind of variation, both the predicative formulation "x is the Y" and the appositive formulations "x, the Y,..." or "The Y, x" should receive a semantic analysis like $Y(x)$, e.g., $head\_of\_department(john\_smith)$.

A large number of so-called *diathesis alternations* such as dative shift or passivization have been discussed in the literature (e.g. Levin, 1993). They can be seen as a kind of semantic normalization. Since textual entailment is concerned with determining to what

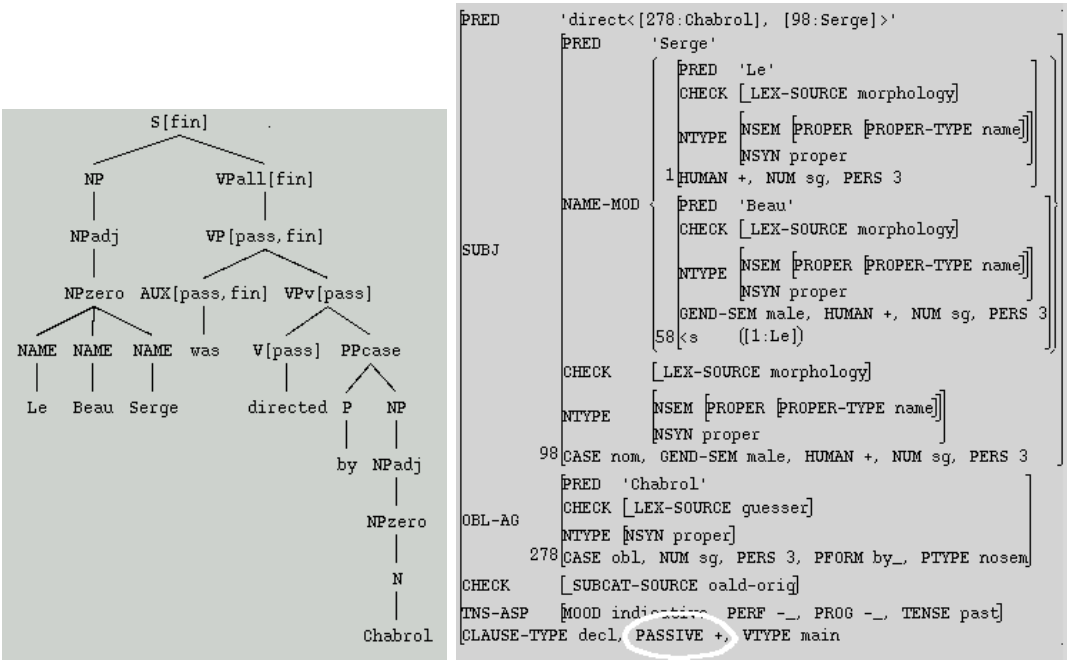


Figure 3.3.: LFG analysis of (3.17) – c-structure (left) and f-structure (right).

degree two sentences are compatible in meaning, such alternations are important clues within the entailment decision.

We will now introduce LFG, the grammar framework we will use. As a further illustration of alternations, we show how active-passive diathesis can be treated within LFG. In subsequent chapters, we will discuss more and different types of alternations. As example sentence, we take (3.17) from the RTE-3 test corpus, the passive formulation of (3.9).

(3.17) "Le Beau Serge" was directed by Chabrol.

LFG

Within LFG, the two main levels of analysis are the constituent level, captured in the so-called *c-structure* and the functional level represented by the so-called *f-structure*. We will mostly be concerned with the f-structures, which are centered around dependency analyses in terms of grammatical functions. An LFG analysis of (3.17) is displayed in Figure 3.3. The c-structure on the right contains basically the same type of information as we have seen in the Collins parse of (3.9) in Figure 3.1. Apart from the pure constituent structure, additional grammatical features such as finiteness or voice are attached to the verbal constituents. The verb phrase *was directed by Chabrol* has been

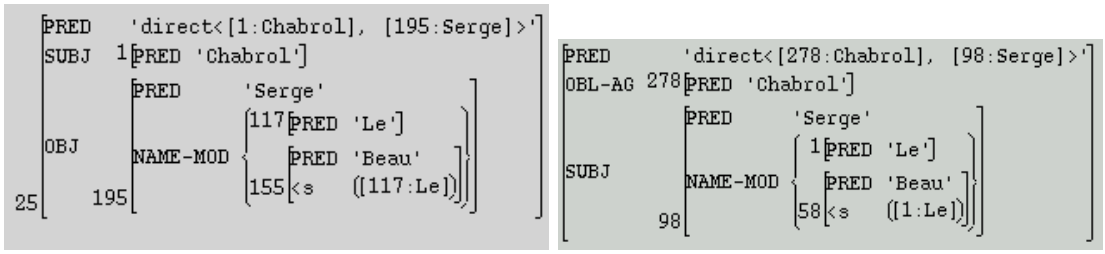


Figure 3.4.: Abbreviated f-structures for (3.9) (left) and (3.17) (right).

recognized as being in passive voice (`VP[pass,fin]`). The corresponding f-structure is shown on the right of Figure 3.3. It contains attribute-value structures for the functionally relevant parts of the sentence, in this case the subject (`SUBJ`) with the predicate (`PRED`) feature value *Serge* and the oblique argument (`OBL-AG`) with the predicate feature value *Chabrol*.

For building syntax-semantics interfaces, the *feature values* are of special interest, especially those of the predicate (`PRED`) feature. Throughout this thesis, we will refer to both the values of the predicate features and also the complete corresponding attribute-value structures as *(LFG) predicate*. The main predicate of (3.17) is $direct(Chabrol, Serge)$, e.g., or for short *direct*.[1] Apart from grammatical functions, f-structures also provide information about grammatical features such as sentence mood and tense, as well as case and number of nouns, determiner types, or types of proper names. We will make use of this information where appropriate.

In order to illustrate how the passive diathesis is accounted for in LFG, Figure 3.4 displays abbreviated f-structures for (3.17) and (3.9). The meaning of an active sentence and the respective passive formulation are for the most part the same. Yet, the surface realization of the arguments is different. The latter is reflected in different functional analyses of both sentences in the given f-structures, e.g., the subject being *Chabrol* in one case and *Le Beau Serge* in the other. The main predicates for both sentences are in fact identical – the LFG analysis has taken care of the alternation.

LFG Parsing Architecture

We use the state-of-the-art LFG parser of Riezler et al. (2002) for syntactic processing. It is based on a hand-coded LFG grammar. A comparative evaluation in Kaplan, Riezler, King, Maxwell III, Vasserman, and Crouch (2004), e.g., shows that it achieves similar parse times, but higher f-scores compared to Collins (1999). The parser includes a named entity recognizer and a morphological guesser to be able to treat names and

[1] Note that the contribution of the PP *by Chabrol* to the semantic representation is the *semantic* head *Chabrol* and not the syntactic head *by*. Most of the time, semantic and syntactic head are identical. From now on, by *head*, we refer to the semantic head.

other unknown words. Another feature which makes the parser robust –supported by LFG's functional paradigm– is the possibility of partial parsing. If a sentence cannot fully be analyzed, the system provides a fragmentary analysis using a special grammar. As the relevant information needed for a textual entailment decision is often contained in a (local) part of a long sentence, a fragmentary analysis of this part can suffice to provide the necessary information. More details on the LFG architecture can be found in Kaplan et al. (2004).

With regard to the problem of ambiguity, the LFG system offers both options we discussed in Section 3.2.1 – either (i) not to resolve the syntactic ambiguity and to further process all possible analyses in parallel or (ii) to select "the best" analysis and discard the alternatives. For (ii), the system includes a stochastic disambiguation component. It supports (i) by representing alternatives (in particular on the level of f-structures) in a *packed* representation format. The re-write system by Crouch (2005), which can be used for subsequent processing, enables working on the packed structures in an elegant and efficient way. As we use this re-write system in our recognizing textual entailment setting, e.g., for projecting semantic analyses and for matching text and hypothesis, it would in principle be possible to work on all analyses of text and hypothesis in parallel. Disambiguation could then take place when text and hypothesis are matched, e.g., by selecting the best fitting analyses, or by computing some kind of mean over all analyses. However, to reduce the number of parameters of the model, we use the stochastic disambiguation component and work on the most probable syntactic analysis for the time being.

3.3. Lexical Semantic Analysis

Within natural language semantics, a standard approach is to represent sentence meaning as a logical formula, typically of predicate logic. For example, translations of the two sentences (3.18) and (3.19) in the style of Davidson (1967) are given in (3.20) and (3.21).

(3.18) John passed away in an accident.

(3.19) John died.

(3.20) $\exists e \exists x \ (pass\_away(e, j) \wedge in(e, x) \wedge accident(x))$

(3.21) $\exists e \ die(e, j)$

These logical representations are normally generated along a syntactic analysis in the way Montague (1973) proposed. This kind of analysis, often called *compositional semantics*, is traditionally mostly concerned with those words and constructions that determine the logical structure of the formula, e.g., introducing logical connectives or scope and modality operators. Open-class words like *pass away* or *accident* in the examples above are typically not further analyzed.

Logic-based compositional semantic analysis has a number of theoretically appealing properties. It is amongst others possible to provide concise, truth-conditional interpretations of the formulae and to use deduction calculi to model inference processes. But in practice, when it comes to processing real-world sentences, this type of analysis faces a number of problems. First, it is difficult to automatically construct an appropriate logical formula for naturally occurring sentences, ambiguity being a central problem. Second, for many tasks like determining that (3.21) can be inferred from (3.20), information about the meaning of expressions like *pass away* and *die* and their relation is needed, e.g., in the form of axioms like (3.22).

(3.22) $\forall e \forall x(pass\_away(e, x) \rightarrow \exists e2(die(e2, x)))$

A complementary approach to natural language semantics, where the meaning of open class content words such as the noun *accident* or the verb *die* is focused, is the area of *lexical semantics*. Lexical semantics typically regards only the word level or local syntactic structures like, e.g., the level of predicate-argument structure. Lexical interpretation, i.e., assigning meaning to words in a given sentence is largely concerned with *word sense disambiguation*, i.e., selecting the appropriate concept among multiple readings of surface words, e.g., *chair* as a piece of furniture vs. a professorship and with identifying relations between predicates and their arguments.

Part of lexical semantics is the study of alternations and generalizations on the level of predicates and argument structure, which are important factors in identifying textual entailment relations. Such generalizations range from lexical variation (`buy` $\equiv$ `purchase`) over argument realization (`marry(X,Y)` $\equiv$ `marry(Y,X)` $\equiv$ `marry(Y`$\bigoplus$`X)`) to more involved relations between complex events (`buy(X,Y,Z)` $\equiv$ `sell(Y,X,Z)`).

In subsequent chapters, we will develop and evaluate a model for textual entailment which is to a large extent based on lexical semantic information. As we will see, the type of inference that is found in the currently available textual entailment corpora is often based on "local" lexical knowledge rather than on the logical structure of the sometimes relatively long sentences. Moreover, lexical semantic analysis is typically scalable by adjusting the granularity and depth of analysis. Note, however, that compositional semantics and lexical semantics are not opposed. They are complementary – a comprehensive description of a sentence meaning would contain both information about (content) word meaning and about the logical relations between the different propositions, entities and properties. Likewise, the fact that compositional semantics is mostly based on logical representation and that most lexical semantic approaches we will discuss do not use logic-based representations is not a fundamental difference between both types of analysis. Their difference is rather the focus of attention. While a prominent task in compositional semantics is to construct comparably few grammar rules that generate the intended logical formulae for given sentences, lexical semantics is concerned with issues related to the potential amount of information that could be contained in open class word meaning representations. To take up the example from above, one could ask what the (defining/relevant/typical/...) properties of dying are or accidents are and how they should be captured. Follow-up questions are how to acquire, store, and access the information in an efficient way.

Lexical meaning can be described along many dimensions and at varying levels of granularity. For example, the meaning of a word can be characterized w.r.t. other words (e.g. lexical relations in WordNet (Miller, Beckwith, Fellbaum, Gross, and Miller, 1990) or verb alternation classes in Levin (1993)), it can be decomposed into primitive semantic predicates, (e.g. *kill* = `CAUSE(BECOME(NOT(ALIVE)))`, see Lakoff, 1972; Schank, 1973; Jackendoff, 1990), or described w.r.t its semantic arguments (e.g. in terms of predicate-argument structure as in the Prague Dependency Treebank, PropBank, or FrameNet, see Žabokrtský, 2000; Kingsbury, Palmer, and Marcus, 2002; Baker et al., 1998). As we want to be able to implement and empirically evaluate our model, we will focus on those frameworks which provide resources for English with non-trivial coverage. They typically model only partial aspects of lexical meaning like concept relations in WordNet and VerbOcean or predicate-argument structures in PropBank and FrameNet. We will present these approaches in turn, and in subsequent chapters, we will show how to combine them in order to be able to access the needed information at different levels of granularity.

3.3.1. WordNet

Hierarchical inheritance structures (IS-A-hierarchies, ontologies, terminologies, semantic networks) have proven efficient and maintainable for representing huge amounts of structured knowledge. Semantic networks had their advent in the field of psycho-linguistic studies (Collins and Quillian, 1969). Originally designed by psycholinguists as an online dictionary being organized in a new, semantically-oriented way (Miller et al., 1990), WordNet has become the largest lexical-semantic resource for English. We will first describe its structure and then show how the information it contains can be accessed and used for natural language processing.

Structure of WordNet

WordNet includes nouns, verbs, and adjectives/adverbs. We will concentrate on nouns and verbs, which are modeled in two separate subsumption hierarchies. The hierarchies are structured by the *hypernymy* relation, which is paraphrased as "is a" for nouns and "is one way of" for verbs. A number of other relations that hold either between synsets (see below) or between words are also represented (e.g. antonomy, meronymy), but we will focus on the hypernymy relation. The hypernyms of the first reading of *cat* are displayed in Figure 3.5. This reads as *a cat is a feline, a feline is a carnivore, etc.* As the hypernymy relation is transitive, it also reads, e.g, *a cat is a vertebrate* or *a cat is an entity.*

Each node in the hierarchy is called a *synset* ("synonym set"), which is the basic unit in WordNet. A synset is a set of synonymous, lemmatized words or collocated multi-word units such as *true cat*. It represents what is often called a *concept* in semantic literature.

At the same time, each synset containing a given lemma represents one *word sense* (or *reading*) of the lemma. As illustration, Table 3.2 lists all synsets for the verb *buy*.

```
{cat, true cat}
   => {feline, felid}
       => {carnivore}
           => {placental, placental mammal, eutherian, eutherian mammal}
               => {mammal}
                   => {vertebrate, craniate}
                       => {chordate}
                           => {animal, animate being, beast, brute, }
                              creature, fauna}
                               => {organism, being}
                                   => {living thing, animate thing}
                                       => {object, physical object}
                                           => {entity}
```

Figure 3.5.: WordNet hypernyms of `cat#n#1`.

Within each part of speech, the synsets are ordered according to frequency. The first synset represents the most frequent sense. So, the most frequent reading of *buy* is the one synonymous to *purchase* and the least frequent is the sense of "being worth". We will use the abbreviation `buy#v#1`, e.g., to refer to the first synsets of the verb *buy*.

Compared to other resources, WordNet has a very good coverage. The current WordNet database contains about 150,000 words organized in over 115,000 synsets. Similar, typically smaller, resources are available for various languages, such as GermanNet for German (Hamp and Feldweg, 1997).

Using WordNet for Natural Language Processing

In recent years, WordNet has proven useful in many different settings and applications. It can be said to be one of the most successful linguistic resources these days. As the machine readable information WordNet provides may seem relatively sparse at first sight, this success might be a bit surprising. Consider Table 3.2 again. From a machine perspective, the gloss and examples are not (easily) accessible. The information that remains is that the verb *buy* has five senses and that the first sense is synonymous to *purchase* and the second sense to *corrupt*, etc. An obvious problem is that senses three to five cannot be distinguished straightforwardly. Leaving this disambiguation problem aside, even if it has been determined that an occurrence of *buy* belongs to the first synset, the information provided by WordNet may not be what one would expect from a meaning representation of the concept of buying – there is no connection to concepts like humans, money, goods, ownership or the like and no mention of transactions involved.

The information that can be accessed includes that *buy* is a synonym of *purchase* and

Synset	Gloss	Example
{buy, purchase}	(obtain by purchase; acquire by means of a financial transaction)	"The family purchased a new car"
{bribe, corrupt, buy, grease one's palms}	(make illegal payments to in exchange for favors or influence)	"This judge can be bought"
{buy}	(acquire by trade or sacrifice or exchange)	"She wanted to buy his love with her dedication to him and his work"
{buy}	(accept as true)	"I can't buy this story"
{buy }	(be worth or be capable of buying)	"This sum will buy you a ride on the train"

Table 3.2.: WordNet senses of the verb *buy*.

that *get* and *acquire* are hypernyms. Moreover, *sister terms*, i.e., other troponyms[2] of *get* like *rent* or *win* can be accessed. They represent related –yet different– concepts, which must not be confused with buying.

This implicit, approximative semantic characterization of concepts, which can be read off WordNet's clear structure, combined with the high coverage, are a key to its success. For many applications, these semantic generalizations seem to be sufficient, an example being *query expansion* in question answering. If a user query contains the words in (3.23), the alternative query (3.24) can be generated varying the verb using synonymy information. (3.25) can be generated using hypernymy information. Likewise (3.26), varying the object.[3]

(3.23) buy, CDs, Berkeley

(3.24) purchase, CDs, Berkeley

(3.25) get, CDs, Berkeley

(3.26) buy, records, Berkeley

Accessing Semantic Information from WordNet. WordNet can in general be used to approximately model *semantic similarity* (and thus *dissimilarity*) of terms by making use of the *meaning postulates* represented by WordNet's relations. For example, Bos (2005) reports a technique of generating logical axioms along the lines of the relations.

[2]Troponymy is the inverse of hypernymy for verbs, paraphrased as "particular ways to".

[3]Some named entities like Berkeley or Beethoven are also included in WordNet. They are related to their super-concepts via the `INSTANCE` relation. In the given example query, *Berkeley*, however, should be present and not be replaced by, e.g., *city*.

Axioms like $\forall x.cat(x) \rightarrow mammal(x)$ are licensed by the hypernymy relation, while, e.g., $\forall x.cat(x) \rightarrow \neg bigcat(x)$ is licensed as both are sister terms. In a related approach, called *eXtended WordNet*, Moldovan and Rus (2001) parsed WordNet's glosses and translated them into a formal representation format they call *logic form*. An example is the logic form (3.28) of the sentence (3.27).

(3.27) The Earth provides the food we eat every day.

(3.28) $Earth : n\_\#1(x1)\ provide : v\_\#2(e1, x1, x2)\ food : n\_\#1(x2)\ we(x3)$
$eat : v\_\#1(e2, x3, x2; x4)\ day : n\_\#1(x4)$

WordNet's structure can also be used to generate qualified similarity judgments, taking into account that, e.g., the concept of cat is more similar to that of mammal than to that of a physical entity. More than 20 different similarity measures have been proposed in the literature, starting from simple (hypernymy) path distance up to measures that include (preprocessed) information from the synsets' glosses (e.g. Resnik, 1995; Lin, 1998; Moldovan and Rus, 2001). Implementations of these semantic distance measures are available in the Perl CPAN archive[4].

Word Sense Disambiguation. WordNet is often used as an interface layer between natural language resources or between natural language and knowledge resources. Examples include linking WordNet and FrameNet or linking WordNet and the SUMO ontology (Shi and Mihalcea, 2005; Niles, 2003). Here, WordNet mainly serves as a reference for *sense disambiguation*.

Word sense disambiguation, i.e., assigning synsets to words occurring in text, has been studied intensively for quite a while (see Agirre and Edmonds, 2006, for an overview), e.g., in the context of the Senseval task (Kilgarriff and Rosenzweig, 2000). However, only few implemented systems are available, such as the one by Patwardhan, Banerjee, and Pedersen (2005). This system is rule-based and follows the common strategy of trying to determine the sense of a target word within a context window of words surrounding it. This works well in cases where context information is available, e.g., the noun *director* is disambiguated correctly by the system for (3.29), but not so well for cases like (3.30), where no appropriate context words such as *movie* are available.

(3.29) Claude Chabrol (born June 24, 1930) is a French movie **director** and has become well-known in the 40 years since his first film, Le Beau Serge.

(3.30) Claude Chabrol is the **director** of Le Beau Serge.

Another class of related systems uses stochastic techniques to learn the appropriate senses from corpora. As word sense disambiguation is usually a resource-intensive and hard task, often, a simple, yet effective approximation is used – the so-called *first sense heuristics* (see e.g. McCarthy, Koeling, Weeds, and Carroll, 2004). The idea is to always assign the first and thus most probable sense to a given word.

[4] `http://search.cpan.org/dist/WordNet-Similarity/`

```
{sell}
    => {exchange, change, interchange}
        => {transfer}
```

Figure 3.6.: WordNet hypernyms of `sell#v#1`.

Limitations

Lexical semantic networks like WordNet are especially well-suited for coding subsumption relations between concepts which are denoted by referential common nouns. Facts like the one that all cats are mammals (see Figure 3.5) or that all cars are vehicles can be coded in a compact way. For verbs denoting events, this scheme is more problematic because events have more internal structure –they happen at a time and a place, participants are involved, they have pre-conditions, sub-parts, and consequences– (see Fellbaum, 1990), which is mirrored in the argument structure of the respective predicates. These features are disregarded in taxonomies like WordNet. Consequently, WordNet's verb hierarchy is by far not as deep as that for nouns, e.g., Figure 3.6 displays the hypernyms of `sell#v#1` (*exchange or deliver for money or its equivalent*).

Likewise, relations between complex event-denoting predicates (*buy* vs. *sell*) cannot easily be represented on the level of words/senses alone. Another limitation of WordNet is the separation of parts of speech. For example, relations between nouns and verbs are not represented in a semantically motivated manner, but only on the basis of derivational morphology. For example, there is a direct link between *seller* and *sell*, but not between *vendor* and *sell*.

3.3.2. Role Semantics

A central issue in capturing the semantics of verbs is to describe how syntactic arguments of the verbs participate in semantic descriptions of the event or state they denote, known as *argument linking*. In earlier times, a number of related approaches have tried to define a set of *semantic roles* (also known as *thematic roles*, *case roles*, *thematic relations* or *case relations*), which describe the relations of verbs and their arguments in a general and compact way. For example, Fillmore (1968) develops a system of so-called *case frames*, formulated in terms of a small set of universal concepts like *Agentive*, *Instrument*, or *Factitive*. Language-specific rules transform the case frames into syntactically correct surface realizations. Different surface realizations like aspectual variations or passivization are explained by different rule applications. Related approaches (e.g. Gruber, 1965; Jackendoff, 1972) differ, e.g., in the number of roles which are assumed and in the way they are identified and described.

However, a convincing "natural set" of thematic roles, which is generally applicable, has never been found. Problematic examples, which speak against the assumption that

a small set of universal roles exists, are reciprocal verbs like *resemble*, where one role would apply twice:

(3.31) [John]$_{\text{THEME?}}$ resembles [Mary]$_{\text{THEME?}}$.

In fact, Fillmore himself soon became one of the critics of this approach and identified as another problem of this approach the amalgamation of grammatical and semantic information. Fillmore (1982) states that on the one hand semantic generalizations between related words such as *buy* and *sell* cannot be represented in a principled way while at the same time semantic distinctions cannot be captured, e.g. the distinction between a GOAL and a RECEIVER in current terms.

Starting from the observation that more flexibility is needed in the argument linking process, Dowty (1991) proposed to use what he calls *proto roles*. These roles like PROTO-AGENT or PROTO-PATIENT are a kind of cluster concept, where individual verb arguments have different "degrees of membership". The proto roles are characterized by a number of semantic features such as *volitional involvement* or *causally effected by the event* and in the linking process, the arguments with the most agentive properties are assigned the AGENT role, and likewise for other proto roles.

A different approach has been pursued especially in the context of building resources containing semantic role information. Instead of defining a small set of universal roles, the two large resources for English, PropBank and FrameNet (Kingsbury et al., 2002; Baker et al., 1998), define specific role sets for single verbs in the case of PropBank and for collections of verbs (and also nouns and adjectives) in the case of FrameNet. Later, we will use FrameNet as semantic resource in our approach to textual entailment. Below, we first give an overview of PropBank and then argue why FrameNet is better suited for our needs. We then present FrameNet's methodology in some detail.

3.3.3. PropBank

In the PropBank ("Proposition Bank"), Kingsbury et al. (2002) provide a manual annotation of all verb instances of the Penn Tree Bank corpus (Wall Street Journal, 1 M words) with semantic role information. PropBank is based on the observation that classes of verbs can occur in different pairs of syntactic frames, which are in a way meaning-preserving (diathesis alternations, e.g., Levin, 1993; Kipper, Dang, and Palmer, 2000). The underlying assumption is that such syntactic alternations are interesting because they reflect semantic properties of the verbs. The aim of PropBank is to provide a semantic annotation which enables studies of such alternation phenomena (Palmer et al., 2005).

Structure of PropBank

An important design decision is that PropBank does not attempt to group verbs to classes. Instead, (most) verbs are treated as separate entities. Modeling generalizations over different predicates is not supported. A second characteristic of PropBank is that the role labels, which are chosen for a given class, do not attempt to provide a semantic

buy.01	sell.01	market.01
Arg0: Buyer	Arg0: Seller	Arg0: Seller
Arg1: Thing bought	Arg1: Thing sold	Arg1: Thing Sold
Arg2: Seller	Arg2: Buyer	Arg2: Buyer
Arg3: Price paid	Arg3: Price paid	Arg3: Attribute
Arg4: Benefactive	Arg4: Benefactive	Arg4: Benefactive

Table 3.3.: PropBank rolesets.

characterization or even formalization of the respective roles. In view of the problems related to providing a uniform set of semantic roles, the roles used in the PropBank are defined for single verb senses as explicated in (Palmer et al., 2005, p. 4):

> "An individual verb's semantic arguments are numbered, beginning with 0. For a particular verb, Arg0 is generally the argument exhibiting features of a prototypical Agent (Dowty, 1991) while Arg1 is a prototypical Patient or Theme. No consistent generalizations can be made across verbs for the higher numbered arguments [...]."

Annotation. Example annotations from PropBank for the verb *sell* are given in (3.32) and (3.33). The annotations show how the active/passive alternation is accounted for by using Arg1 consistently for the patient-like "goods role", although *Marginal operations...* is the surface subject of the passive sentence.

(3.32) A company spokesman declined to comment and said that [the officials]$_{\text{ARG0}}$ who [sold]$_{\text{REL}}$ [shares]$_{\text{ARG1}}$ wouldn't comment.

(3.33) [Marginal operations and assets]$_{\text{ARG1}}$ have been [sold]$_{\text{REL}}$.

The annotation is guided by a collection of so-called *rolesets* which define all potential roles that can apply for a given verb sense.

Rolesets. The corresponding roleset for the verb *say* is displayed in Table 3.3 in the middle column. In addition to a providing a list of all possible arguments for a verb sense, mnemonics ("descriptors") are provided for the annotators. However, the descriptors "have no theoretical standing" (Palmer et al., 2005). Apart from these verb-specific roles, a small set of general roles like temporal modifiers can be applied to any verb, as can be seen in (3.34), where roles like ARGM-TMP are annotated.

(3.34) [They]$_{\text{ARG0}}$ [may]$_{\text{ARGM-MOD}}$ [also]$_{\text{ARGM-DIS}}$ [eventually]$_{\text{ARGM-TMP}}$[sell]$_{\text{REL}}$ [the shares]$_{\text{ARG1}}$ [to third parties]$_{\text{ARG2-TO}}$, but the outside investors who own the remaining 60% of Coldwell Banker have the right to first refusal.

In Table 3.3, we contrast the rolesets of *sell* and *market* to given an example for the missing generalization of roles across predicates. The label Arg3 is used for the price of the goods in the case of of *sell*, but for the product category in the case of *market*, as can also be seen in annotated examples like (3.35) and (3.36).

(3.35) They sell the Crimson 710 [for 6.500 EUR]$_{\text{ARG3}}$.

(3.36) They market the Crimson 710 [as a true high-end product]$_{\text{ARG3}}$.

The missing generalization becomes even more obvious comparing the rolesets of *buy* and *sell* in Table 3.3. Almost none of the commonalities between both events are captured – Arg0 always marks the agentive participant, which is the seller in one case and the buyer in the other. Note that the Arg4s also differ, in a situation where John bought a car from Mary, the beneficent roles in (3.37) and (3.38) are different.

(3.37) [John]$_{\text{ARG0}}$ bought [the car]$_{\text{ARG1}}$ [for his mother]$_{\text{ARG4}}$.

(3.38) [Mary]$_{\text{ARG0}}$ sold [the car]$_{\text{ARG1}}$ [for her mother]$_{\text{ARG4}}$.

Usage and Limitations

PropBank currently provides annotations, rolesets, and also syntactic realization patterns (*framesets*) for about 3.300 verbs. So far, it has predominantly been used for its intended application of studying role labeling methods (e.g. Gildea and Hockenmaier, 2003; Pradhan et al., 2005). Due to the (semantically) theory neutral approach, PropBank has rarely been used for tasks which require modeling complex semantic interactions, e.g., for information access or even in multi-lingual context. Exceptions, where PropBank has been used in limited ways, are (Surdeanu, Harabagiu, Williams, and Aarseth, 2003; Qiu, Kan, and Chua, 2006). Moreover, as PropBank does neither group predicates, nor guarantee generalizations of role labels across predicates, it is not obvious how to integrate PropBank with information on a conceptual layer. Although, e.g., WordNet could provide the information that `sell#3` and `market#3` are closely related (direct hypernyms), the problems remain how to represent this information in PropBank's flat structure and how to ensure a sound role inventory.

For modeling textual entailment, it is important to capture relations across different verbs, as the verb used in text and hypothesis are often not identical. Therefore, PropBank's type of analysis only covers some of the data, namely those cases in which the verb is identical. For example, the relation between the passive sentence (3.39) and (3.40) from the RTE-2 corpus can be established using PropBank. The Arg labels can be used to identify the corresponding semantic roles and additional sources can then be accessed to check the compatibility of the role fillers.

(3.39) Before reconstruction began, [the Reichstag]$_{\text{ARG1}}$ was wrapped by [the Bulgarian artist Christo and his wife Jeanne-Claude]$_{\text{ARG0}}$ in 1995, attracting millions of visitors.

(3.40) [Christo]$_{\text{ARG0}}$ wraps [German Reichstag]$_{\text{ARG1}}$

(3.41) [They]$_{\text{ARG0}}$ sell the dvd as a special edition [for 20 \$]$_{\text{ARG3}}$.

(3.42) [They]$_{\text{ARG0}}$ market the dvd [as a special edition]$_{\text{ARG3}}$.

However, cross-predicate variation is a frequent phenomenon in textual entailment, as can be seen in (3.41) entailing (3.42). Arg0 and Arg1 roles should be reliably comparable across verbs, in this case identifying the agent. However, the fact that the Arg3 label is assigned to semantically different roles, could be taken as evidence here that both sentences are not compatible in meaning because *20 $* and *special edition* are incompatible. This is however misleading. Moreover, complex relations like the one between *buy* and *sell* are not represented at all. Neither are variations across different parts of speech such as nominalizations. So, PropBank would probably be of limited applicability for our purpose. Instead, we will use FrameNet as semantic framework.

3.3.4. FrameNet

The *Berkeley FrameNet Project* (Baker et al., 1998) is concerned with the construction of a resource which –like PropBank– provides a role semantic characterization of a core vocabulary of English, exemplified by annotated instances. However, FrameNet's approach differs from that of PropBank in a number of important points. First, FrameNet is based on a semantic paradigm, namely *frame semantics* (see Petruck, 1996, for an overview), which goes back to Fillmore (1976). As one consequence, semantic relations among similar predicates and across complex situations like *buy* and *sell* are represented. Second, FrameNet annotates not only verbs, but also nouns, adjectives, and constructions. Third, instead of annotating one particular corpus, which bears the risk of missing relevant configurations, FrameNet annotation proceeds in a lexicographic, top-down fashion – first, the semantic classes to be annotated are chosen, and then, representative example sentences from different corpora are annotated. We will give a short introduction to Frame Semantics and then describe the current FrameNet Project and the data they provide in detail.

Frame Semantics

While Fillmore is well-known for his work on case grammar, he soon realized a number of problems with this kind of approach (see Section 3.3.2). In (Fillmore, 1982, p.p. 115), he reports how the idea of a new type of semantics developed:

> "[I]t came more and more to seem that another independent level of role structure was needed for the semantic description of verbs in particular limited domains. One possible way [...] [is] deriving sets of truth conditions for a clause from semantic information individually attached to given predicates; but it seemed to me more profitable to believe that there are larger cognitive structures capable of providing a new layer of semantic role notions in terms of which whole domains of vocabulary could be semantically characterized."

Fillmore calls these conceptual structures *frames* and develops the theory of *frame semantics*. Fillmore (1985) compares frame semantics to lexical field theory and argues that most words can only be understood with reference to related words from the same domain, which typically stand in paradigmatic relations (e.g. *Thursday*, *week-end*; *befriedigend* vs. *ausreichend* [German school grades]). Another, related observation is that language learning proceeds "scenario-driven" (Fillmore, 1977). These characteristics of (lexical) semantics are reflected in a double function that the frames fulfill – on the one hand, they bundle related words, on the other hand, they define the events which have to be referred to in order to make sense of the words. In (Lowe, Baker, and Fillmore, 1997, p. 3), frames are described to:

> "have many properties of stereotyped scenarios—situations in which speakers expect certain events to occur and states to obtain. In general, frames encode a certain amount of 'real-world knowledge' in schematized form."

Framing principles. Fillmore (1985) states that frames exist largely independently from language and are related to notions like *schema, script, scenario, ideational scaffolding, cognitive model*, or *folk theory* as known from literature in natural language understanding, cognitive psychology, or artificial intelligence. He sees frames as *tools for the description and explanation of lexical and grammatical meaning*. Semantically opposing words like *up* and *down* typically constitute one frame. The question as to what a frame precisely circumscribes can be a pragmatic decision (Fillmore, 1985, footnote on p. 229):

> "[. . .] sharing semantic content is no guarantee of membership in a single interpretive frame. In my view, such words as *skip, hop, leap*, etc., reflect separate frames, each representing its own schema of pedal locomotion. There is no context-free frame within which these terms occupy different 'slots', though such a frame could easily exist if there arose, for sports purposes, say, a need for stipulating precise distinctions among them."

Important notions adopted from cognitive psychology are *focus* and *background*, which are used to explain, e.g. how *buy* and *sell* describe identical events while *profiling* either the buyer or the seller. Another notion adopted from related fields is that of *prototypicality*. Frames are meant as describing prototypical situations and actual uses may well deviate.

Frames and Frame Elements. Each frame is associated with a set of semantic roles or *frame elements (FEs)* in FrameNet terminology, that represent the participants or propositions involved. Table 3.4 shows the definition of the frame STATEMENT from the database of the FrameNet project with the (core) FEs SPEAKER, MESSAGE, ADDRESSEE, and TOPIC. The frame definition also comprises a natural-language description of the situation which is described, as well as a list of *target words* that potentially *evoke* a frame are given in the form of so-called *lexical units* (LUs), triples of a lemma, a part

Frame: STATEMENT		
	This frame contains verbs and nouns that communicate the act of a SPEAKER to address a MESSAGE to some ADDRESSEE using language. A number of the words can be used performatively, such as *declare* and *insist*.	
Core Frame Elements	SPEAKER	The Speaker is the person who produces the Message (whether spoken or written). It is normally expressed as the External Argument of predicative uses of the target word, or as the Genitive modifier of the noun.
	MESSAGE	The Message is the FE that identifies the content of what the Speaker is communicating to the Addressee. It can be expressed as a clause or as a noun phrase.
	TOPIC	The Topic is the subject matter to which the Message pertains. It is normally expressed as a PP Complement headed by "about", but in some cases it can appear as a direct object.
	MEDIUM	Medium is the physical entity or channel used by the Speaker to transmit the statement.
Lex. Units	acknowledge.v, acknowledgment.n, add.v, address.v, admission.n, admit.v, affirm.v, affirmation.n, allegation.n, allege.v, announce.v, announcement.n, assert.v, assertion.n, attest.v, aver.v, avow.v, avowal.n, ...	

Table 3.4.: A (partial) frame definition.

of speech, and the frame. A feature of frames is that words of different parts of speech, e.g., deverbal nouns and the respective verbs are represented together.

Linking information is also given in the FrameNet database, e.g., that the SPEAKER can be realized as external object or by-PP in the case of *say*. Here, one can see how FrameNet defines a semantic role concept – by referring to linguistic entities and concepts/scenarios at the same time.[5]

Like PropBank roles, Frame elements are local to particular frames (Baker et al., 1998), although their names are sometimes used in multiple frames. As notation, we will sometimes use the frames as prefix to distinguish roles, e.g., we will write GIVING.DONOR and TRANSFER.DONOR to distinguish the DONOR roles of GIVING and TRANSFER. FrameNet distinguishes a number of different FE types, which we will describe below.

An interesting theoretical question is how a frame and its roles relate, e.g., which of both is primary. In fact, both are linked in a "chicken-egg relationship" – a huge part of the frame identity is constituted in the roles it has and the roles themselves

[5] Fillmore (1977) describes these two levels as being a "general representation of all of the essential aspects of events" and "a particular perspective on an event of the type dictated by a case frame", respectively. This distinction is also mirrored in the current FrameNet resource in the distinction between non-lexical frames such as COMMERCE_GOODS_TRANSFER and perspectivized, lexicalized frames COMMERCE_BUY and COMMERCE_SELL, which are annotated with linking information.

are meaningless without the frame and the associated linking information. From a machine perspective, frames and roles alone are a kind of markup that cannot be further interpreted. In line with Fillmore's original theory, it is important to describe for a given sentence why the utterer chose frame X (and LU y) and which roles are realized by which constituents. We will try to elaborate this interplay of a frame and its roles by looking at the task of semantic interpretation, i.e., assigning the correct frame (and roles) to a given verb (and its arguments).

Semantic interpretation. Sometimes, properties of the role fillers can be used to disambiguate the frame and sometimes, intrinsic knowledge of the frames and the given filler is needed to choose the right frame. For example, FrameNet distinguishes the frames REMEMBERING_EXPERIENCE, which relates to procedural memory and REMEMBERING_INFORMATION, which relates to declarative memory. Both frames can be evoked by the verb *remember*.

(3.43) I remember [how I managed to unlock the screw]$_{\text{EXPERIENCE}}$.
(REMEMBERING_EXPERIENCE)

(3.44) I remember [that I managed to unlock the screw]$_{\text{MENTAL\_CONTENT}}$.
(REMEMBERING_INFORMATION)

(3.45) Who remembers [me being a total dork and going up to everyone saying "Hey, I'm Ashley Hunt from Chicago"!?!]$_{\text{EXPERIENCE}}$.
(REMEMBERING_EXPERIENCE)

(3.46) Bill remembers [her as smarter than she is]$_{\text{MENTAL\_CONTENT}}$.
(REMEMBERING_INFORMATION)

Clear indications as to which of both frames fits is given in the distinction between *remember how* and *remember that* as in (3.43) and (3.44). In contrast, it is much harder to distinguish (3.45) and (3.46) (from the FrameNet corpus), where the surface realization of the role fillers do not contain any clues. Here, reference to the frame definition is needed. Finally, the LUs are another important factor in the characterization of a frame. In the running example, e.g., the complex verb *look back* can only evoke REMEMBERING_EXPERIENCE, but not REMEMBERING_INFORMATION.

It really is the interplay between the frame definition, the defined roles, the linking information, and the (grounding in) target words, which characterizes a frame.

Frame inheritance. Not only the words that evoke frames form a "net" structure. The frames themselves can inherit from each other, e.g. to express elaboration between concepts. Multiple frame inheritance ("blending") makes it possible to describe meaning decomposition as has been shown for some example frames in Fillmore and Atkins (1998): A CONVERSATION frame inherits from both RECIPROCITY and TALK. In turn, the frame QUARREL inherits from the CONVERSATION frame and a CONTENTION frame which maps RECIPROCITY into "Opposition". Berkeley FrameNet defines a number of so-called *frame relations*, which are discussed in detail below.

The Berkeley FrameNet Project

The Berkeley FrameNet project has been creating a huge database of frame descriptions and annotated example sentences for an English core vocabulary. In 1997, the Berkeley FrameNet-1 project (Baker et al., 1998) started to implement the central ideas of frame semantics by developing methods and tools for a frame-based description of the syntactic and semantic valency of English words. In this first project phase, a number of *semantic domains* like HEALTH CARE, PERCEPTION, MOTION were selected to guide the definition of new frames – much in the spirit of the original frame semantics. In the ongoing FrameNet-2 project (Fillmore, Wooters, and Baker, 2001), the frames are defined separately without reference to semantic domains. Additionally, a more elaborate way of linking frames via different *frame relations* (see below) has been put to use.

The FrameNet project concentrates on what Fillmore called *cognitive* frames, but does not make any further claims about the cognitive status of frames. FrameNet deliberately aims at a coarse grained level of verb sense distinctions as it has been noted that other resources' distinctions are sometimes overprecise, containing pragmatic and world-knowledge aspects. On the other hand, FrameNet intends to maximally separate the semantic roles per (newly defined) frame first and then to eliminate redundancy at a later stage. The current FrameNet database contains more than 800 frames of general conceptual classes, e.g. AWARENESS, COMMERCIAL_TRANSACTION, or THEFT with more than 10.000 LUs and 135.000 annotated example sentences.

Types of Frame Elements. FrameNet distinguishes between different FE types, which indicates their status within the given frame. FrameNet distinguishes between *core*, *peripheral* (non-core), *extrathematic*, and *core-unexpressed* FEs. They are defined in Ruppenhofer, Ellsworth, Petruck, and Johnson (2006) as follows. A *core* frame element is one that instantiates a conceptually necessary component of a frame, while making the frame unique and different from other frames. *Peripheral* FEs mark such notions as TIME, PLACE, MANNER, MEANS, DEGREE, and the like. They do not uniquely characterize a frame, and can be instantiated in any semantically appropriate frame. Peripheral frame elements do not introduce additional, independent or distinct events from the main reported event. In contrast, *extra-thematic* frame elements situate an event against a background of another state of affairs, either of an actual event or state of the same type, as illustrated with ITERATION, or by evoking a larger frame within which the reported state of affairs is embedded. The value *core-unexpressed* is a special notational shorthand. It is assigned to FEs that behave like core frame elements in the frame where they are marked as core-unexpressed but which, counter to expectation, may not be used for annotation in descendants of that frame. Frame elements marked as core-unexpressed will thus not necessarily be listed among the FEs in descendant frames.

Examples of core (3.47), peripheral (3.48), and extrathematic (3.49) role annotations from the FrameNet corpus are given below.

(3.47) ["I am engaged"]$_{\text{MESSAGE}}$, [she]$_{\text{SPEAKER}}$ announced happily. (STATEMENT)

(3.48) "I am engaged", she announced [happily]$_{\text{MANNER}}$. (STATEMENT)

(3.49) [For many years]$_{\text{ITERATION}}$, he walked to the forum alone. (SELF_MOTION)

Not all frame elements are always overtly realized. They can remain unexpressed like the message in (3.50).

(3.50) [Pat]$_{\text{SPEAKER}}$ already told [me]$_{\text{ADDRESSEE}}$ []$_{\text{MESSAGE}}$. (STATEMENT)

Semantic types. More recently, FrameNet has started to assign so-called *semantic types* to FEs in order to add information like selectional preferences. For example, the semantic type `sentient` is assigned to the SPEAKER of STATEMENT. For the LUs, no further semantic type information such as WordNet senses or the like is available.

Frame relations ("FrameNet hierarchy"). In the current FrameNet database, different types of frame-to-frame relations are specified. The two central frame relations are the *Inheritance* and *Subframe* (formerly: Composition) relation, which are complemented by the *Uses* relation. The relations *Causative_of*, *Inchoative_of*, *Precedes* and *Perspective_on* have been added more recently (Petruck, Fillmore, Baker, Ellsworth, and Ruppenhofer, 2004; Ruppenhofer et al., 2006).

Frame inheritance Frame inheritance is a relation between a parent frame and a child frame where the latter is an elaboration of the former. Any semantic characterization, in particular the role inventory of the parent applies to the child frame as well, possibly being more specific. The child frame can also define additional roles or characteristics. Often, inherited roles are renamed. For example, the frame ARREST inherits the roles AGENT and PATIENT from the frame INTENTIONALLY_AFFECT (renamed into AUTHORITIES and SUSPECT), and adds the roles CHARGES and SUSPENSE.

The Subframe relation The Subframe relation is used to model abstract "scenario frames", such as CRIMINAL PROCESS or EMPLOYMENT. Scenario frames represent complex events with subframe relations holding between the scenario frame and frames that describe (temporally ordered) sub-events. For example, the frame CRIMINAL_PROCESS has the subframes ARRAIGNMENT, ARREST, SENTENCING, and TRIAL. Subframes usually inherit roles from their super frame, e.g. CHARGE and DEFENDANT of ARRAIGNMENT inherit from the respective roles of CRIMINAL_PROCESS.

Precedes This relation can be used to further specify a temporal order on sub-frames in a complex scenario, e.g., BEING_EMPLOYED precedes QUITTING.

The Uses relation The uses relation holds between a specific frame and a more schematic frame that it references. For example, COMMERCE_BUY uses COMMERCE_GOODS-TRANSFER because "An act of buying is not a complete transfer but it cannot be fully understood other than against the background of a goods transfer that is part of a commercial transaction." (Ruppenhofer et al., 2006).

Perspective_on The relations is described in Ruppenhofer et al. (2006) as: "This relation (new in Release 1.3) is a refinement of the more general Using relation [...]. The use of this relation indicates the presence of at least two different points-of-view that can be taken on the neutral frame. For example, the MEASURE_SCENARIO, in which an Entity's VALUE for some ATTRIBUTE is described, can be viewed either from the point-of-view of exact measurement (e.g. "Joey weighed 7 pounds.") or as a relative measure (e.g. "Joey was heavy."). The FEs in the two cases are quite different, so the words should not be included in the same frame, but they do make reference to the same scene."

Causative_of and Inchoative_of The annotation of FrameNet with relations between causative, inchoative and stative frames has only begun. Examples 3.51 through 3.53 evoke frames, which stand in the respective relations.

(3.51) They raised the oil price. (CAUSE_CHANGE_OF_SCALAR_POSITION)

(3.52) The oil price is rising. (CHANGE_POSITION_ON_A_SCALE)

(3.53) The oil price is high. (POSITION_ON_A_SCALE)

We will discuss the ontological status of the different relations forming FrameNet's graph structure in the end of the next section.

Using FrameNet for Natural Language Processing

As FrameNet is a semantic resource which has proven to be largely language-independent (Boas, 2005), a number of projects are investigating the use of FrameNet frames for languages other than English, such as Spanish (Subirats and Petruck, 2003), Japanese (Ohara, Fujii, Ohori, Suzuki, Saito, and Ishizak, 2004), or German in the SALSA project (Erk, Kowalski, Padó, and Pinkal, 2003). So far, most attention has been spent on manual and automatic frame semantic annotation, e.g., SALSA extends the German TIGER treebank (Brants, Dipper, Hansen, Lezius, and Smith, 2002) with frame semantic annotations (Burchardt, Erk, Frank, Kowalski, Pado, and Pinkal, 2006a), and devised automatic annotation methods we will present below.

FrameNet has –until today– been used in applications only in restricted ways, the main reason probably being its limited coverage (see below). For example, in the context of Question Answering, Narayanan and Harabagiu (2004) use it as additional feature if available and Frank et al. (2006) use it in a closed domain.

As for resources like WordNet, the semantic information provided by FrameNet is in the first place provided for human inspection and interpretation. Although most information is also provided in a machine-readable format (the glosses are still plain text), additional effort is needed to automatically access and process it. We will now show what kind of semantic generalizations FrameNet can provide, then turn to the question of automatic frame annotation and finally discuss some issues concerning to the immediate utility of the information.

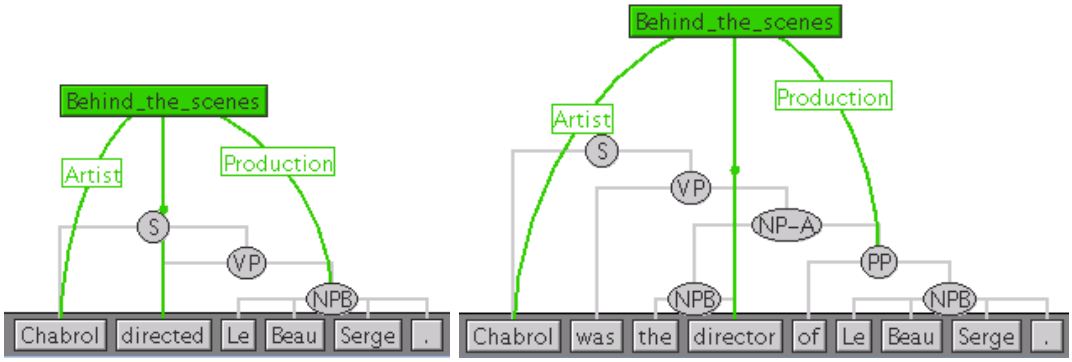


Figure 3.7.: Frame-based normalization over part of speech (*direct* vs. *director*).

Semantic generalization. A frame analysis of a predicate and its arguments captures linguistic variations on a semantic level, which goes beyond the kinds of variations which can be treated, e.g., on grammatical grounds. It is, among others, possible to provide a normalization over different parts of speech as illustrated in in Figure 3.7. As a matter of course, both sentences displayed get an identical frame semantic analysis. The verb *direct* in the left sentence evokes the frame BEHIND_THE_SCENES which talks about the film business. The PRODUCTION role has been assigned to the noun phrase *Le Beau Serge*, the ARTIST role to the prepositional phrase *by Chabrol*. The nominal formulation on the right receives an identical frame analysis.[6] In fact, this sentence pair was taken from a textual entailment corpus (slightly shortened). Entailment can be confirmed here straightforwardly using frame semantics and the "built-in" knowledge about prototypical situations.

(3.54) through (3.57) illustrate how frame analysis provides a normalization over a number of linguistic alternatives of formulating the same content – varying lexicalization, parts of speech, and voice. The sentences are taken from the WWW and talk about the same event – the takeover of the Rover company by BMW. They all involve the frame COMMERCE_BUY, which is displayed in Table 3.5.

(3.54) [BMW]$_{\text{BUYER}}$ bought [Rover]$_{\text{GOODS}}$ [from British Aerospace]$_{\text{SELLER}}$.

(3.55) [Rover]$_{\text{GOODS}}$ was bought [by BMW, which financed the new Range Rover]$_{\text{BUYER}}$.

(3.56) [BMW's]$_{\text{BUYER}}$ purchase [of Rover]$_{\text{GOODS}}$ [for $1.2 billion]$_{\text{MONEY}}$ was a good move.

[6] The question as to whether the preposition *of* is included in the PRODUCTION role filler or not is a matter of taste as the semantic head is *Le Beau Serge*. However, for deeper semantic analysis, as well as for building syntax-semantics interfaces and for lexicographic research, this information might be relevant.

Frame: COMMERCE_BUY		
	These are words describing a basic commercial transaction involving a buyer and a seller exchanging money and goods, taking the perspective of the buyer. The words vary individually in the patterns of frame element realization they allow. For example, the typical pattern for the verb BUY: BUYER buys GOODS from SELLER for MONEY. Abby bought a car from Robin for $5,000.	
Core FEs	BUYER	The Buyer wants the Goods and offers Money to a Seller in exchange for them.
	GOODS	The FE Goods is anything (including labor or time, for example) which is exchanged for Money in a transaction.
Periph. FEs	SELLER	The Seller has possession of the Goods and exchanges them for Money from a Buyer.
	MONEY	Money is the thing given in exchange for Goods in a transaction.
	RATE	In some cases, price or payment is described per unit of Goods.
	⋮	
LUs	buy.v, purchase.v, purchase_((act)).n ...	

Table 3.5.: Definition of COMMERCE_BUY.

(3.57) [BMW, which acquired [Rover]$_{\text{GOODS}}$ in 1994]$_{\text{BUYER}}$, is now dismantling the company.

By the way, the last example, (3.57), also shows *nested* frame elements, where the GOODS role is embedded within the BUYER role, as is the frame-evoking target word. With frame semantics, it is also possible to generate nested frame structures, e.g., the MESSAGE of a STATEMENT frame typically contains frame-evoking words itself. Likewise, it is possible to have frame roles which point to words within other (frame's) roles as in the example above. By now, the potential of frame (element) nesting has been addressed, e.g., in Burchardt et al. (2005b) and Pado and Erk (2005), but it has not been intensively studied so far. We will show some examples of how it can be used in inferences for discourse analysis in Chapter 7.1.2.

Automatic Frame Assignment

In this section we will discuss the automation of frame and role assignment. Typically, the assignment of frames and the assignment of roles are seen as two separate, subsequent tasks. We will discuss both in turn.

Frame assignment. Frame assignment can be seen as a word sense disambiguation task comparable to the task of assigning WordNet synsets (see Erk, 2005). A target expression may be listed as a lexical unit of several frames. Each of these frames can be

seen as a sense of the target expression. For example, the verb *skim.v* is listed as a lexical unit for four frames (all examples are from the FrameNet corpus, partly abbreviated):

(3.58) READING: Skimming *a chapter for its main idea may be done over coffee.*

(3.59) REMOVING: *Remove the vanilla pod,* skim *the jam, and let it cool.*

(3.60) SCRUTINY: *She* skimmed *through the newspaper clippings.*

(3.61) SELF_MOTION: *We* skimmed *across the surface of that sodding lake whilst all around us gathered the dark hosts of hell.*

The frames READING, REMOVING and SELF_MOTION constitute clearly distinguished senses of *skim.v*. READING and SCRUTINY are somewhat harder to distinguish as far as *skim.v* is concerned, even though they describe different situations in general. In this example, the number of senses listed for the verb in FrameNet is comparable in number to the sense distinctions made in WordNet. However, as we will detail in Chapter 4, automatic frame assignment is a big issue in cases where FrameNet's coverage is not that good.

Role assignment. Role assignment is probably the easier of the two tasks. The challenge is to identify and model the syntactic and possibly semantic clues that are needed to assign role labels to the constituents which fill argument positions of a given frame target. So far, much research has been conducted in this direction (e.g. Gildea and Jurafsky, 2002; Baldewein, Erk, Pado, and Prescher, 2004), prominently in the context of Senseval-3 (Mihalcea and Edmonds, 2004).

Available tools. A system, which combines both automatic frame and role assignment is the Shalmaneser system by (Erk and Pado, 2006). The performance of shalmaneser as reported in Erk and Pado (2006) is quite good, exceeding 0.90 accuracy on frame assignment, 0.90 F-Score on role recognition, and 0.80 in role labeling. The system use statistical classification and is available pre-trained on the FrameNet corpus. It is part of a flexible "tool-box" for semantic parsing, which includes an interface to the Collins parser. In fact, Figure 3.7 shows frame semantic analyses generated by Shalmaneser. Shalmaneser outputs SALSA/TIGER-XML, an easy to use format for representing among other things syntactic and frame semantic information (Erk and Pado, 2004).

Limitations of FrameNet

Two main issues which impair the (immediate) usability of FrameNet are its limited coverage and the difficulty of interpreting the different types of frame relations.

Coverage. The FrameNet workflow proceeds in a top-down fashion – one frame is created at a time. First, a frame and its LUs are created and then, representative corpus instances are annotated. As the process is labor-intensive, FrameNet will remain incomplete in several respects for the next future. Three types of coverage problems can be distinguished:

Missing LUs: A frame does not (yet) list all relevant words which can evoke it. For example, *bargain* is not listed in FrameNet at all, although it would fit in one sense in the frame MAKE_AGREEMENT_ON_ACTION. Because of FrameNet's top-down approach, this type of missing coverage should not occur in theory.

Missing frames: This type of missing coverage occurs if there are prototypical situations which are not described by any existing frame. For example, in the current FrameNet release, there is no frame covering the meaning of *strategy* in the sense that an actor has a specific plan or scheme of how to fulfill a goal. Still, there is the frame PROJECT which overlaps the missing frame, and there is the very general frame INTENTIONALLY_AFFECT, the missing frame would probably inherit from.

Missing frame relations In contrast to ontologically organized resources like WordNet, many frames in FrameNet have been defined but not yet included in the inheritance hierarchy. Approximately 40% of the frames in the current database are unique beginners in this sense – although some are related to other frames via relations other than inheritance.

The problem of missing senses is pervasive and is especially challenging for systems being trained for automatic frame assignments. To take up the example from above again, the adjective *strategic* is currently only listed for the frame WEAPON. To illustrate the problem, in Burchardt et al. (2005a), we also found that only 10.7% of the then available 8000 LUs were ambiguous at all, and the baseline for the disambiguation task (assign each LU its most frequent frame) was already at 93% f-score. This is a problem of the FrameNet corpus, not of the FrameNet approach as such, as experiments with a snapshot of the German SALSA corpus have confirmed (Erk, 2005). Also, some frames have no annotated examples, and hence cannot be learned in a supervised learning setting, among them important frames like the POSSESSION frame for *have.v*.

As we will detail in Chapter 4, the coverage problem can be alleviated using WordNet's larger coverage to heuristically assign frames to senses, which would remain uncovered otherwise. We will also touch the somewhat broader issue of how the SUMO knowledge ontology can be used to cover gaps in FrameNet.

Impact of different relation types. Within natural language processing tasks, it is often necessary to measure *semantic similarity* between, e.g., known classes of entities and a new entity occurring in a text. Already in the case of nouns –which often refer to relatively concrete objects– it is not easy to define a comprehensive measure, as can be seen in the vast number of such measures which have been defined for WordNet. A main

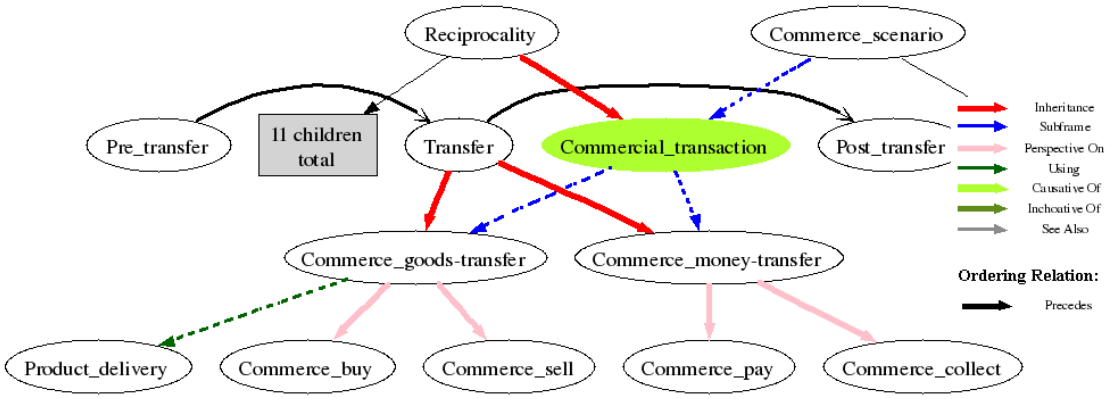


Figure 3.8.: Frame relations of COMMERCIAL_TRANSACTION (screenshot of FrameNet's FrameGrapher).

difficulty is how to delimit concepts or concept clusters in the face of different levels of granularity of sense distinctions, which can be found in given resources.

Modeling *semantic similarity* for events is a bigger challenge, as they form a heterogeneous class. If we look at FrameNet, we find frames that have different ontological status in that they describe, e.g., complex events (COMMERCIAL_TRANSACTION), sub-events (COMMERCE_GOODS-TRANSFER), linguistic perspectives on events (COMMERCE_BUY), or partial aspects of events (RECIPROCALITY). Additional complexity is introduced by the fact that the semantic roles have to be taken into account as well. This overall complexity surfaces in the different types of frame relations, which have to be interpreted in different ways. Without awareness of the impact of different relation types, it is impossible to derive clear similarity judgments. As illustration, consider the two related sentences, (3.62) and (3.63), which instantiate the frames COMMERCE_BUY and COMMERCE_PAY, respectively.

(3.62) [John]$_{\text{BUYER}}$ bought [a car]$_{\text{GOODS}}$ [from Mary]$_{\text{SELLER}}$ [for 10.000$]$_{\text{MONEY}}$.
(COMMERCE_BUY)

(3.63) [John]$_{\text{BUYER}}$ paid [Mary]$_{\text{SELLER}}$ [10.000$]$_{\text{MONEY}}$ [for a car]$_{\text{GOODS}}$.
(COMMERCE_PAY)

As can be seen in Figure 3.8 on the bottom, both frames are connected via perspectivization links up to COMMERCE_GOODS-TRANSFER and COMMERCE_MONEY-TRANSFER, respectively. At this level, there are two possible further (pairs of) links – subframe links to COMMERCIAL_TRANSACTION and inheritance links to TRANSFER. Without going into detail here, depending on which relation is considered, contradictory evidence can be generated concerning the similarity of the sentences. We will come back to this in Section 7.1.

Sumo class	Killing	Death
Axiom	(=> (and (instance ?KILL Killing) (patient ?KILL ?OBJ)) (exists (?DEATH) (and (instance ?DEATH Death) (experiencer ?DEATH ?OBJ) (causes ?KILL ?DEATH))))	(=> (and (instance ?DEATH Death) (experiencer ?DEATH ?AGENT)) (holdsDuring (FutureFn (WhenFn ?DEATH)) (attribute ?AGENT Dead)))
Transliteration	if ?KILL is an instance of killing and ?OBJ is a patient of ?KILL, then there exists ?DEATH so that ?DEATH is an instance of death and ?OBJ experiences ?DEATH and ?KILL causes ?DEATH	if ?DEATH is an instance of death and ?AGENT experiences ?DEATH, then Dead is an attribute of ?AGENT hold during after the time of existence of ?DEATH

Figure 3.9.: SUMO axioms.

3.4. Ontological Resources

In the introduction of this Chapter, we stated that additional ontological knowledge is sometimes needed in the course of natural language processing. After introducing central linguistic resources, we can now substantiate this observation. For example, to establish a relation between the textual entailment pair in (3.64)-(3.65), it is necessary to relate *murder* and *dead*, i.e., to access the knowledge that being murdered results in being dead.

(3.64) John Lennon was **murdered** by a deranged fan in New York City on 8 December 1980 after he returned home from a recording session.

(3.65) John Lennon is **dead**.

However, neither FrameNet nor WordNet provides the necessary information. A frame analysis provides the frames KILLING for *murder* and DEAD_OR_ALIVE for *dead*. But they are not related via any frame relations. In WordNet, *murder* is a hypernym of *kill*, which stands in the *causes* relation to *die*. But no relation exists between *die* and *dead*. As derivationally related forms of *die*, only *dead* as noun and *deadness* are represented.

A provisional solution to this problem of missing ontological knowledge in natural language processing systems is to try and manually add it for re-occurring cases, as, e.g., Bos and Markert (2006) do. However, the coverage that can be achieved this way is limited, and the result might not easily be re-usable. A more principled solution

is to extract the needed information from available knowledge sources like the SUMO ontology (see below). In SUMO, the missing link to relate, e.g., *murder* and *dead* can be found in the combination of two axioms displayed in Figure 3.9. In short, the left axiom states that killing someone causes his death and the right axiom states that one is dead after ones death.

3.4.1. Upper Ontologies

Knowledge is often coded in the form of ontologies. As the amount of knowledge which can be coded is of course immense, a common division is made between *upper (level) ontologies* (top-level ontologies) representing general knowledge which is applicable in many domains and *domain ontologies* representing domain-specific knowledge. While providing an in-depth survey of available ontologies goes beyond the aim of this thesis, a general observation we made is that domain ontologies tend to contain a high proportion of "nominal" concepts (geographical names, types of weapons, biomedical terms, etc.) while upper ontologies contain a much larger proportion of "verbal" concepts (typical situations, events, scenarios). As the task of recognizing textual entailment as specified by the given data does not relate to a specific domain and as we are especially interested in using semantic information on the level of predicate-argument structure, we will concentrate on upper level ontologies.

Ontologies are basically hierarchically organized graph structures. Their definitorial part typically consists of *classes* (concepts) that are connected via *relations* such as `inheritance` or `instantiation`. Other relations such as `causes` in the axiom displayed in Figure 3.9 on the left can provide further, non-hierarchical links between classes. Concrete *facts* are coded as *individuals* instantiating one or more classes and they can be linked to other individuals via the admissible relations. Most available ontologies are coded in some variant of first order logic. Although reasoning in general is undecidable, axiomatic information like the fact that being murdered implies being dead can be integrated in logic-based natural language processing systems as Bos and Markert (2006) have shown.

Yet, within this thesis, we will concentrate on the question of how to *access* ontological knowledge via natural language interfaces in order to enrich the analyses we get using linguistic resources. We will include ontological knowledge in our model for textual entailment and thus provide "ready-made" input for future research on logic-based processing, e.g., using theorem proving.

Accessing Ontological Knowledge

Knowledge ontologies are mostly studied and developed within the knowledge engineering or artificial intelligence community. Consequently, the design and structuring principles are made for representing information on a *conceptual* level rather than on the level of word meaning as linguistic resources do. A fundamental problem limiting the usability of ontologies in natural language processing is that knowledge ontologies often do not provide a link to natural language. In fact, mostly not even annotated (example)

sentences are available. In other words, syntax-semantics interfaces, which provide a link between language and conceptual knowledge, are needed. They are a key requirement for any further research concerning the extraction and application of the knowledge contained in ontologies. We will discuss this issue in detail in Chapter 4. A first step towards a syntax-semantics interface is formed by the mappings from WordNet synsets to ontology classes, which exist for several ontologies, most notably DOLCE, Cyc, and SUMO, which we will present below. Another well-known "linguistically motivated ontology", the Generalized Upper Model (Bateman, Henschel, and Rinaldi, 1995), a descendant of the Penman Upper Model, does not provide a mapping to WordNet.

Dolce

The DOLCE ("Descriptive Ontology for Linguistic and Cognitive Engineering") ontology was developed within the WonderWeb project (Masolo, Borgo, Gangemi, Guarino, Oltramari, and Schneider, 2003). It is formalized in first-order logic with modals and time; 1.000 of WordNet 1.6's noun synsets have been linked to it. The developers call DOLCE a *foundational ontology* and intend it to be a starting point and a reference for a whole library of ontologies for various purposes within a "Semantic Web" scenario. Gangemi, Guarino, Masolo, Oltramari, and Schneider (2002) describe its main purpose as:

> "[...] to negotiate meaning, either for enabling effective cooperation among multiple artificial agents, or for establishing consensus in a mixed society where artificial agents cooperate with human beings."

DOLCE has been used in a study, which detected some inconsistencies in WordNet 1.6. Moreover, some proposals for a re-structuring of the noun taxonomy have been made (Oltramari, Gangemi, Guarino, and Masolo, 2002). However, for the kind of automatic processing we intend, DOLCE seems less suitable. One reason is the level and fine-graininess of the analysis, which may be due to the "cognitive bias" of DOLCE (Masolo et al., 2003):

> "DOLCE has a clear cognitive bias, in the sense that it aims at capturing the ontological categories underlying natural language and human commonsense. We believe that such bias is very important for the Semantic Web (especially if we recognize its intrinsic social nature [...]). We do not commit to a strictly referentialist metaphysics related to the intrinsic nature of the world: rather, the categories we introduce here are thought of as cognitive artifacts ultimately depending on human perception, cultural imprints and social conventions (a sort of "cognitive" metaphysics)."

An illustration of the types of categories DOLCE assumes is given in the analysis (3.66) which is discussed in Gangemi et al. (2002).

(3.66) This rose is red.

The DOLCE analysis of (3.66) involves three individuals (called *particulars* here), reflecting the way in which qualities are modeled in DOLCE, assuming the existence of qualities and quality regions:

1. A Non-agentive Physical Object representing the rose (`rose1`),
2. the color of `rose1`, a Physical Quality (qt_c(rose1)), and
3. the color quale of $\texttt{qt}_c\texttt{(rose1)}$ which is a point in the red color space ($\texttt{ql(qt}_c\texttt{(rose1))} = \texttt{color\#1}$).

Without going into more detail, the type of information provided here is very different from the semantic information we get from the linguistic resources. It is hard to imagine how information on this fundamental level can automatically be integrated with a comparably shallow natural language semantic analysis. Moreover, it is not obvious how a syntax-semantics interface for DOLCE could be designed.

Cyc

The Cyc (from "en**cyc**lopedia") project (Lenat, 1995) was initiated by Doug Lenat in the mid-1980's and as of today is the largest endeavor attempting to formalize everyday common sense knowledge. Cyc used to be an expensive commercial product and only a small part, *OpenCyc*, was freely available which did not suffice to assess the usability of Cyc.

By the time of writing this thesis, a much larger version of OpenCyc has been released, complemented by *ResearchCyc* which includes a database of 300.000 concepts connected by 26.000 relations, and access tools including some sort of natural language parser. Because this is a very recent development, it was not possible to discuss this resource here, but we intend future research in this direction.

SUMO

Among the upper ontologies, SUMO is currently the most promising one. SUMO (Niles and Pease, 2001), the "suggested upper merged ontology" models objects and processes and implicitly provides a coarse characterization of relevant roles for each concept, as well as sortal constraints for roles. Both types of information are provided within *inference rules* (*axioms*). For example, the axiom of the class `Death` displayed in Figure 3.9 on the right constraints the filler of the `experiencer` role (represented by the variable `?AGENT`) to be dead after the event.

SUMO contains about 1000 concepts and is complemented by the "mid-level" ontology MILO (Niles and Terry, 2004) containing another 1500 concepts.[7] For our purposes, SUMO is especially interesting because the conceptual classes for events are "compatible" with FrameNet frames, as they often describe prototypical situations and their participants. SUMO explicitly defines an asymmetric relation type `CaseRole` which can

[7] Henceforth, by SUMO we mean a combination of SUMO and MILO.

Super classes	Sister classes
Entity	Birth
Physical	Breathing
Process	Death
InternalChange	Digesting
BiologicalProcess	Ingesting
PhysiologicProcess	LayingEggs
OrganismProcess	Mating
Death	RecoveringFromIllness
	Replication

Figure 3.10.: Partial SUMO class hierarchy for Death.

be linked to FrameNet roles. The latter point is important as SUMO itself does not provide a full natural language interface. At least, the SUMO concepts are linked to WordNet synsets (Niles and Pease, 2003).

Structure of SUMO. The basic ontology consists of classes hierarchically related by the subclass and instance relations. For example, Figure 3.10 on the left shows the super classes of Death. It might be surprising that Process is a subclass of Physical. The idea behind this classification is that processes are "physical reality" in that they occur at a certain place and time. Figure 3.10 on the right shows the sister concepts of Death, i.e., all subclasses of OrganismProcess. Here, one can observe a certain inhomogeneity in terms of importance of the classes and also a partiality as there are more organism processes one can imagine. A reason for imbalance in SUMO is the fact that it is in fact a blend of a number of freely available ontologies and that it was "constructed with reference to very pragmatic principles" (Niles and Pease, 2001). Imbalance in granularity and coverage limitations, however, are issues one must accept if it comes to modeling large amounts of knowledge.

Both hierarchical relations subclass and instance are instances of BinaryRelation[8]. The class BinaryRelation is also instantiated by various other, non-hierarchical relations like inverse or properPart which connect classes along various dimensions. The relations are further classified as to whether they are transitive, reflexive, and the like. For us, the subclass CaseRole of BinaryRelation is of special interest. This class is instantiated by relations like agent, patient, or destination which seem to correspond to the linguistic concept of semantic role we have discussed in Section 3.3.2. SUMO provides 18 instances of CaseRole, some of which are very general like agent; others are much more specific like invadingVirus or gainsControl. The concept of CaseRole,

[8]More precisely, subclass and instance are instances of BinaryPredicate which is an instance of both BinaryRelation and Predicate. However, the status of Predicate is not documented. We will therefore stick to the more general BinaryRelation.

```
(=>
(and
    (instance ?JOIN JoiningAnOrganization)
    (agent ?JOIN ?AGENT)
    (patient ?JOIN ?ORG)
    (instance ?AGENT Agent)
    (instance ?ORG Organization))
(member ?AGENT ?ORG))
```

Figure 3.11.: SUMO axiom of class `JoiningAnOrganization`.

however, does not cover all relations which correspond to the linguistic concept of semantic role. Some SUMO relations such as `represents`, which links the proposition of class `Communication` to the class `Proposition`, are not included in `CaseRole`, but inherit directly from `BinaryRelation`.

SUMO Axioms. Figure 3.11 shows another example SUMO axiom providing information, which is needed, e.g., to establish the entailment relation in (3.67)-(3.68), namely the information that after joining an organization, one is member of the organization.

(3.67) T: Olson, 62, previously worked as a partner at Ernst & Young LLP, before **joining** the Fed board in 2001, to serve a term ending in 2010.

(3.68) H: Olson is a **member of** the Fed board. – Yes (IE)

The axioms are given in prefix notation. The example is a conditional statement with a conjunction of five conditions in the antecedent, which have to be met for the inference to hold – The event itself, represented by the variable `?JOIN` must be an instance of the class `JoiningAnOrganization` and the agent and patient (represented by the variables `?AGENT` and `?ORG`) must be instances of the classes `Agent`, and `Organization`, respectively. The consequent of the statement expresses that the agent and the patient stand in a `member` relation after the event. SUMO contains roughly 4.000 axioms. The number and usefulness of axioms per class again differs, a problematic example being the axioms for the class `Transfer`. While one axiom states that the agent and patient must not be equal, four remaining axioms relate to different aspects related to blood transfer – information one would rather expect in a domain ontology.

WordNet-SUMO linking. SUMO had originally been linked to WordNet version 1.6 (Niles and Pease, 2003), but the linking has constantly been updated to meet the current WordNet version. Three types of links have been used to characterize different relations of SUMO classes and concepts denoted by WordNet synsets – *synonymy*, *hypernymy*, and

instantiation. In the case of synonymy, both concepts are (almost) identical in meaning like the synset `plant#2` and the class `Plant`. An example for a hypernymy relation is the link between the synset `nunnery#1` and the SUMO class `ReligiousOrganization`, which is broader in meaning, also covering non-christian organizations. An example of instantiation is the synset `brain_death#1` which instantiates the SUMO class `Death` rather than being a subclass.

Using SUMO for natural language processing. So far, only parts of SUMO have been used in natural language processing applications, e.g., Tatu and Moldovan (2006) base their temporal reasoning on the respective SUMO classes, which in turn are based on Allen (1984). However, a prerequisite for using resources like SUMO on a large scale within natural language processing is the availability of proper syntax-semantics interfaces. While the existing WordNet interface makes it possible to assign SUMO classes to predicates, the information which is included in the axioms, i.e., relations between classes and their arguments, as well as sortal information, cannot be accessed in this way. In Chapter 4, we will detail how we can link SUMO and FrameNet in order to use FrameNet as a natural language interface to SUMO.

Part II.

Modeling Textual Entailment

4. Combining Lexical and Ontological Resources

In this chapter, we explore how the coverage issues which impair the usability of FrameNet can be attenuated by combining it with other resources, in particular with WordNet and the SUMO upper ontology.

We will explore the interoperability of the resources as follows. 4.1 will shortly illustrate the coverage issues and give an impression of the potential of interaction between FrameNet and other resources. In Section 4.2, we will show how the large coverage of WordNet can be used to successfully define a Detour to FrameNet that is able to account for missing LUs (incomplete lexicon). To some extent, the problem of missing frames (senses) is also tackled as this approach can assign abstract frames to words for which a specific frame is missing. This problem will be brought up in Section 4.3 again, where we integrate FrameNet and the SUMO knowledge ontology. As a first step, we will discuss the possibility of linking FrameNet and SUMO such that the former can be used as a natural language interface for the latter. In the future, we envisage a standalone natural language interface to SUMO to access background knowledge in general to possibly bridge gaps in FrameNet's coverage in particular.

4.1. Interoperability of FrameNet with Other Resources

The Berkeley FrameNet repository is incomplete in several respects. In Section 3.3.4, we identified missing LUs (incomplete lexicon) and missing frames (senses not covered) as major types of incompleteness. The problem of coverage is manifest. In a study we conducted on the 574 sentences of the RTE-1 development corpus (Dagan et al., 2006), with an average of 16.24 words/sentence, the statistically trained Shalmaneser system yielded an assignment of only 2.7 frames and 3.6 FEs per sentence.

To get an impression of coverage issues and the potential of interaction between FrameNet and other resources, consider Figure 4.1, which provides an analysis of (4.1).

(4.1) John is tasting wine.

The lower part of the figure shows a straightforward frame analysis of the sentence on the basis of the current Berkeley lexicon. The middle shows a frame analysis using the *Detour* approach (see below), based on the WordNet synsets displayed in the upper middle of the figure. The top of the figure shows a SUMO analysis based on the existing WordNet-SUMO linking.

In the FrameNet repository, the verb *taste* is annotated for three frames: Appearance, Perception_experience, and Perception_active. The latter frame describes

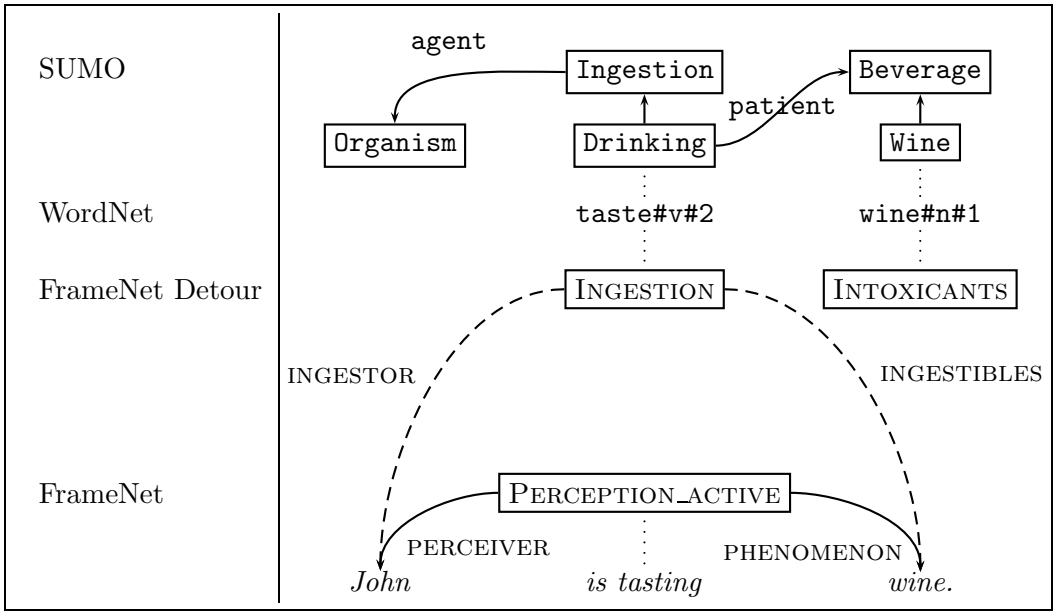


Figure 4.1.: Analyses of (4.1).

"perception words whose perceivers intentionally direct their attention to some entity or phenomenon in order to have a perceptual experience." It covers the given situation, but is still very general. A more specific frame is INGESTION, but it cannot be assigned since the *taste* is not (yet) annotated as LU for this frame. The noun *wine* is not included in the FrameNet lexicon at all. This might also be the case of a missing LU or even of a missing frame.

Both issues can be accounted for by taking a "Detour via WordNet". Given the synsets `taste#v#2` and `wine#n#1`, the frames INGESTION and INTOXICANTS can be assigned (the two second best frames suggested by the Detour for *wine* are FOOD and SUBSTANCE).[1]

The frames INGESTION and INTOXICANTS describe only some aspects of the sentence's core meaning. A SUMO analysis –triggered by the same WordNet synsets as the Detour– provides the additional information that drinking is a kind of ingestion where the "patient" is a beverage and that wine indeed is a beverage. Yet, the figure suggests a link between the `patient` of `Drinking` and `Wine` by conflating different types and tokens to improve readability. Actually, the instances of `Drinking` and `Wine` triggered by the respective synsets are (initially) unrelated. What is missing is an automatic way of arriving at a fully linked SUMO analysis for a given sentence.

Analysis of this simple sentence illustrates is that information which is potentially relevant for inference tasks like checking textual entailment is not provided by one resource

[1] The Detour itself can only suggest frames, the roles have to be assigned otherwise, as indicated by the dashed lines.

or on one level of description alone. Some information is accessible only in their interplay. In the past, some attempts for linking the given resources have been reported. Related work includes a semi-automatic linking of WordNet, VerbNet, and FrameNet reported in Shi and Mihalcea (2005). It treats only verbs and is based on older versions of the resources. Scheffczyk, Pease, and Ellsworth (2006) manually linked FrameNet's semantic types to appropriate SUMO classes, thus focusing on role *fillers*. This is complementary to the task of linking frames and SUMO *classes* we will describe below.

4.2. A WordNet Detour to FrameNet

In this section we present an approach of using WordNet in order to attenuate FrameNet's coverage issues, in particular that of missing LUs. Missing LUs are a severe issue for statistically trained systems, which fail if they encounter a word that is not annotated in FrameNet. The Detour is rule-based and uses WordNet to generalize over a given target word in order to compensate for the missing LU and to assign the appropriate frame. The central algorithm exploits the fact that sense discrimination in WordNet is in general more fine-grained than in FrameNet. So, the assignment task is a often a "many-to-one" problem. In general, there are a number of WordNet-related words available for a given target that are listed in FrameNet's LUs for the intended frame. We first present the main algorithm and then evaluate the implemented Detour system.

4.2.1. The Detour Algorithm

As an illustration how WordNet-based frame assignment works, consider (4.2).

(4.2) Ostriches bury their heads in the sand.

For the target word *bury*, the word sense disambiguation system of Patwardhan et al. (2005) determines the correct WordNet synset `bury#v#3` (*"place in the earth and cover with soil"*). While *bury* is not yet listed in FrameNet's LUs, several WordNet "relatives" of this target are listed as LUs for the frame PLACING (e.g. *lay, put, place*). Using this information, we can access this frame. As there are other frames containing some of the related words (e.g. ATTACK is evoked by *set* and *lay*), a weighting mechanism is used to determine the best fitting frame(s). In the given example, ATTACK is correctly weighted much lower than PLACING by the weighting algorithm of the Detour main algorithm.

The complete Detour algorithm is described below. Pseudo code can be found in Figure 4.2.

1. For a given target word[2], a set of *WordNet relatives* containing all synonyms and hypernyms plus the respective antonyms[3] of these words is computed.

[2] In the following, by *word*, we either mean a word sense, e.g., if WordNet distances are mentioned, or an unambiguated lemma, e.g., as occurring in an LU.

[3] The inclusion of antonyms is effective here because words and their antonyms are typically LUs of the same frame such as *rise* and *fall* for CHANGE_POSITION_ON_A_SCALE.

2. All *candidate frames* evoked by the WordNet relatives are computed. These are all frames that have any of the respective words as LU.

3. In order to select the best frame(s), all candidate frames are weighted according to the function explained below.

For a given target word and a frame F out of the candidate frames, the weighting function basically sums over all of the target word's WordNet relatives that evoke F ($wn\_relative \in Evoking(F)$):

$$Weight(F) = \frac{1}{|Evoking(F)|} * \sum_{wn\_relative \in Evoking(F)} \frac{similarity(wn\_relative, target\_word)}{spreading\_factor(wn\_relative)}$$

Main factors of the weighting function are the *similarity* between the target word and the WordNet relative and what we call the *spreading factor* of the WordNet relative. As a *normalization*, the the number of WordNet relatives evoking the frame ($|Evoking(F)|$) is factored out. Details for these factors are provided below.

Similarity We take the square of the inverse of the WordNet path distance between the target word and the WordNet relative as similarity measure. The longer the path between a WordNet relative and the target word, the lower the similarity and thus the weighting for F. Other similarity measures can easily be plugged in in the implementation as modules from the Perl CPAN archive mentioned in Section 3.3.1 are used.

Spreading factor By *spreading factor* of a WordNet relative we mean the number of frames evoked by that word. It indicates how much evidence the word provides for a frame F under consideration. If there is more than one frame evoked by the word, F only gets the respective share. For example, *go* is listed as LU for three frames (MOTION, COMPATIBILITY, NAME_BEARING) and thus has spreading factor 3. In the weighting of either frame, *go* thus only contributes a share of 1/3.

Normalization The absolute number of WordNet relatives evoking a frame is factored out by a division of the summed weight over the WordNet relatives by the number of relatives ($|Evoking(F)|$). This makes the weights comparable across different runs of the algorithm.

4.2.2. The Detour System

The Detour algorithm has been implemented in the publicly available Detour system.[4] To arrive at a dense frame annotation in applications, we combine Shalmaneser (see Section 3.3.4) and the Detour system. Figure 4.3 shows an automatic Frame assignment for (4.3) combining Shalmaneser and the Detour system.

[4] `search.cpan.org/~reiter/FrameNet-WordNet-Detour/`, an online demonstrator can be accessed at `www.coli.uni-saarland.de/~albu/cgi-bin/FN-Detour.cgi`

1. $WordNet\_relatives = \{w \mid w \in Target\_synset\} \cup \{w \mid \exists\ Synset :$
   $hypernym(Synset, Target\_Synset) \wedge w \in Synset\}$
   $WordNet\_relatives = WordNet\_relatives \cup \{w' \mid$
   $\exists\ w \in WordNet\_relatives \wedge antonym(w', w)\}$
2. forall F in $Frames$
   $Evoking(F) = \{\}$
   end
   forall F in $Frames$, forall W in $WordNet\_relatives$:
   if W is a LU of F then
   $Evoking(F) = Evoking(F) \cup \{W\}$
   $spreading\_factor(W)$ += 1
   end
   end, end
3. forall $Synset, Synset'$:
   $similarity(Synset, Synset') = (\frac{1}{WordNet\_Path\_Distance(Synset, Synset')})^2$
   end
   forall F in $Frames$:
   $Weight(F) = \frac{1}{|Evoking(F)|} * \sum_{wn\_relative \in Evoking(F)} \frac{similarity(wn\_relative, target\_word)}{spreading\_factor(wn\_relative)}$
   end
4. Return frame(s) with highest weight.

Figure 4.2.: Complete Detour algorithm.

(4.3) A July 31 bombing at Hebrew University in Jerusalem killed nine people, including five Americans.

The frame ATTACK has been assigned by the Detour system (indicated by a small flag in Figure 4.3), the frame KILLING and both semantic roles have been assigned by Shalmaneser. This automatic analysis is close to optimal. It would be perfect if the core-role KILLING.CAUSE role had been assigned to the bombing by Shalmaneser.

4.2.3. Evaluation

In this section, we evaluate the Detour system on FrameNet data and against the FrameNet-WordNet verb sense mappings of Shi and Mihalcea (2005).

Experiments on FrameNet data

For proper evaluation of the system, we would need a "realistic" gold standard – a corpus where target words are annotated with their WordNet sense and with correct FrameNet frame(s). Additionally, in order for the Detour to be effective, (many of) these words should be missing in the FrameNet lexicon.

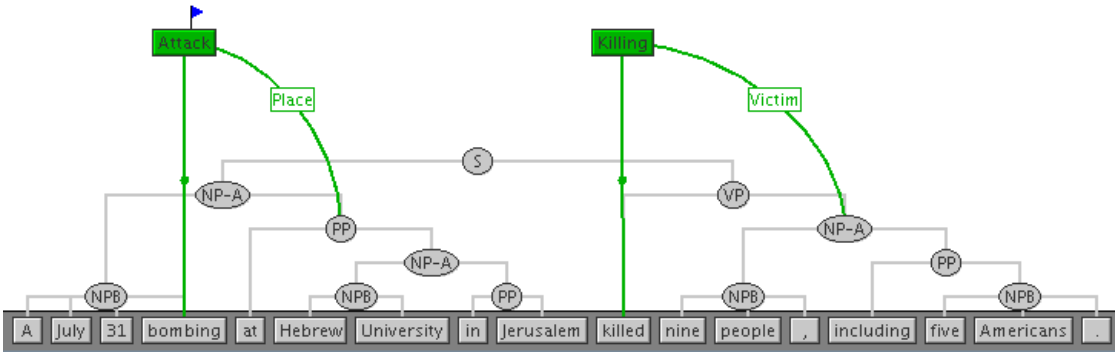


Figure 4.3.: Analysis of (4.3) using the Detour system plus Shalmaneser.

	Frames per synset		
	none	1	>1
Total	13%	71%	16%
Result contains gold standard	-	38%	7%

Table 4.1.: Frame assignment of detour-only system (FrameNet corpus).

As such a resource does not exist, we performed an evaluation using held out examples. To this end, we took the FrameNet example sentences ("FrameNet corpus") as gold-standard and ran the system in a *detour-only* mode that simulates a situation in which the respective target words are missing in FrameNet. This effectively forces the algorithm to take the detour via synonyms, antonyms, and hypernyms. We have evaluated on roughly 80.000 frame-annotated instances (60.000 verb, 20.000 noun, 20.000 adj./adv).

Instead of using an external word sense disambiguation system, we followed the first sense hypothesis and assumed the first, most frequent, WordNet synset for each target word. Two reasons for doing so are first, that using available word sense disambiguation systems did not show a clear effect and second, that the computational complexity of the experiments was reduced drastically.

Results

Table 4.1 shows the system's performance on the FrameNet corpus. The first row shows in how many cases no, one, or more than one frame was assigned, thus measuring recall. The second row indicates whether the gold standard frame is contained in the system result, which is a weak measure of precision.

The detour-only system has high recall – In 87% of the cases, it is able to assign one or more frames to a target word unknown in FrameNet. For the complete data set, the overall precision in terms of containment of the gold standard frame is at 45%. If we

Gold standard	Frames assigned by system	Frequency
MANUFACTURING	INVENTION	19
	INTENTIONALLY_CREATE	12
	BUILDING	11
	CAUSE_TO_START	5
	GETTING	4
	TRANSFORMATION	1

Table 4.2.: Frames assigned by detour-only system.

consider only cases where frames were assigned, precision is at 51%.

Note that in all these cases, statistically trained systems would not be able to assign any frame at all. The Detour successfully assigns the gold standard frame in almost half of the cases. Moreover, the remaining cases are not necessarily severe errors – it is still to be determined how close these assignments are to the gold standard. Since there is no formal measure of frame distance, we inspected sample data manually. Table 4.2 lists the gold standard MANUFACTURING together with frames the Detour suggests. Most of the frames are either from the same domain as the gold standard frame or semantically compatible. Inspection of more examples revealed that the Detour frames are semantically closely related in many cases, often only differing in aspect, perspective or specificity (e.g. CHOOSING vs. DECIDING, AMOUNTING_TO vs. ADDING_UP, or TRAVEL vs. MOTION).

In addition to the detour-only condition, we also tested the "unimpaired" system on the FrameNet corpus. Coverage goes up to 96%. For 83% the gold standard frame is contained in the set of assigned frames (for 67% the gold standard frame is unambiguously assigned). This result indicates that the system introduces a significant amount of noise – one potential error source being the missing word sense disambiguation. But as long as we do not have a formal measure of frame distance, it is hard to determine how often frames which have been assigned instead of the gold standards are true errors. For example, as we have discussed in context of (4.1), the gold standard frames for *taste* are PERCEPTION_ACTIVE, PERCEPTION_EXPERIENCE, and APPEARANCE. But we would not consider it an error if the Detour assigned INGESTION to *taste* in (4.4) from the FrameNet corpus, although this frame is not in the gold standard.

(4.4) Samples of wine were tasted to demonstrate what made a 'good' wine.

Experiments With a Manually Verified Resource

Shi and Mihalcea (2005) provide a manually verified annotation of a sample of 3,094 verbs from FrameNet LUs with WordNet synsets. As this makes it possible to test the Detour without the issue of disambiguation, we also evaluated it against this annotation. Table 4.3 lists the results of running the detour system on the respective verbs, again in

	Gold standard contained in	
	1^{st} result	$1^{st} \cup 2^{nd}$ result
Full system	83%	96%
Detour-only	49%	64%

Table 4.3.: Precision of Detour system compared to (Shi and Mihalcea, 2005).

full mode and in the limited detour-only mode. The table also shows the figures for a relaxed task, where the second best results of the Detour are also taken into account.

Compared to the evaluation the FrameNet data described above, precision is only slightly better (e.g. 49% vs. 45% in detour-only mode). The relaxed task leads to an improvement of precision by roughly 14% in both modes. This latter result indicates the Detour weighting mechanism can still be optimized as many gold-standard frames are contained in the result weighted only second best. Again, there is no measure that allows an analyses as to how much worse higher weighted result are in these cases.

Comparable performance of the Detour on non-disambiguated FrameNet data and on the manually verified data by Shi and Mihalcea (2005) might indicate low quality of the latter. In order to asses the reliability of the manually corrected annotation, we inspected a small random sample of 20 verbs (see Reiter, 2006, for details). In 70% of the cases, the frames heuristically assigned by Detour seemed similar or better than the manually approved frames. Only in 30% of the cases, the frame(s) assigned by the Detour were worse than the one assigned by Shi and Mihalcea (2005). More experiments will be needed before it is possible to draw final conclusions.

Length of the Detour

To asses the effect of the detour-only mode, we compare the full system and the detour-only version on unseen text. We measure the WordNet path distance between the target synset and the synset(s) that finally evokes the frame. This allows an estimation of the "length" of the detour. We distinguish three cases:

Synonymy: The frame evoking synset is the target synset.

Direct hypernymy: The frame evoking synset is a direct hypernym of the target.

Transitive hypernymy: The frame evoking synset is a transitive hypernym of the target.

We ran both system versions on 560 sentences (3.800 instances of nouns, verbs, and adjectives) of the first development set of the RTE-1 (Dagan et al., 2006) data. Table 4.4 gives an overview of the distribution of the distance between the respective target synset and the synset(s) that finally triggered the frame assignment. The effect of switching from full to detour-only mode is a 21% drop of synonym cases. Two thirds of these cases move to the direct hypernym class. So, in the detour-only system, in almost two thirds

	Synonym	Direct hypernym	Transitive hypernym
Full system	54%	18%	27%
Detour-only	33%	31%	35%

Table 4.4.: Distance between frame evoking and target synset (RTE data).

of all cases, the frame evoking synset is a synonym or direct hypernym of the target. For the remaining transitive hypernym cases, we also computed the average WordNet path distance between frame evoking synset and target, which is about 3 for both system modes. As the WordNet verb hierarchy is typically not very deep, the frames we assign in this category are comparably general. For example, for the target `life#n#6` which is the synset for the lifespan of e.g. a battery, we assign the frame QUANTITY, evoked by `measure#n#3`.

4.2.4. Discussion

Different types of missing coverage in FrameNet have different practical impact on the task of assigning a frame to a given lemma. In cases where a suitable frame is available, the Detour can "bridge" missing LUs. This naturally enlarges the choice space for the disambiguation task. For example, the noun *bank* is only annotated for the frame RELATIONAL_NATURAL_FEATURES, i.e., in the river bank meaning. Given the synset `bank#n#2`, the Detour adds the competing frames INSTITUTIONS, ORGANIZATION, BUSINESSES, INTENTIONALLY_CREATE (via *establishment*), and AGGREGATE (via *group*). As we have shown, the Detour weighting algorithm can select the best fitting frame in many cases.

Moreover, the Detour can sometimes (heuristically) account for missing frames by assigning more general or strongly related frames. For example, there is no frame that describes a concept of acting strategically in order to reach some specific goal. The Detour assigns the frame PROJECT to `strategy#n#1` because the related words `plan#n#1`, `program#n#2`, and `scheme#n#1` are listed for this frame. Although this is not a 100% fit, it might suffice for approximate semantic modeling.

A remaining issue is the treatment of cases, where only very few WordNet relatives are available for a target word, or cases, where WordNet relatives evoke very many frames. In both cases, evidence for all frames is typically low. We have conducted first experiments for improving precision by including various thresholds into the weighting algorithm. But as soon as they become effective, a large amount of recall is traded with no clear optimal setting (see Reiter, 2007).

In cases, where there is no fitting frame available in FrameNet for a word (sense) under consideration, the Detour should not suggest any frame (even if the target is annotated for some frame in another reading). This is a problem for the Detour (as is for related systems like Shalmaneser as well). While it should in principle be possible to

devise thresholds that constrain the frame assignment if there is too little evidence, it is hard to find a clear boarder between rare positive cases and errors/noise. It would be helpful to have some control mechanism to correctly reject uncovered targets in advance. First investigations of this problem are reported in Erk (2006) under the heading *outlier detection*.

Future efforts should also be directed to the improvement of external word sense disambiguation systems. In certain cases, close WordNet relatives map to distinct frames. Thus, errors in the initial assignment of synsets immediately affect the quality of frame assignment.

To conclude this section, as frames are general conceptual classes and thus to a large extent language-independent, the method described carries over to other languages that dispose of a counterpart of WordNet and FrameNet LU lists. We have tested our system on the German GermaNet (Hamp and Feldweg, 1997) with German LUs from SALSA. However, as both resources are currently much smaller than WordNet and the English FrameNet lexicon, the results have to be taken with some care.

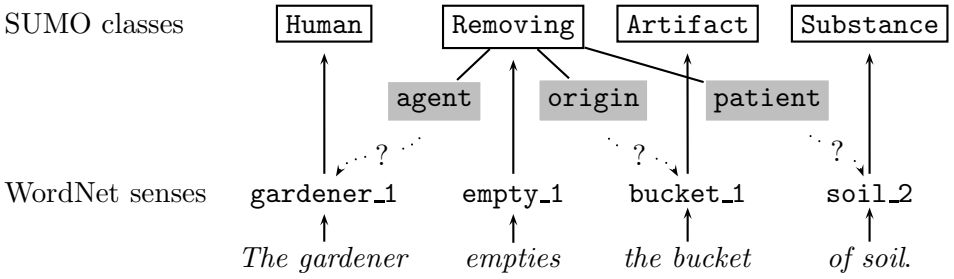


Figure 4.4.: WordNet/SUMO analyses.

4.3. Interfacing FrameNet and SUMO

Throughout this thesis, we approach the phenomenon of textual entailment using natural language analysis mainly on the levels of grammatical description and lexical semantics. As we have argued, for a comprehensive model of textual entailment – like for many other natural language processing tasks – it will at some point be necessary to include additional, ontological knowledge. In this section, we will take SUMO as a typical representative of an upper level knowledge ontology and explore ways of accessing the knowledge contained. For a full integration of the type of knowledge provided by knowledge ontologies, we have to address two subtasks:

Natural language interface As the ontologies are developed in the context of knowledge engineering, they typically do not have a (full) natural language interface. SUMO, e.g., only provides access to its classes via the existing WordNet linking (Niles and Pease, 2003). The axioms cannot be automatically accessed so far. Once we can provide full SUMO analysis of natural language phrases and sentences, we can enrich frame semantic analyses with additional knowledge contained in SUMO. For example, we can heuristically derive "bridging inferences" from the axioms.

Reasoning In order to fully benefit from this knowledge, however, additional state-of-the art reasoning capabilities are needed as the ontological knowledge is mostly given as some variant of predicate logic.

Approaching both tasks would go far beyond the scope of this thesis. In the following we want to approach the first task by showing how FrameNet can be used as an interface to SUMO and how both resources could be linked with manageable manual effort.

```
(=>
 (and
     (instance ?REMOVE Removing)
     (origin ?REMOVE ?PLACE)
     (patient ?REMOVE ?OBJ))
 (and
     (holdsDuring
        (BeginFn
            (WhenFn ?REMOVE))
        (located ?OBJ ?PLACE))
     (holdsDuring
        (EndFn
            (WhenFn ?REMOVE))
        (not
            (located ?OBJ ?PLACE)))))
```

Figure 4.5.: SUMO axiom from class `Removing`.

4.3.1. Accessing SUMO

Figure 4.4 illustrates the issue of accessing the information contained in SUMO for the simple sentence (4.5).

(4.5) The gardener empties the bucket of soil.

What is displayed is the current state of the art – The headwords are mapped to appropriate WordNet senses, which in turn are linked to SUMO concepts, via the available WordNet-SUMO mapping. However, the SUMO axioms for `Removing` refer to participants like `agent` or `patient` and `origin` as, e.g., in the axiom displayed in Figure 4.5. But the axioms cannot be filled with the appropriate arguments since information about the regularities of the mapping between linguistic surface and SUMO arguments is not yet available.

In contrast, FrameNet defines a frame EMPTYING and also provides linking information for the respective roles in the form of syntactic realization and valency patterns as displayed in Table 4.5 for the verb *empty*. From these valency patterns we can read off that the pattern "X empties the Y", where the AGENT is realized as external NP (subject) and the SOURCE as (direct) object NP is the most frequent usage of the verb *empty* in the FrameNet annotations. The pattern instantiated by (4.5) is the second pattern in the table. If we now had the information that the frame EMPTYING is compatible in meaning with the SUMO class `Removing` and that the frame elements AGENT, SOURCE and THEME map to the SUMO relations `agent`, `origin`, and `patient`, respectively, we can build a syntax semantics interface for SUMO. Consequently, a linking of both resources would make it possible to automatically access SUMO axioms as we can use

Number	Agent	Source	Theme
(14)	NP Ext	NP Obj	INI –
(1)	NP Ext	NP Obj	PP[of] Dep
(2)	CNI –	NP Ext	INI –
(1)	CNI –	NP Ext	PP[of] Dep
(1)	CNI –	NP Obj	INI –
(1)	CNI –	NP Obj	PP[of] Dep
(1)	PP[by] Dep	NP Ext	INI –

Table 4.5.: FrameNet valency patterns for the verb *empty* (frame EMPTYING).

available frame assignment methods.

The interface would enable accessing those SUMO classes that have a counterpart in FrameNet. As a next step, one should generalize this to a standalone interface for SUMO, which can then also be used to provide analyses for predicates which are not covered by FrameNet.

Our plan for the rest of this section is to present a semi-automatic approach for linking FrameNet frames and roles to SUMO classes and relations. We will also present preliminary evaluation results derived from an exemplary linking of a sample of frames and roles. However, our focus will more be on exploring the feasibility of this endeavor than on presenting an ultimate result. As it is a natural division, we will treat the linking of frames to classes and roles to relations separately and discuss both subtasks in this order.

4.3.2. Linking Frames and SUMO Classes

Linking frames and SUMO classes is the easier part of linking both resources as we can draw back upon the existing WordNet-SUMO mapping. Still, as FrameNet does not (yet) provide a sense disambiguation for the lemmas in the LUs in terms of WordNet, we are facing a disambiguation problem again. A more fundamental problem is that we do not have any SUMO- (and frame-) annotated gold standard which we could use to evaluate and refine the mapping to be devised. For the time being, we therefore propose a straightforward mapping where disambiguation is based on redundancy. As illustrated in Figure 4.6, we can map a given frame FRAME_X onto one or more SUMO classes (CLASS_X, ...) by taking all possible WordNet synsets for all of the frame's LUs and following the existing WordNet-SUMO mapping. The best class is determined by a majority voting. By design, this method generates one or more SUMO class for all frames that have at least one LU that is contained in WordNet. This applies to most frames (715 out of 795).

In order to obtain a qualitative evaluation and also to address the question of how both resources relate, we automatically generated SUMO classes for 50 randomly selected

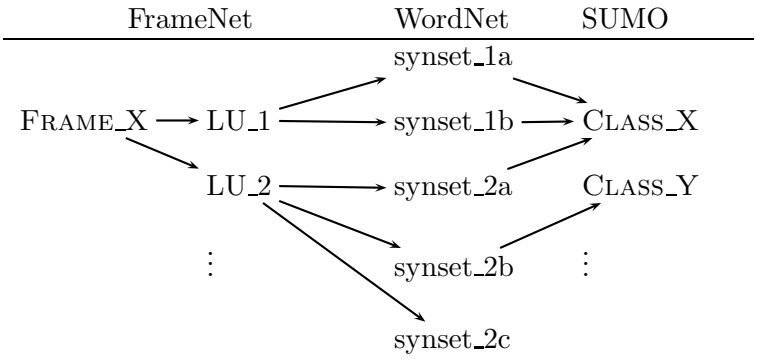


Figure 4.6.: Mapping frames to SUMO classes.

frames and manually checked the result. Table 4.6 lists the results for all frames that have been mapped to SUMO *classes*. The other half of the frames has been mapped to SUMO attributes and relations, which have a different ontological status in SUMO. For the qualitative comparison, we will restrict ourselves to those frames that map to SUMO classes, as they are most similar to frames. We classified the relation between the frame and the assigned class into four categories as can be seen in Table 4.6. The three categories "same range", "frame broader", and "class broader" describe relations where the frame and the class either overlap fully or to a high degree. Almost 80% of the cases fall into these categories. 20% of the cases relate frames and classes that are not (or only loosely) related in meaning. Most of these errors are due to noise or disambiguation problems and can probably be largely eliminated by refinement of the mapping algorithm, e.g., by introducing a weighting mechanism like the one used in the Detour. Some errors like the mapping of the frame Finish_competition to the much more specific class `Game` are harder to prevent. The frame contains a number of LUs like *tie* and *draw*, which are linked to `Game` via the WordNet-SUMO mapping, but both differ on the level of specifity in a way that we consider the mapping inadequate.

Overall, results of automatically mapping frames to SUMO classes are very promising. More research will be needed to find out to what extent the mapping of frames to other SUMO objects such as relations and attributes are plausible (see Reiter, 2006, for a discussion).

4.3.3. Linking Roles and SUMO Relations

The mapping of frame semantic roles to SUMO relations is a "many-to-few" mapping as the number of role-like relations in SUMO is in the magnitude of tenth while the number of frame elements is in the magnitude of thousands. As it would be a huge effort

Frame	Class	Same Range	Frame Broader	Class Broader	Miss
ACTIVITY_ONGOING	IntentionalProcess			x	
	Keeping				x
BECOMING	Process			x	
CALENDRIC_UNIT	TimeInterval	x			
CARDINAL_NUMBERS	Device				x
CHANGE_POSITION_ON_A_SCALE	Increasing		x		
CHOOSING	Selecting	x			
COOKING_CREATION	Cooking	x			
EDUCATION_TEACHING	Communication			x	
	Learning				x
FINISH_COMPETITION	Game				x
	Lost		(Attr.)		
	Won		(Attr.)		
INTENTIONALLY_ACT	IntentionalProcess	x			
INTENTIONALLY_CREATE	ContentDevelopment		x		
	Creation	x			
	Manufacture		x		
JUDGMENT_COMMUNICATION	Communication			x	
LEADERSHIP	Guiding			x	
LOCATIVE_RELATION	Process				x
MOVING_IN_PLACE	Motion			x	
	Rotating		x		
ORIGIN	Nation		x		
PERSONAL_RELATIONSHIP	SocialInteraction	x			
POLITICAL_LOCALES	City		x		
POSITION_ON_A_SCALE	Collection				x
SCRUTINY	Investigating	x			
SELF_MOTION	Walking		x		
STATEMENT	Stating		x		
		7	10	6	6

Table 4.6.: Automatic frame - SUMO class mapping.

to manually map all roles, we suppose to make use of FrameNet's and SUMO's hierarchy to reduce the number of items to be linked. Our idea for a semi-automatic approach is to link both resources only on the top level of the respective hierarchies.

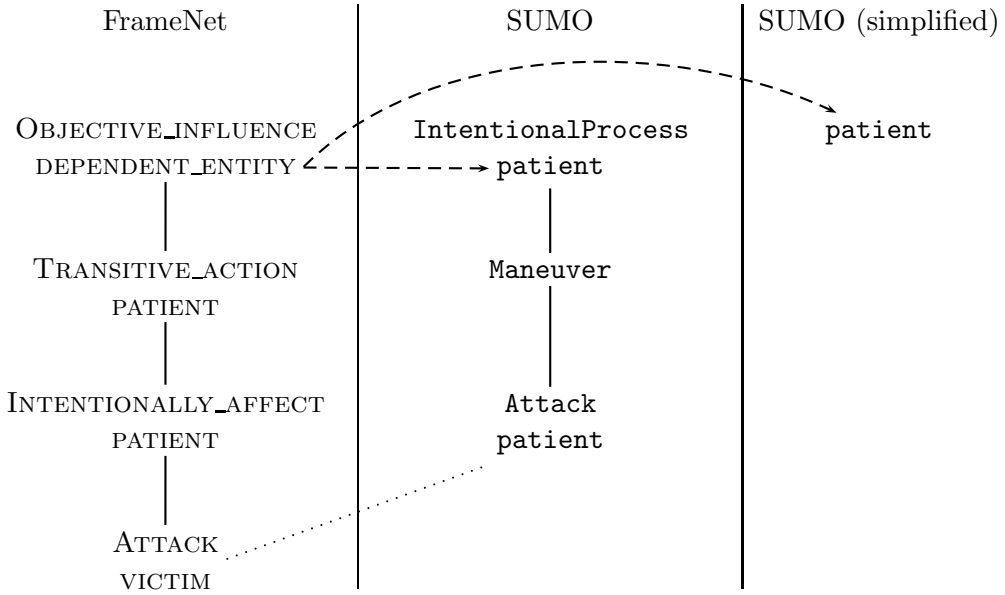


Figure 4.7.: Mapping frame elements to SUMO relations.

If we only link the most general ancestor of a given frame element to a suitable top-level relation in SUMO, it should be possible to induce mappings for frames elements of more specific frames as illustrated in Figure 4.7 (left and middle). After mapping the OBJECTIVE_INFLUENCE.DEPENDENT_ENTITY role to `IntentionalProcess.patient` (lower dashed arrow), we should be able to induce a mapping between more specific roles and relations, e.g., ATTACK.VICTIM and `Attack.patient` (dotted line).

As the relations in SUMO we consider comparable to semantic roles form a small, universal set, which is used at all levels of specifity (see below), we can further simplify the linking effort on the SUMO side by mapping top-level FrameNet roles directly to SUMO relations like `patient` in Figure 4.7 on the right. This patient can then directly be accessed, e.g., by the class `Attack` in a concrete link triggered by the frame ATTACK.

The quantitative dimension of the linking task can be adjusted to the user needs – Linking a considerable portion of roles can be achieved with moderate effort making use of the hierarchy; a complete linking is a larger effort. More concretely, if we restrict ourselves to core frame elements, the current FrameNet database (Release 1.3) contains 1527 *root elements*, i.e., frame elements that do not "inherit" from any other frame element. Figure 4.8 schematically displays a number of frames A ... Z with their respective elements A1 ... Z3. The frame elements printed in boldface are root elements. As we want to gain from inheritance, we will focus on those frames that are hierarchically linked to other frames. In the example displayed in the figure, we would thus disregard

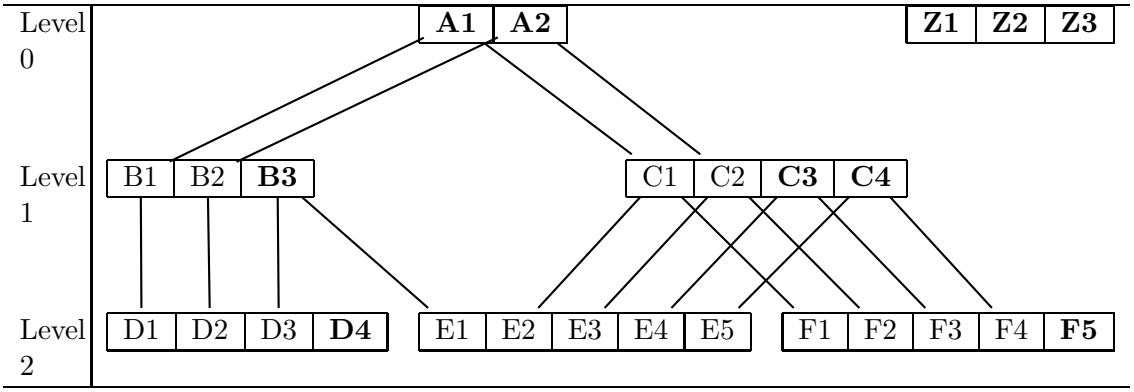


Figure 4.8.: Frame element inheritance.

frame Z for the time being. In FrameNet, 499 of the 795 frames are linked via inheritance or subframe relations, the two hierarchical relations which are relevant in this context. Of these 499 frames, 57 frames are *root frames*, i.e., they do not inherit from another frame, like the frame A in the figure.

If we map the 167 root elements of the root frames to SUMO, we cover 925 core elements and 98 non-core elements via inheritance. In the running example, by linking A1, we cover A1, B1, D1, C1, E2, and F1, for instance. For FrameNet, the ratio of linked elements to covered elements is about 1:6. The core elements covered are roughly 2/3 of all core elements (1472) defined by the 499 linked frames. Here, one can see a huge gain of using the hierarchy. If we continue the linking iteratively on deeper levels of the hierarchy, the ratio of linked to covered elements naturally decreases. In FrameNet, the ratio on the "first inheritance level" (frames B and C in Figure 4.8) is 1:1.6 (307 linked root elements cover 505 elements). Consequently, if a complete coverage of all (core) elements of the hierarchically linked frames or even all frames is intended, more effort is needed.

So far, we have assumed that all frame elements have a counterpart in SUMO. In fact, in a respective study on a set of 142 root frame elements, only 69 could successfully be linked to counterparts in SUMO, 73 remained unlinked. Still, the majority of (inherited) core elements (1.83 of 2.94 on average) had been linked. As one would expect, some frame elements have been linked to multiple relations. For our sample of 50 annotations from the FrameNet corpus we discussed in the previous section, we generated 13 different frame element mappings onto one or more SUMO role. The mappings, which can be found in Table 4.7, are by and large appropriate. The multiple mapping for BECOMING.ENTITY shows the difficulty of deciding for a unique relation. The FrameNet database defines this frame element to be the "Entity which undergoes a change, ending up in the FINAL_STATE or FINAL_CATEGORY". The role `patient` is defined as "a participant [...] that may be moved, said, experienced, etc.", `experiencer` as "[the

Frame Element	Counterpart in SUMO
STATEMENT.MESSAGE	`represents`
STATEMENT.SPEAKER	`agent`
JUDGMENT_COMMUNICATION.EVALUEE	`experiencer`
CHOOSING.COGNIZER	`agent`
INTENTIONALLY_CREATE.CREATED_ENTITY	`patient`
POLITICAL_LOCALES.LOCALE	`located`
SELF_MOTION.GOAL	`destination`
SELF_MOTION.TIME	`time`
SELF_MOTION.SOURCE	`origin`
SELF_MOTION.SELF_MOVER	`agent`, `patient`
BECOMING.ENTITY	`patient`, `experiencer`
EXPECTATION.COGNIZER	`agent`
SCRUTINY.COGNIZER	`agent`

Table 4.7.: Frame elements mapped to SUMO.

experiencer] experiences the Process". From these definitions, one can only speculate about how the usage of the roles is intended precisely.

The reasons why frame elements remain unlinked are heterogeneous. One type of incompatibilities are frame elements which are based on linguistic considerations that do not fit into the ontological categories in SUMO. For example, the frame INTENTIONALLY_ACT has two core roles, AGENT and ACT. The former can be linked to `agent` in SUMO. The latter has no counterpart in SUMO. In FrameNet, it is also often absorbed in more specific descending frames as illustrated in (4.6) and (4.7). The frame INTENTIONALLY_CREATE inherits from INTENTIONALLY_ACT, but it does not provide a role like ACT.

(4.6) [John]$_{\text{AGENT}}$ [produces jazz records]$_{\text{ACT}}$. (INTENTIONALLY_ACT)

(4.7) [John]$_{\text{CREATOR}}$ produces [jazz records]$_{\text{CREATED\_ENTITY}}$. (INTENTIONALLY_CREATE)

Issues related to the role concept in SUMO. Until now, we have treated SUMO relations as more or less direct counterpart of semantic roles in FrameNet. Yet, the "role concept" in SUMO differs, which leads to some problems for the kind of mapping we envisage. In SUMO, the relations we consider compatible with the concept of semantic roles do not have a special status – a common class representing just them cannot be found. Many role-like relations like `agent` and `patient` are instances of the class `CaseRole`, but `CaseRole`'s documentation restricts it to physical objects. The relation `causes`, for instance, is not in the class `CaseRole`, but in `BinaryPredicate`, a superclass of `CaseRole`. `BinaryPredicate`, however, also contains more "technical" relations structuring SUMO.

COMMUNICATION	ContentDevelopment	Communication
COMMUNICATOR	agent	agent
MEDIUM	patient?	patient
MESSAGE	patient?	
TOPIC		

Table 4.8.: SUMO classes related to frame COMMUNICATION.

Moreover, relations are not "defined" on certain classes. Instead, they are simply used within axioms referring to a given classes. As the axioms of mother classes are inherited by subclasses, the roles applicable to a more general class should also apply at more specific classes. However, no information is provided on how axioms and relations behave precisely in the course of inheritance.[5]

In order to get an overview of the "role-like" propertied used in actual SUMO axioms, we semi-automatically extracted candidates from the axioms. It turned out that only a relatively small inventory of 14 relations is used: `agent`, `attends`, `causes`, `destination`, `experiencer`, `instrument`, `located`, `origin`, `patient`, `realization`, `refers`, `represents`, `result` and `subProcess`.

All in all, the SUMO role concept based on a small set of widely applicable relations is somehow comparable with the linguistic concept of "case roles" and in fact the problems that occur here unfortunately confirm this impression. As we will see, the interpretation of, e.g., what a `patient` is in fact differs for different classes, which makes the linking decisions much more difficult. As a problematic example, we want to discuss the issues arising if we want to link the frame COMMUNICATION to any of its two candidate classes `Communication` or `ContentDevelopment`. The frame defines the core elements listed in Table 4.8 on the left. First of all, in both cases, only two of the core roles have suitable direct counterparts in SUMO. The COMMUNICATOR can consistently be mapped to `agent`. Mapping the other roles is more involved. In the case of `ContentDevelopment`, the problem is that the `patient` relation points to the class `ContentBearingObject` which covers both, MEDIUM and MESSAGE. Therefore, deciding for linking one or the other is difficult. In the context of the class `Communication`, the `patient` is clearly used in the sense of a MEDIUM. Here, the issue of consistently capturing a rich and diverse concept of semantic roles using a small set of universals shows up. The issue becomes even more severe if we take inheritance in SUMO into account. The upper part of Figure 4.9 displays a central axiom of the class `Communication`. For better readability, the role relations are displayed as a graph in the lower part of the figure together with some role relations of two (transitive) superclasses `ContentBearingPhysical` and `SocialInteraction`. A first observation is that the message of the communication, which seemed to be missing, is "embedded" under the medium – the `refers` relation of the `patient` points to the mes-

[5] As a consequence, if, e.g., a class like `Vehicle` uses a property `price` and a subclass `Car` also does, both prices' values could be different for a concrete instance of the latter.

```
(subclass Communication ContentBearingProcess)
(subclass ContentBearingProcess ContentBearingPhysical)
(subclass Communication SocialInteraction)

(=>
    (instance ?COMMUNICATE Communication)
    (exists (?PHYS ?ENTITY ?AGENT1 ?AGENT2)
        (and
            (refers ?PHYS ?ENTITY)
            (patient ?COMMUNICATE ?PHYS)
            (instance ?AGENT1 CognitiveAgent)
            (agent ?COMMUNICATE ?AGENT1)
            (instance ?AGENT2 CognitiveAgent)
            (destination ?COMMUNICATE ?AGENT2))))
```

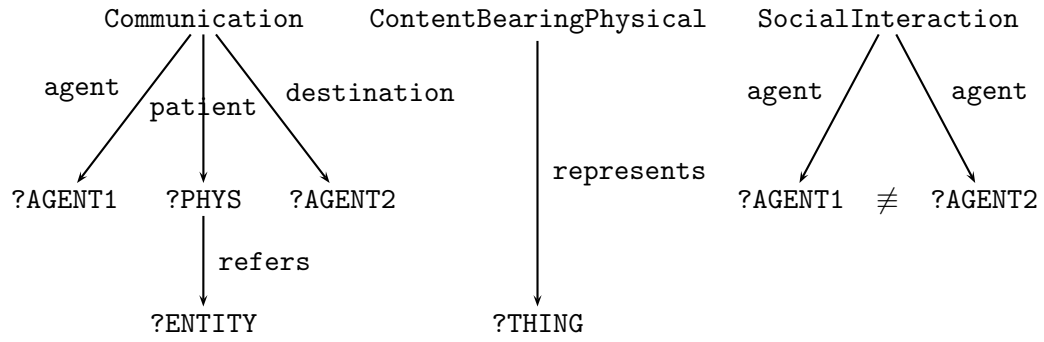


Figure 4.9.: SUMO class `Communication`.

sage. Yet, in the superclass `ContentBearingPhysical`, the message is realized directly by the relation `represents`. The superclass `SocialInteraction` in turn uses two `agent` relations, which should correspond to `agent` and `destination` in `Communication`. It is unclear, however, which agent corresponds to what. Summing up, we encounter a vast number of direct, indirect or inherited role-like relations for the class `Communication` and a stricter modeling is needed to arrive at a consistent overall picture.

4.3.4. Discussion

Our general impression is that the knowledge contained in SUMO and FrameNet bear enough resemblance and offer sufficient connecting factors to be tightly linked. While it should not be too too difficult to access knowledge contained in the SUMO axioms to provide semantic information not contained in FrameNet, some theoretical questions

regarding, e.g., the relation between frames and SUMO attributes have to be further investigated. A wider research question concerns a common representation of knowledge contained in FrameNet and in SUMO. While SUMO is logic-based, where, e.g., variables denote events and objects like `(patient ?BUY ?OBJECT)`, the basic unit in FrameNet are frames, roles, and possibly complex strings as fillers like *the red car in the yard* for the GOODS in a COMMERCE situation, and there is no standard translation into logic available.

Our first results in linking FrameNet and SUMO are promising. The mapping of frames onto classes can be automated. A considerable number of Frame Elements can also be mapped onto SUMO relations with moderate effort. Yet, for now, a more comprehensive linking of roles would have to be conducted "locally", i.e., on the roles a given frame and a given class. So, currently, we cannot fully exploit the possibility of reducing the manual effort by utilizing the hierarchical structures. It would certainly be helpful if not indispensable to annotate a corpus with SUMO classes and relations to be able to study role relations in more detail, in particular if inheritance is involved. This corpus could then also annotated with frames and roles for a contrastive comparison.

4.4. Summary of this Chapter

This chapter has been concerned with interoperability of FrameNet with other resources pursuing the overall goal of alleviating different coverage issues. We presented the Detour system that accounts for missing LUs within existing frames. It successfully manages to assign correct or approximate frames in many cases by making use of WordNet's better coverage. It is used in the SALSA RTE system we present in the next section. We also explored combining FrameNet with the SUMO ontology to use the former as a natural language interface for the latter. A tight linking would allow to access "world knowledge" from SUMO that goes beyond what FrameNet provides. Ultimately, it would also be helpful to derive a standalone SUMO interface to cover gaps in FrameNet's coverage. We have shown that both resources bear enough resemblances for the semi-automatic algorithm we proposed for linking to work. Still, some details have to be resolved, especially within the structure of SUMO, before one arrives at the representations and coverage that would support large-scale automatic systems like the SALSA RTE system. We therefore only include SUMO *classes* in the SALSA RTE system (by way of the existing WordNet mapping) and leave a full integration of SUMO for future work.

5. The SALSA RTE System

In this Chapter, we describe the main contribution of our thesis – an approach to textual entailment, which was implemented in the SALSA RTE system. In previous chapters, we have introduced textual entailment, a promising inference framework to be used in natural language processing applications. We argued that the level of predicate-argument structure is well suited for representing interesting commonalities and divergences between text and hypothesis pairs while being manageable in terms of complexity. FrameNet turned out to be the most promising resource for the given task out of available large resources capturing predicate-argument structure for English. It models variation across predicates in a principled way and also contains some amount of background knowledge – about typical situations, their participants, and relations. Still, FrameNet so far has not been used for the task of modeling entailment (or comparable open domain inference tasks). Coverage issues are probably the main reason. We have explored possibilities of interfacing FrameNet with related resources in order to alleviate coverage problems. The combination with WordNet implemented in the Detour system has proved the most successful.

In this chapter, we will present an approach to textual entailment that is centered around frame semantics, projected from a grammatical analysis. We show how graph-based meaning representations of texts and hypotheses can be constructed that contain information from LFG, FrameNet, and WordNet. The hybrid graphs keep the different information sources separate. Still, interaction between the different layers is exploited for a number of refinements and normalizations. Textual entailment "reasoning" is done in a two-step procedure. First, a graph matching algorithm detects and marks compatible parts in the meaning representations of hypothesis and text. Second, a statistical model is trained on textual entailment corpora to decide entailment. The system design is prepared for future extension. It includes some additional information derived from the SUMO knowledge ontology and also approximates a few "deep" semantic phenomena like negation and modality.

This chapter is structured as follows. In Section 5.1, we will give an overview of the basic architecture of our system. In Section 5.2, we will describe the linguistic analysis of text and hypothesis. Section 5.3 is concerned with the matching algorithm for computing the directed overlap of the hypothesis with the text. Section 5.4 deals with the machine learning architecture.

5.1. Basic Architecture

Textual entailment is typically treated with a combination of information from different sources (see Section 2.6). Our focus is on predicate-argument structure. The linguistic

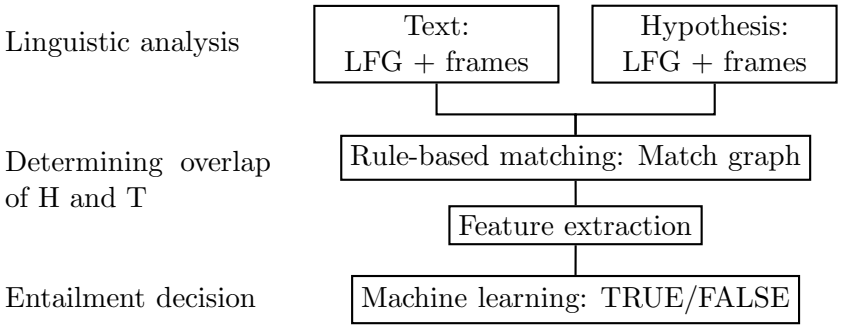


Figure 5.1.: Basic architecture of the SALSA RTE system.

analysis component of the SALSA RTE system is centered around frame semantics (see 3.3.4), which describes predicate-argument structure. The level of grammatical functions serves as basic layer. It accounts for a number of more syntax-near variations, which are indicative for textual entailment (see Section 3.2.2). The respective analyses are provided by the wide-coverage, deep grammatical LFG framework.

Figure 5.1 gives an overview of the concrete modules and architecture of the SALSA RTE system. It instantiates the standard RTE architecture described in Section 2.5.1. The main stages are (i) a linguistic analysis of text and hypothesis, (ii) the computation of the "structural and semantic overlap" of hypothesis and text, and (iii) a statistical entailment decision. In-line with related approaches, we approximate textual entailment by comparing "meaning representations" of hypothesis and text. The basic assumption is that the more of the meaning of the hypothesis is covered by the text, the more probable entailment holds.

In this chapter, we describe the approach and implementation. We will discuss theoretical considerations and practical issues as well as implementation details in one go. Our principal focus is on the question as to whether it is possible to successfully center an architecture for recognizing textual entailment around frame semantic analysis. More precisely, (i) whether it is practically feasible to integrate frame semantic and LFG information in a real system that has sufficient coverage, and (ii) whether it is possible to identify patterns in the LFG/frame analysis that support or reject entailment. With regard to statistical feature modeling, we are interested in gathering insights and observing tendencies rather than striving for optimized performance. In this component, we follow current "best practice" and rely on standard machine learning tools.

5.2. Linguistic Analysis

In this section, we will present the details of the system's linguistic analysis component (see Figure 5.1). Text and hypothesis are both analyzed separately. The linguistic analysis itself can be divided into two main stages, which are displayed in Figure 5.2. Below,

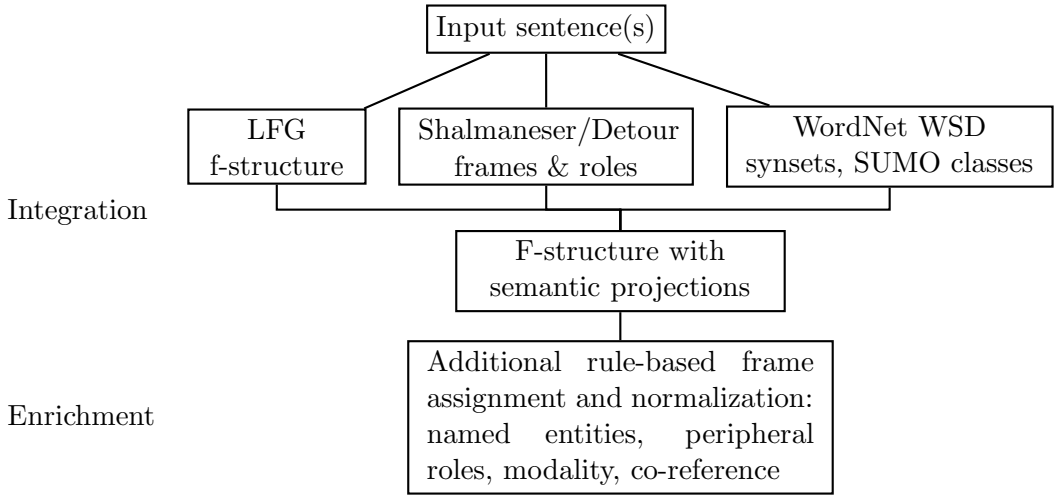


Figure 5.2.: Linguistic analysis.

we will explain these stages in detail. In the first stage (Section 5.2.1), information, which is generated by different sources in isolation (LFG, Detour, Shalmaneser, WSD system), is integrated into one layered representation format. In the second stage (Section 5.2.2), this integrated linguistic analysis is further enriched and normalized. Finally, Section 5.2.3 gives a short overview of the technical realization of the component.

5.2.1. Integration of Linguistic Analysis Components

The three primary types of linguistic analysis are provided by resources we have introduced in Chapter 3 – (i) an LFG analysis by the XLE parser, (ii) a frame semantic analysis by the Shalmaneser and Detour systems, and (iii) WordNet synsets and SUMO classes by components of Patwardhan et al. (2005); Niles (2003). The three analysis tools are run independently and in parallel on the input. The primary output for a given sentence consists of (i) LFG c-structure and f-structure, (ii) Collins parse with frames, and (iii) ontological information in tuple form (word, synset, SUMO class).

These different types of information are only partly compatible. We created an interface, which uses LFG f-structure as backbone and follows the LFG projection strategy to integrate all information into a uniform representation, which keeps the different layers apart.

Within LFG, c-structure and f-structure are linked by a so-called *projection*. This architecture has been extended to a semantic projection, e.g., by Halvorsen and Kaplan (1988); a frame-semantic projection has been proposed by Frank and Erk (2004). In the SALSA RTE system, we implemented a frame semantic projection and generalized this approach by devising a second projection from the frame semantic layer onto the ontological layer. The resulting structure is a tripartite graph like the one shown in

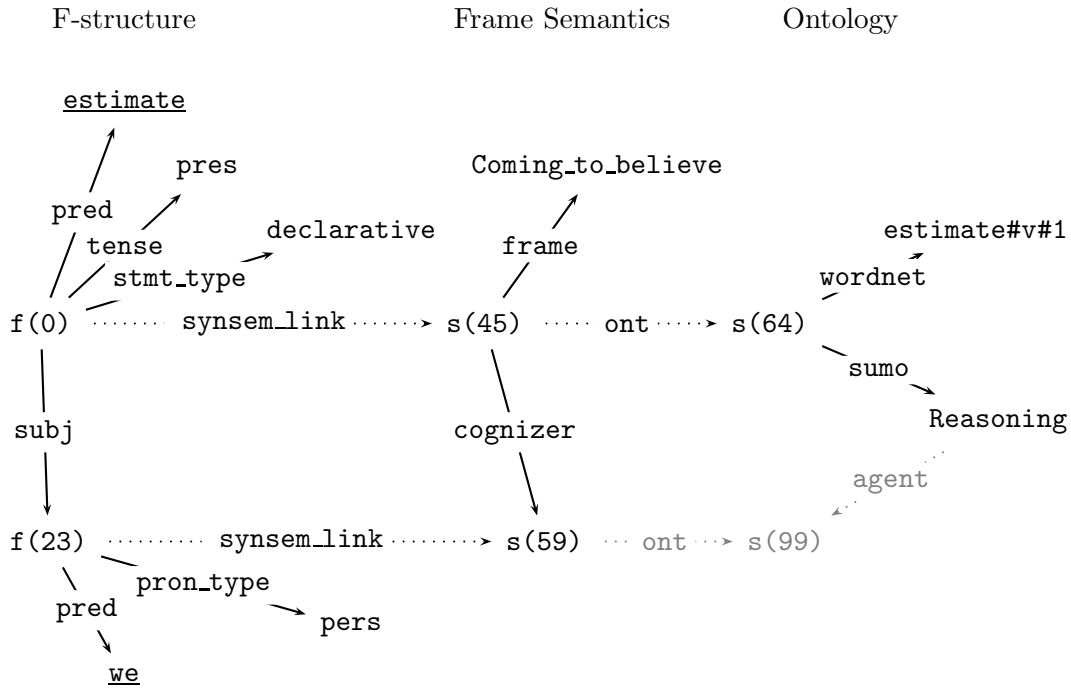


Figure 5.3.: Analysis of *We estimate*.

Figure 5.3 for the small fragment *We estimate* of (5.1).

(5.1) We estimate that the Big Bang happened 20 years ago.

The left-hand side of Figure 5.3 shows a graph representation of (most parts of) an f-structure for the given fragment (the dotted edges can be ignored for the moment). Nodes representing f-structure predicates are labeled with indices of the form `f(n)`. The first two words of (5.1) are represented by the nodes `f(0)` (= predicate `estimate`) and `f(23)` (= predicate `we`). Both are linked via a `subj` edge because that the latter is the grammatical subject of the former. Both predicates are additionally annotated with grammatical information via edges to nodes labeled with atomic values such as `pres` for the tense of the verb.

The dotted edge labeled `synsem_link` represent the projection from f-structure to frame semantics. Semantic nodes are labeled `s(m)`, e.g., the semantic node projected by the main predicate is `s(45)`. Like for the f-structure nodes, features of the semantic nodes are represented as edges to atomic values. For example, `Coming_to_believe` is the `frame` of `s(45)`. Semantic roles are represented as re-entrances like the `cognizer` edge pointing from `s(45)` to `s(59)`, the semantic node projected by `we`. In the case of this pronoun, the semantic projection is "empty" as no (frame) semantic information is available.

The semantic nodes are linked to ontological information via a second projection (`ont`). For example, the ontological node projected by `s(45)` is `s(64)`. It provides the corresponding WordNet synset (`estimate#v#1`) and SUMO class (`Reasoning`). As the design of a reasoning module for SUMO knowledge is beyond the scope of this thesis, an integration of a FrameNet-SUMO role mapping like the one presented in Chapter 4 (the gray part in the figure) is left for future work.

We generate a tripartite graph like the one we have shown above for each text and hypothesis.[1] More formally, each graph consists of a set of nodes *Nodes* and a set of edges *Edges* such that:

$$
\begin{aligned}
&Nodes = G \cup F \cup O \qquad (5.2)\\
&(G = \text{grammatical nodes}, F = \text{frame-nodes}, O = \text{ontological nodes})\\
&Edges = FF \cup SP \cup SR \cup OP \cup OK\\
&FF \subseteq G \times G \text{ (f-structure features)}\\
&SP \subseteq G \times F \text{ (semantic projection)}\\
&SR \subseteq F \times F \text{ (roles)}\\
&OP \subseteq F \times O \text{ (ontological projection)}\\
&OK \subseteq O \times O \text{ (ontological knowledge)}
\end{aligned}
$$

This formalization will be relevant in Section 5.3, when we describe the matching algorithm comparing hypothesis and text.

Interface Design

Interfacing the two central layers of the model, LFG and frame semantics, is difficult and error-prone. The task is to port the frame and role annotation generated by Shalmaneser onto LFG f-structures. As both types of information are linked to different syntactic parses (Collins vs. LFG c-structure), issues range from different tokenization and treatment of abbreviations, numbers, etc. to different syntactic analysis caused by ambiguity or fragmentary parses. Figure 5.4 shows an example of a fragmentary LFG c-structure and the corresponding Collins parse. Shalmaneser assigned the FE EMPLOYER to the prepositional phrase (PP) *for the Les Paul Legacy* of the Collins parse. In the corresponding LFG parse, however, the company name has not been completely recognized. Therefore, a matching prepositional phrase cannot be found in the LFG parse.

A fundamental design decision concerns the interface layer used for exchange between the different resources. Three natural choices are given below.

Syntactic constituents A straightforward idea is to induce a mapping of corresponding syntactic constituents of both parses. This works fine if the parses are by and large

[1] While the basic linguistic components work sentence-based, it is straightforward to extend this type of representation to multi-sentence fragments. In the system implementation, the respective graphs representing single-sentences are tied together with a new top node. We will discuss further potential of frame semantics for extending semantic analysis across sentence boundaries in Chapter 7.

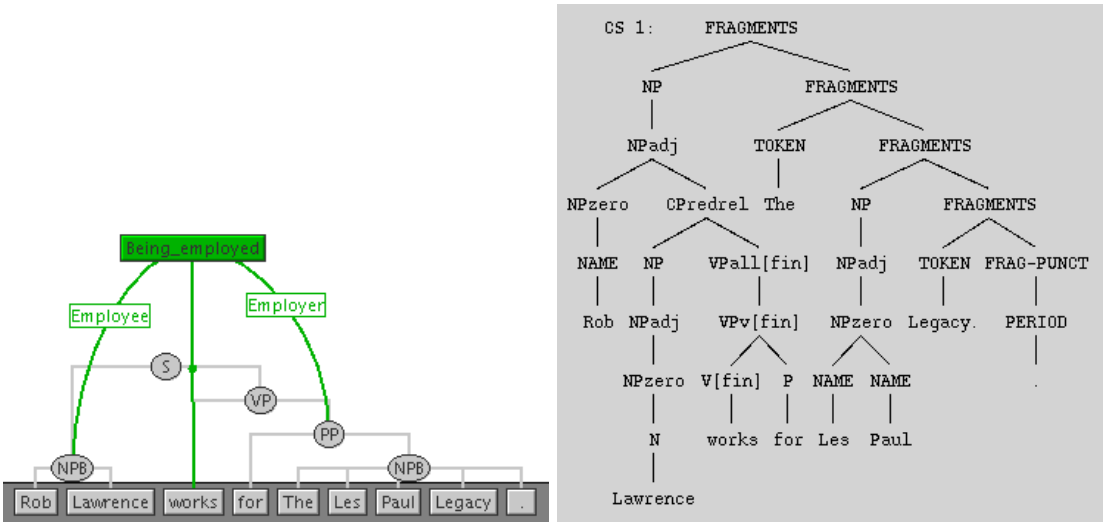


Figure 5.4.: Collins parse and fragmentary LFG c-structure.

the same, i.e., (i) both parsers exhibit the same level of granularity and (ii) both parsers analyze actual sentences similarly. Neither of these conditions is met if we compare the Collins and LFG parser. The c-structures are more fine-grained than the relatively flat Collins parses. Moreover, differences in parse results are found relatively frequently. Above, we have already illustrated a problem that can occur in case of a fragmentary parse. Comparable problems can occur, e.g., if ambiguities are resolved differently.

Head words/lemmas Another option is to use the head words of the constituents that realize frames and roles in the Collins parse to find corresponding words (predicates) in the LFG c-structures (f-structures). While this is robust with respect to different parses, information about the *exact* yield of the original constituents can get lost.

Surface spans As parsers typically provide span information, it is also possible to use this as interface layer. One problem is that different parsers exhibit systematic differences, e.g., regarding the question of how punctuation such as hyphenation is counted. Another problem are actual parse errors, which lower the robustness of this approach.

We have weighted all options against each other. For the given parsers, using head words/lemmas works best. As we focus on the (semantic) head words of role fillers anyway, the danger of "cutting off" peripheral parts of the role fillers' yield is not an issue at all.

To conclude this part, implementing the ontological projection was less difficult. As it is realized as a subsequent step to the semantic projection, most problematic cases have been treated within the semantic projection already. In short, if the semantic projection succeeds, the ontological projection typically also does.

5.2.2. Enrichment and Normalization

In this section, we will be concerned with the second processing step of the linguistic analysis component (see Figure 5.2) – enrichment and normalization of the analyses. Newly extracted or refined information can be represented in two ways:

1. New information can be included by generating, deleting or modifying existing types of representations such as frames, SUMO classes, or LFG f-structure elements. This has the advantage that subsequent processing steps, which treat these structures according to their "standard interpretation", need not be adjusted. An example where we proceed like this is the introduction of new frames based on information from the LFG named entity recognizer.

2. Alternatively, special operators can be introduced in the graph for marking certain phenomena. In order for them to become effective, it is necessary to provide a suitable "operational interpretation" in subsequent reasoning steps. The treatment of modality will follow this scheme – modality markers are introduced on certain nodes in the graph and in the subsequent comparison of hypothesis and text, matching is blocked if modalities are incompatible.

In this section, we will demonstrate a number of enrichments and normalizations of both types. In Chapter 2, we have argued that the levels of linguistic analysis and phenomena observable in textual entailment data that can influence the entailment decision are manifold. A comprehensive treatment of all of them is currently out of reach. In our refinement stage, we focus on special treatment of the topics listed below. Their choice is motivated on the one hand by the existence of respective corpus examples. On the other hand, we chose topics for which suitable information is provided by the given resources.

- Named Entities
- Peripheral roles
- Temporal order of events
- Modality
- Co-reference

This list is incomplete, of course. The architecture is open to and prepared for future extension. Technically, all refinements have been realized using the XLE re-write system.

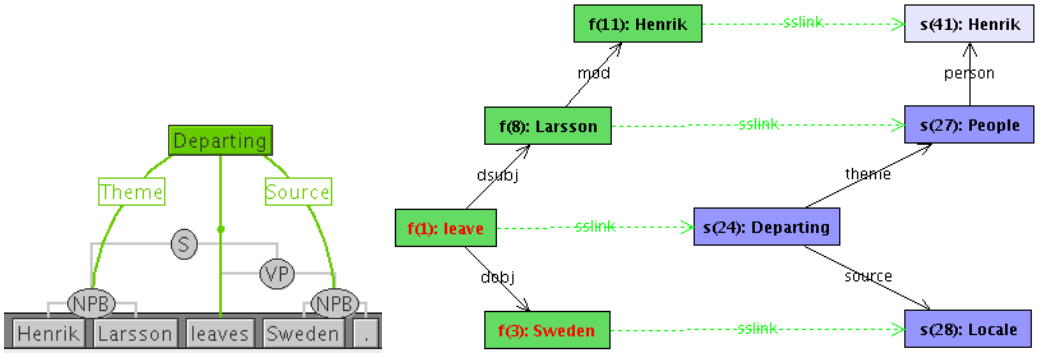


Figure 5.5.: Semantic enrichment: named entities.

For illustration, we will present some of the respective re-write rules where appropriate. The complete re-write grammar we generated consists of roughly 500 lines of code, containing about 45 rules and 20 templates and macros. The XLE system automatically expands this hand-written code into 400 basic re-write instructions.

Named Entities

FrameNet defines a frame PEOPLE, which is evoked by words like *woman* or *fellow* referring to human beings. Proper names are not listed in the LUs for this frame. In order to provide a uniform semantic representation, we assign this frame to individuals referred to by proper names as well. To this end, we use information provided by the LFG parser's named entity recognizer. Figure 5.5 on the left shows an initial frame analysis of (5.3) by Shalmaneser, where *Larsson* is not assigned a frame.

(5.3) Henrik Larsson leaves Sweden.

The LFG named entity recognizer tags the respective predicate as `human`. On the basis of this information, we generate an enriched frame semantic representation. It is displayed in Figure 5.5 on the right. *Larsson* evokes a PEOPLE frame and the name modifier *Henrik* is the filler of the PERSON role. Technically, this is achieved by two rewrite rules. Simplified versions of the rules are displayed below.[2] The rules are triggered when their antecedent, i.e., the part above the arrow (`==>`) is met. A `+` indicates that the fact must be present in the analysis (e.g. `+'S::'(X,SemX)`). In the simplest case, the facts on the right of the arrow are added to the analysis.

```
+human(X,+), nsem_proper_type(X,name)
+'S::'(X,SemX)
==> frame(SemX,'People'), 'Person'(SemX,SemX).
```

[2]Throughout this section, most rules displayed have been simplified for better readability.

This first rule is triggered if an f-structure node `X` has the property `human` and the value `name` is found under a certain path, which is "hidden" in the macro `nsem_proper_type`. In the consequent of the rule, a new PEOPLE frame is assigned to the semantic projection (`S::`) of this f-structure node. A self-referential PERSON role pointing back to the frame is generated, too (not shown in the figure above). This facilitates later access of all constituent parts of the name. A second rule translates all LFG name modifiers into PERSON roles of the frame:

```
+frame(SemX,'People'), +'S::'(X,SemX),
+name_mod(X,Set), +in_set(Y,Set), +'S::'(Y,SemY)
==> 'Person'(SemX,SemY).
```

It is triggered if a PEOPLE frame `SemX` already exists and its respective f-structure node `X` is connected to another node `Y` via a name modifier (`name_mod`) edge. This configuration is found in the example above after the first rule has generated the PEOPLE frame. *Henrik* is analyzed as name modifier of *Larsson* by the LFG. In such a case, a new PERSON role is generated that connects the semantic projections of both nodes (`SemX, SemY`). In the figure, the result is the PERSON role between `s(27)` and `s(41)`.

We designed comparable rules that translate various other types of information provided by the named entity recognizer into frame semantics. For example, in (5.3) above, the information that *Sweden* is a `location` triggers the generation of a LOCALE frame, as can be seen in Figure 5.5 as well.[3] Other examples are the frames BUSINESS, triggered by the category `company`, LEADERSHIP, triggered by `title`, or POLITICAL_LOCALES, triggered by `country` and `city`. The complete grammar snippet for treating named entities is listed in Appendix B.

Peripheral Roles

In Section 3.2.2, we observed that modifiers like temporal adverbials can be equally important for entailment decisions as proper arguments. The same holds for the level of semantic arguments, where peripheral semantic roles can be as important as core roles for the detection of entailment. However, in the FrameNet corpus peripheral roles like TIME are underrepresented, i.e., annotated much less systematic than core roles like AGENT. This observation can be underpinned by comparing the FrameNet annotation with an RTE corpus we manually annotated for evaluation purposes (see Section 6.4.1). Table 5.1 contrasts the distribution of a selected set of frame elements in the FrameNet corpus with the distribution in the RTE corpus (in the table, we have conflated frame elements from different frames like PRACTICE.AGENT and DARING.AGENT into AGENT in order to show the general tendency).

In the RTE annotation, peripheral roles are annotated much more often than in the FrameNet annotation. For example, the PLACE role is annotated about three times as

[3] A normalization, which can also be seen Figure 5.5, is that we provide a uniform grammatical analysis as explained in Section 3.2.2 – passivization is normalized using "deep" subject `dsubj` and "deep" object `dobj` functions.

Role type	Role name	Percentage in FrameNet annotation (322337 roles)	Percentage in RTE annotation (3346 roles)
Core	Agent	4.4	3.6
	Theme	3	2.5
Periph.	Time	1.6	6.1
	Place	0.9	3.2
Extrath.	Internal_cause	0.2	0
	Iteration	0.01	0

Table 5.1.: Distribution of selected roles.

often as in the FrameNet corpus (3.2% vs. 0.9%) while the percentage of core roles such as THEME is comparable (2.5% vs. 3%). As a consequence of this imbalance in the FrameNet corpus, Shalmaneser, which is trained on this data, is not able to assign the peripheral roles as reliably as core roles.

LFG provides a semantic classification of modifiers, which we can use to recover the central peripheral roles TIME and PLACE on grounds of LFG adjunct classes `time` and `loc`. The implementation comprises several rules and rule macros. We will show one of the five respective rules and macros as an example below.

```
(+adjunct(X,Y) | +obl(Z,X) ),
+ptype(Y,sem),
((+psem(Y,PS), +in_set('loc-dir',PS)) | (+psem(Y,PS), +in_set('loc',PS))),
+obj(Y,Z),
+ntype(Z,NT), +nsem(NT,NS), +proper(NS,PR), +proper_type(PR,location),
+'S::'(X,SemX), +'S::'(Y,SemY)
==> location(SemX,SemY).
```

This rule treats cases, where a proper name within a prepositional phrase has been recognized as the name of a location as in *(live) in Tokyo*. In brief, this rule fires if a predicate `X` has an adjunct or oblique argument `Y`, which is a locative preposition like *in* (`|` represents disjunction). Moreover, the object of the preposition (`Z`) has to be classified as a proper name of type `location`. In this case, we add the information that the semantic projection of `Y` (`SemY`) fills the LOCATION role of the semantic projection of `X` (`SemX`).

The manual creation of such rules is relatively time-consuming as it requires insight into a number of LFG-specific details on the one hand and at the same time inspection of a large number of examples is needed in order to cover the most frequent cases. For future development of this or comparable systems, tools supporting the task of defining such rules would be very helpful, possibly with a graphical user interface. Another option could be to try to extract and learn such rule patterns from (manually) annotated data.

A Larger Pattern for RelativeTime

Even though the RTE corpora are assumed not to require reasoning about time, some related approaches implemented special components for modeling, e.g., temporal sequence of events (e.g., Tatu et al., 2006, see Section 2.5.3). While it is not a central aspect in FrameNet either, there is a frame Relative_time, which describes relative order between two events. On this basis, we implemented an analysis of relational time adverbs such as *when* or *during*.

The example also illustrates how a compact rule *template* can translate complete constructions into complex frame configurations. In the code snippet below, the respective rule template (operator '::') is listed together with some of the predicates that trigger it.

```
time_adjunct(TimeAdjunct) ::
+pred(X,TimeAdjunct),
+adjunct(X,Set), +in_set(Y,Set)
+obj(Y,Z),
'S::'(X,SemX), 'S::'(Y,SemY), 'S::'(Z,SemZ)
==> time(SemZ,SemX), frame(SemY,'Relative_timeDEF'),
    'Focal_occasion'(SemY,SemX), 'Landmark_occasion'(SemY,SemZ).

time_adjunct(after).
time_adjunct(as).
time_adjunct(before).
...
```

The result of an application of this template is probably best explained by way of an example. Figure 5.6 shows a partial analysis of (5.4). The frame Relative_time is assigned to *as*, the focal_occasion is *(the news) come(s)* and the landmark_occasion is *(doctors) warn(ed)*. Incidentally, this figure also shows our representation of multiple frame assignment. We assume LFG's way of representing sets, with the predicate `in_set`. This can be seen for *warn*, which is linked to two frames.

(5.4) The news comes as doctors in Hong Kong warned that people who survive Sars may suffer permanent lung damage and may suffer a relapse.

Note that we are still concerned with semantic analysis and refinement. Implementations of temporal *reasoning* would be situated in subsequent processing components. Along these lines, it is possible to model other more complex relations between frames like causation, given that the respective information is either already provided by LFG or can be acquired elsewhere. Ideally, theoretical work on textual entailment would substantiate and inspire work on future extensions of this type.

Modality

While phenomena like quantification or scope ambiguities do not play a central role in the current textual entailment data, modality occurs relatively frequently, but is typically

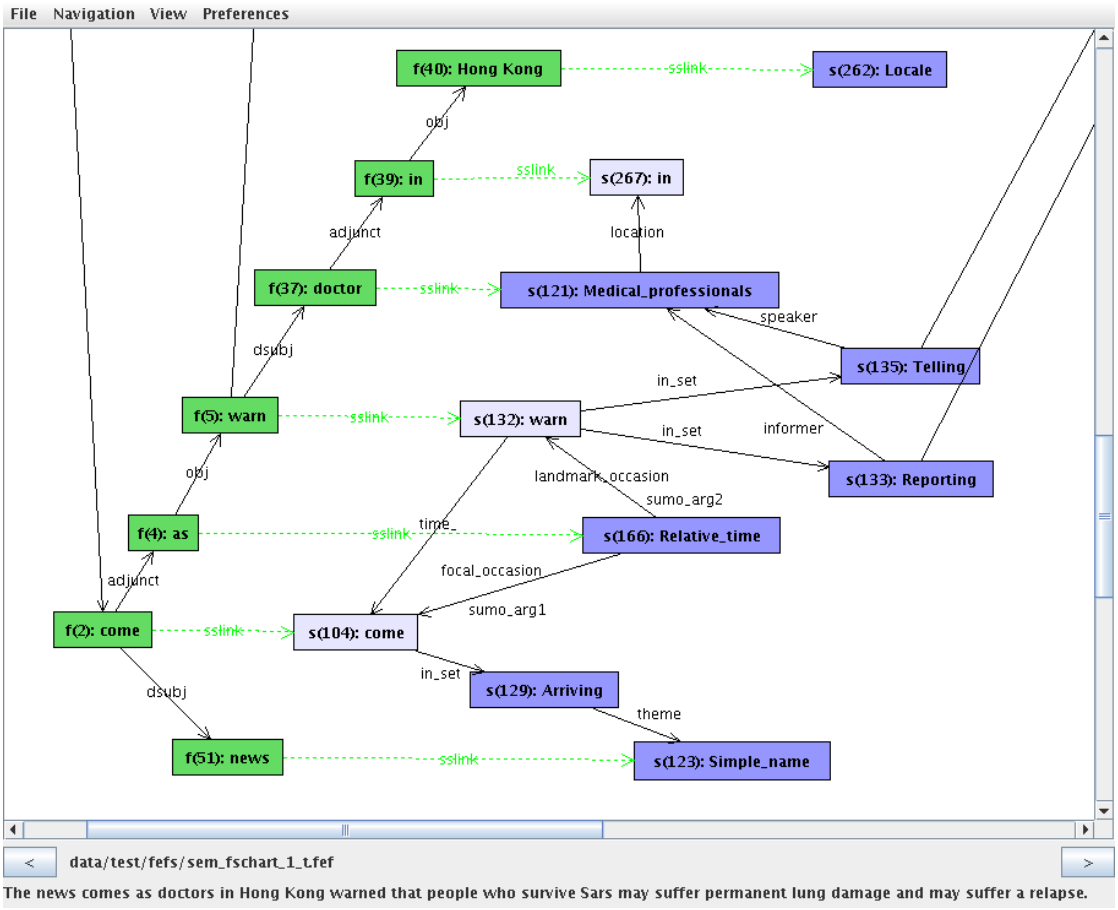


Figure 5.6.: Semantic enrichment: RelativeTime.

neglected by existing systems (Pennacchiotti, 2007). By the time of writing of this thesis, Nairn, Condoravdi, and Karttunen (2006) published an approach for a "shallow", yet comprehensive treatment of modality on the level of dependency trees. In fact, what we had implemented can be seen as a simplified variant hereof. Instead of modeling the interaction between embedded modal operators within a sentence, we only consider one operator at a time and later compare the operators of text and hypothesis. Inspired by the two classical operators of modal logic, we introduce three different modality operators:

`box` This operator represents logical necessity. It is triggered by predicates such as *must*, *have to* or *fut*, the LFG marker for will/shall-future.

`dia` The diamond operator represents logical possibility, triggered by *can*, *may*, or *might*. In conditional *if-then* construction, we also mark both parts as `dia`.

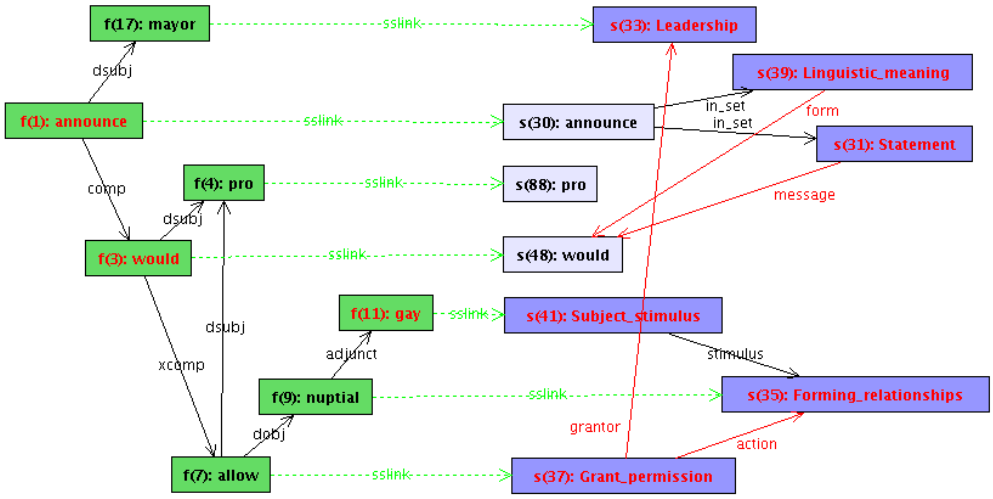


Figure 5.7.: Analysis of example (5.5).

`neg` This operator marks negation.

Every node embedded under a node introducing a modality is marked with the respective label described above. For example, in Figure 5.7, an analysis of (5.5), the modality of type `dia` is introduced by the node `f(3)`. Therefore, nodes `f(4)-f(11)`, as well as the semantic nodes `s(35)-s(88)` are marked `dia`. At the same time, the nodes `f(17)` for *mayor* and the corresponding frame LEADERSHIP are outside the modal scope and thus unmarked. In the comparison of hypothesis and text, we later check whether the modalities of potentially corresponding nodes and edges are compatible (see Section 5.3.7).

(5.5) The mayor announced he would allow gay nuptials.

Formally, we extend the set of nodes of the graphs by elements of $Mod = \{box, dia, neg\}$ and extend the set of edges by edges M from any node to these atomic values ($M \subseteq Nodes \times Mod$).

Co-reference

Co-reference refers to a number of linguistic phenomena that have in common that the referents of certain linguistic expressions are identical. While some tasks like pronoun resolution in general or the identification of nominal chains are intricate, a number of regular phenomena can be treated on the basis of a grammatical analysis. It is, e.g., possible to model appositions like in (5.6), where one entity is referred by different descriptions. Likewise, the antecedent nouns for relative pronouns like in (5.6) can be identified.

(5.6) **US Federal Reserve boss, Alan Greenspan**, sees increased US trade protectionism and ever-larger budget deficits as the biggest threats to the US economy.

(5.7) Scientists have discovered a **gene that** produces a hormone that raises the life expectancy in mice by 30 percent.

LFG marks appositive constructions as `parenthetical`. In the system, this information triggers a respective re-write rule. We then represent these phenomena by introducing a special operator, `coref`, which links the respective f-structure nodes. Thus, the set of edges in the formal graph model is extended by edges Cor connecting LFG f-structure nodes ($Cor \subseteq G \times G$).

Still, it would be beneficial to have a general treatment of pronouns, as well. For the time being, we have implemented a rough heuristics, which marks all preceding nouns of a pronoun as possible antecedents with another special operator. As there has been some recent progress in the performance and availability of broad-coverage pronoun resolution systems (e.g., Morton, 2000; Uryupina, 2007), inclusion of an external pronoun resolution system would be an option for future works.

5.2.3. Implementation of the Linguistic Analysis Component

The linguistic analysis component is for the most part implemented in the XLE rewrite system of Crouch (2005), which is a convenient interface to f-structure information. The architecture is modular such that it is easy to change or replace certain modules. For exchange with other components like Perl Scripts or Prolog programs we need for subsequent computations, we transform the XLE internal format into a simple Prolog-like format we call *FEF* (Frame Exchange Format). This format is not only well suited for manual inspection, it also makes it easy to share data. An example FEF is presented in Appendix C. Incidentally, figures like 5.6 in this section are screenshots from a FEFViewer, we designed.[4]

5.3. Determining Directed Overlap of Hypothesis and Text

After text and hypothesis have both been analyzed linguistically, they are interpreted with respect to their potential entailment relation (see Figure 5.1). This "entailment reasoning" includes among other things an interpretation of the special operators marking modality and co-reference. We take directed overlap of hypothesis and text as entailment measure. It is computed by a graph matching algorithm. Section 5.3.1 introduces the notion of directed overlap as entailment measure. In Section 5.3.2, we present the graph matching strategy and then go through different match types from LFG node matching (Section 5.3.3) to semantic role matching (Section 5.3.6). We finally illustrate how context-dependent inferences on background knowledge can be incorporated into the matching architecture in Section 5.3.8.

[4]Thanks to Alexander Koller, who implemented the FEFViewer.

5.3.1. Directed Overlap as Entailment Measure

Textual entailment is typically approximated via some sort of "semantic similarity" between text and hypothesis (cf. Pennacchiotti, 2007). This is effectively often implemented by computing a variant of *directed overlap* of (meaning) representations of text and hypothesis.[5] The overlap is *directed* for two reasons. First, textual entailment is a directed notion – on a conceptual level, an occurrence of *vehicle* in the hypothesis is covered by an occurrence of *car* in the text, but not vice versa. This directedness immediately carries over to the computation of overlap. Schematically, `(car,vehicle)` would belong to the overlap of a given hypothesis and text, while `(vehicle,car)` would not.

Second, it is decisive how much material of the hypothesis can be related to material in the text. Ideally, one would use meaning representations that establish a linear correlation of overlap and entailment. This would mean that the more material of the hypothesis overlaps with the text, the more probable entailment holds. The only issue would be finding an appropriate threshold for the entailed/non-entailed decision. Yet, the question of how different types of information precisely contribute to the task of detecting textual entailment is largely an open research question. We allow access to all information we have represented on the different analysis layers and use statistical techniques such as feature selection to identify the most relevant factors.

5.3.2. Matching Strategy

We compute overlap between hypothesis and text by detecting matches between nodes and edges. The matching process is implemented in the re-write system we used for the processing of the linguistic analyses, too. The matching result is stored in a third structure we call *match graph*. Matching nodes and edges in text and hypothesis license new nodes and edges in the match graph. In the following, we will distinguish elements of the three graphs by subscripts $text|hyp|match$.

Structure of the match graph. The match graph is a collection of matching nodes, which can be connected by edges. Formally, we represent it by two sets $Nodes_{match}$ and $Edges_{match}$. Each element of $Nodes_{match}$ is a triple of the form $< match_{TYPE}, n_{text}, n_{hyp} >$, which represents the type of match and and a pair of matching nodes – one from the text and one from the hypothesis such as $< match_{frame}, s(23), s(45) >$. These triples are the nodes of the match graph. Their design makes it possible to trace back (i) which are the two original nodes and (ii) what type of matching licensed the node. The edges in $Edges_{match}$ consist of *pairs* of matching nodes. They are also labeled with a match type. The match graph is the union of all different types of matching nodes and matching edges:

[5] As the concept of textual entailment is theory-neutral, some approaches leave implicit what is computed precisely. Formal notions that are used alternatively to *overlap* include *embedding* and *alignment*.

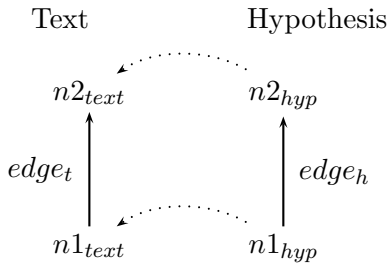


Figure 5.8.: Edge match requires matching nodes.

$Nodes_{match} := NM_{lfg\_pred} \cup NM_{lfg\_coref} \cup NM_{lfg\_anaph} \cup NM_{lfg\_wn} \cup NM_{frame\_id} \cup NM_{frame\_rel} \cup NM_{frame\_detour} \cup \{< dummy, nn, nn >\}$

$Edges_{match} := EM_{lfg\_all} \cup EM_{lfg\_subcgf} \cup EM_{lfg\_modgf} \cup EM_{role\_strict} \cup EM_{role}$

Below, we will present the different match types one by one. Incidentally, the "dummy" node in $Nodes_{match}$ serves as target node for matching FE edges with unmatching fillers (see below).

The matching process. Edges can only match if both source and target node match. This configuration is illustrated in Figure 5.8, where dotted lines represent node matches. For this dependency, the matching process is divided into two subsequent steps – first all node matches are established and then edges are matched. We do not prevent nodes and edges in text and hypothesis from matching multiple other nodes. This matching strategy is declarative in contrast to procedural approaches that implement, e.g., a top-down strategy which embeds the hypothesis into the text using a tree editing algorithm. We understand the match graph as a collection of factors possibly relevant for entailment.

Node and edge matches represent different aspects and degrees of similarity. We classify matches according to their levels of analysis, e.g., grammatical or semantic. Where appropriate, we define different matching conditions, e.g., two frames can match *exactly* if their names are identical, but two frames can also match *heuristically* if they are related via a frame relation. The different match types are represented by different labels of the matches recorded in $Nodes_{match}$ and $Edges_{match}$ such as $< match_{frame-id}, N_{text}, N_{hyp} >$.

All three graphs, the match graph and the original graphs of text and hypothesis are input of the subsequent feature extraction component. Later, we will see that the size of connected sub-graphs in the match graph is a good indicator for entailment. This could be taken as a starting point for the development of "textual entailment algorithms". Below, we will detail the matching process, focusing on central match types related to LFG f-structure, frame semantics, and the treatment of modality.

5.3.3. LFG Node Matching

On LFG f-structure level, we measure the "lexical and structural" similarity of hypothesis and text. To this end, we consider matching nodes (*predicates*) and matching edges (*features*). Node matches of different types are established under the conditions below. Different match types are defined as follows:

Identity If two predicate nodes in text and hypothesis are identical, the matching nodes are included in NM_{lfg_pred} as described formally below:

$< match_{lfg\_pred}, n_{text}, n_{hyp} > \in \ NM_{lfg\_pred}$ if $n_{text} \in G_{text}, n_{hyp} \in G_{hyp}$ and $pred(n_{text}) = pred(n_{hyp})$[6]

Note that description of the graphs of hypothesis and text like G_{text} refer to definition (5.2).

Structural relatedness This covers cases where two nodes in either text or hypothesis are labeled as co-referential in constructions like appositions or relative clauses, as explained above. A node from the respective sentence can be identified to match both the identical phrase and the co-referential one. For example, a node labeled *invasion* in the hypothesis can be matched with the nodes *invasion* and *which* in the text, if they occur in the relative construction *invasion, which was sponsored* [...]. The co-reference can occur either in text or hypothesis:

$< match_{lfg\_coref}, n_{text}, n_{hyp} > \in \ NM_{lfg\_coref}$ if $n_{text} \in G_{text}, n_{hyp} \in G_{hyp}$ and
$(\exists nt \in Nodes_{text} : (coref(nt, n_{text}) \wedge label(nt) = label(n_{hyp}))) \ \vee$
$(\exists nh \in Nodes_{hyp} : (coref(nh, n_{hyp}) \wedge label(nh) = label(n_{text})))$

Anaphoric relatedness This match type considers nodes which are marked as co-referential by the preliminary treatment of anaphora. As in the case of structural relatedness above, a node can match both, the antecedent and the anaphora in the paired sentence. The definition is similar to the one for structural co-reference above, only the match type is NM_{lfg_anaph} here and the predicate tested for is ant_set instead of $coref$.

Semantic relatedness In order to allow some semantic variation, this type of node matching is triggered whenever two dissimilar predicates are related in WordNet with a path length of at most 2 in the hypernym hierarchy. This method flexibly abstracts away over part of speech and particular readings in terms of synsets. It effectively determines the shortest path between two lemmas in WordNet. For example, this allows to relate *use (estrogen)* and *take (estrogen)*.

[6] As the only purpose of this notation is to provide clarity, we keep it semi-formal and leave out definitions of self-explanatory elements like the function *pred*, which returns the predicate values of f-structure nodes.

$< match_{lfg\_wn}, n_{text}, n_{hyp} > \in \ NM_{lfg\_wn}$ if $n_{text} \in G_{text}, n_{hyp} \in G_{hyp}$ and $WN\_distance(pred(n_{text}), pred(n_{hyp})) \leq 2$

In the implementation, WordNet distance ($WN\_distance$) between identical lemmas is set to infinite, thus preventing spurious matches. It is computed in an undirected manner, thus measuring similarity rather than entailment. A future possibility is to experiment with a directed notion.[7]

The latter two match types (anaphoric and WordNet relatedness) are "weaker" than the first two in the sense that they are based on heuristics. This bears the danger that arbitrary matches between nodes of text and hypothesis are established. To avoid over-generation, we constrain these match types in such a way that they only apply if a subsequent edge match is justified. This prevents unrelated nodes from completely different parts of text and hypothesis from matching. So, after edge matches are established, spuriously matching nodes have to be removed. To this end, we define a new set of matching nodes $Nodes'_{match}$, where the unwanted nodes do not occur:

$Nodes'_{match} := Nodes_{match} \ - \ \{< match_{lfg\_wn}, n_{text}, n_{hyp} > \in \ Nodes_{match} |$
$\neg \exists Edge \in Edges_{match} :$
$(from(Edge, < match_{lfg\_wn}, n_{text}, n_{hyp} >) \ \vee \ to(Edge, < match_{lfg\_wn}, n_{text}, n_{hyp} >)\}$

In short, the new definition of matching nodes requires that each node of match type lfg_wn in the match graph has an incoming or outgoing edge. The predicates *to* and $from$ select the match node triples that constitute the endpoints of the edge. Match nodes licensed by anaphoric relatedness are treated analogously.

5.3.4. LFG Edge Matching

Feature (edge) match is triggered only if both the node annotated with the feature and the node representing the value match. We distinguish the following three types of feature matches:

All This category marks all matching edges ("features") from LFG analysis. Edge matching LFG nodes is defined as follows:

$< match_{lfg\_all}, < m1, n1_{text}, n1_{hyp} >, < m2, n2_{text}, n2_{hyp} >> \in EM_{lfg\_all}$ if
$< m1, n1_{text}, n1_{hyp} > \in Nodes_{match}, \ < m2, n2_{text}, n2_{hyp} > \in Nodes_{match}$ and
$\exists edge_t \in FF_{text}, \exists edge_h \in FF_{hyp}$:

[7] Already in the undirected variant, the number of WordNet matches is low, e.g., in the RTE-2 test set containing 800 sentence pairs, we count 85 occurrences (that meet the further condition of licensing a new edge, see below). Features of this low frequency are typically ignored by machine learners. Therefore, it will be necessary to combine this feature with other, related features to become effective in machine learning.

$from(edge_t, n1_{text}) \wedge to(edge_t, n2_{text}) \ \wedge \ from(edge_h, n1_{hyp}) \wedge to(edge_h, n2_{hyp}) \ \wedge \ label(edge_t) = label(edge_h)$

Basically, two matching nodes must be found such that the respective original nodes in text and hypothesis are connected by an edge with identical label (see Figure 5.8 for a graphical display). Note again that description of the graphs of hypothesis and text like FF_{text} refer to definition (5.2).

Subcategorized This match type marks only the governable LFG functions, including normalized grammatical functions `subj, obj, dsubj, dobj`, as well as oblique arguments and complements `obl, comp, xcomp`.

The formal definition is like the one above, but the match type is EM_{lfg_subcgf} and the additional constraint is added that $label(edge_t)$ $(= label(edge_h))$ must be in $\{subj, obj, dsubj, \ldots\}$.

Modifying This marks only the modifying "semantic functions" adjectives, modifiers, and possessive relations (`adj, mod, poss`). Again, the definition is like the one above. The match type is EM_{lfg_modgf} and the constraint is that $label(edge_t)$ $(= label(edge_h))$ must be in $\{adj, mod, poss\}$.

All different match types are represented separately such that it is possible, e.g., to derive interdependencies among them in the statistical model.

5.3.5. Frame Matching

In the simplest case, a frame in the hypothesis is matched with a frame with an identical label in the text. For frame semantic nodes, the matching is defined as:

$< match_{frame\_id}, n_{text}, n_{hyp} > \in \ NM_{frame\_id}$ if $n_{text} \in F_{text}, n_{hyp} \in F_{hyp}$ and $frame(n_{text}) = frame(n_{hyp})$

This applies only in cases, where a frame in the hypothesis and in the text are identical, which is a very strong condition. First of all, according to classical entailment, a frame in a hypothesis might be more general than a frame in the text. For example, "X is walking" entails "X is moving", and indeed, the respective frames SELF_MOTION and MOTION stand in an inheritance relation. This pattern should apply to textual entailment, too. Therefore, we added a match type based on hierarchical information, checking direct inheritance:

$< match_{frame\_rel}, n_{text}, n_{hyp} > \in \ NM_{frame\_rel}$ if $n_{text} \in F_{text}, n_{hyp} \in F_{hyp}$ and $superframe(frame(n_{hyp}), frame(n_{text}))$

Moreover, as different frames often highlight certain meaning components of situations (e.g. Ellsworth, Erk, Kingsbury, and Pado, 2004), it can easily happen that the frames

assigned in text and hypothesis differ slightly while being largely compatible in meaning. Therefore, a more realistic matching condition would also identify highly similar frames as matching. Ideally, the reliability of the match could be calculated from the "semantic distance" of the two frames under consideration.

To date no frame semantic distance measures have been defined. In the absence of a reliable existing frame distance measure, we implemented a straightforward, heuristic assessment of frame distance. Given two frames, the algorithm checks whether one frame is reachable from the other via FrameNet's Inheritance, Subframe, and Using relations. It measures frame similarity in an undirected manner for these three relations. The algorithm also checks whether the two given frames stand in a Causation or Inchoative_of relation like the frames CAUSE_CHANGE_OF_SCALAR_POSITION, CHANGE_OF_SCALAR_POSITION and POSITION_ON_A_SCALE, respectively. This is done in a directed way. The resulting match types are added to the match graph:

$< match_{frame\_detour}, n_{text}, n_{hyp} > \in \ NM_{frame\_detour}$ if $n_{text} \in F_{text}, n_{hyp} \in F_{hyp}$ and
$\neg superframe(frame(n_{hyp}), frame(n_{text})) \ \wedge$
$detourRel(frame(n_{hyp}), frame(n_{text}))$

Example relations from the textual entailment corpora that are derived by this algorithm include, e.g., SPEAK_ON_TOPIC vs. TOPIC, as well as DEAD_OR_ALIVE vs. DEATH. We will come back to the algorithm again in the evaluation of Shalmaneser and Detour in Section 6.4.

5.3.6. Role and Filler Matching

To reduce computational complexity, role matches are established simply by checking identity of role names. Extension of the role matching mechanism to compatible roles with different names (of non-identical, related frames) is non-trivial in the re-write system. This would require access to the FrameNet hierarchy at compute time (or extensive pre-compilation). It is left for future work.

Given that two roles match, we distinguish two different match types according to whether the fillers also match:

Strict match This condition holds if two frames match and the names of the roles under consideration are identical in hypothesis and text. The role fillers must also match. The formal definition is this:

$< match_{role\_strict}, < m1, n1_{text}, n1_{hyp} >, < m2, n2_{text}, n2_{hyp} >> \in EM_{role\_strict}$ if
$< m1, n1_{text}, n1_{hyp} > \in Nodes_{match}, \ < m2, n2_{text}, n2_{hyp} > \in Nodes_{match},$
$role_t \in SR_{text}, \ role_h \in SR_{hyp}$ and
$from(role_t, n1_{text}) \ \wedge \ to(role_t, n2_{text}) \ \wedge \ from(role_h, n1_{hyp}) \ \wedge \ to(role_h, n2_{hyp}) \ \wedge$
$label(role_t) = label(role_h)$

We check whether we find a pair of matching frame nodes in text and hypothesis such that they are connected by edges with identical labels.

Loose match This condition generalizes the one above such that filler match is not required. It is practically only a test as to whether a given role is realized in both text and hypothesis at a matching frame. The formal definition just checks whether outgoing roles from a matching node in text and hypothesis have identical labels. We reserve a special node with match type "dummy" in the match graph as endpoint of the edge for the graph to be well formed. In fact, this is the only exception from the general constraint that matching edges must be connected by two matching nodes.

$< match_{role}, < m1, n1_{text}, n1_{hyp} >, < dummy, nn, nn >> \in EM_{role}$ if
$< m1, n1_{text}, n1_{hyp} > \in Nodes_{match},\ role_t \in SR_{text},\ role_h \in SR_{hyp}$ and
$from(role_t, n1_{text})\ \wedge\ from(role_h, n1_{hyp})\ \wedge\ label(role_t) = label(role_h)$

As the overall number of matching frames and thus roles available for matching is still comparably low, we did not implement any more elaborate role matching techniques for the time being. After a full integration of the FrameNet hierarchy and the developments of frame distance measures, experiments with similarity measures for role fillers like Levenshtein distance or WordNet-based measures will make sense. Moreover, it will also be important to include FrameNet's role-mapping information to be able to compare roles across different frames with non-identical names. In theory, role filler match or filler incompatibility should be an important factor for entailment checking.

5.3.7. Modality

The match types we have described so far detect *similar* material in hypothesis and text, which may indicate true entailment. In contrast, the detection of inconsistent modality types in text and hypothesis indicates negative entailment. Nodes from hypothesis and text can match only if both nodes are marked with identical modality. In case of modality mismatch, we remove all matches embedded under the respective node and instead mark the nodes with a new match type. So, the detection of modality has a double effect. On the one hand, the respective nodes are marked, which might allow the statistical model to derive certain regularities. On the other hand, structures embedded under mismatching modality are blocked from any other kind of matching. This effectively reduces the size of the match graph.

Formally, we generate new sets of matching nodes and edges, where unwanted matches are removed and mismatches are marked by a new match type $match_{mod\_mismatch}$:

$Nodes''_{match} := \{< m, n_{text}, n_{hyp} > \in Nodes'_{match} |\ modality(n_{text}) = modality(n_{hyp})\}$
$\cup\ \{< match_{mod\_mismatch}, n_{text}, n_{hyp} >\ |\ \exists < m, n_{text}, n_{hyp} > \in\ Nodes'_{match} |$
$modality(n_{text}) \neq modality(n_{hyp})\}$

$Edges'_{match} := \{< m, < m1, n1_{text}, n1_{hyp} >, < m2, n2_{text}, n2_{hyp} >> \in Edges_{match} |$
$modality(n1_{text}) = modality(n1_{hyp}) \ \wedge \ modality(n2_{text}) = modality(n2_{hyp})\}$

The function *modality* returns the modality annotation of a given node and 0 if no modality is annotated.

5.3.8. Context-dependent Enrichment with Background Knowledge

In Section 3.4.1, we argued that textual entailment sometimes requires "bridging inferences" using information beyond the level of frame semantics. Additional knowledge can easily be integrated in the matching architecture, as we have shown with a few hand-coded axioms. For example, one axiom encodes the background knowledge that *joining X* can result in *(being) member of X*. Certainly, there are also cases of joining that do not result in membership as in *join a chat room*, where one becomes visitor rather than member. In order to avoid overgeneration, we restrict these "heuristic inferences" to apply only if both sides of the axiom are fully instantiated by text and hypothesis. These axioms are interesting as they can mutually disambiguate information from text and hypothesis. For example, while *have* can in general not always be interpreted as referring to possession (*a day has 24 hours.*), in (5.8)-(5.9), it is clear that *has* in the hypothesis means possession.

(5.8) T: The twin buildings are 88 stories each, compared with the Sears Tower**'s** 110 stories.

(5.9) H: The Sears Tower **has** 110 stories.

Just as a proof of concept, we added four handwritten axioms to our system, two of which are sketched above. They apply in the magnitude of 2,5% of the sentence pairs in RTE sets. One idea for future research would be to explore the possibility of either deriving such rules from manual annotations or from other resources. One option could be to try to parse SUMO axioms and to semi-automatically translate them into rewrite rules.

5.4. Feature Extraction and Statistical Entailment Decision

In accordance with the vast majority of approaches to textual entailment, the final entailment decision is made in a machine learning setting. To this end, features are extracted from text, hypothesis, and the match graph (Section 5.4.1). They are then taken to train a statistical model (Section 5.4.2) on the RTE development corpora, where the correct entailment values are annotated.

5.4.1. Feature Extraction

In the feature extraction step (see Figure 5.1), a feature vector is computed for each text/hypothesis/match graph triple. The basic features we extract can be classified

according to their (i) level of representation (lexical, syntactic, semantic), (ii) *source* (text, hypothesis or match graph), and as to (iii) whether they represent a *proportional measure* (hypothesis/text, match/hypothesis ratio).

Extraction of the feature values is for the most part achieved simply by *counting*, e.g., the number of frame matches, and if applicable by computation of proportional ratios. Therefore, most features have integer or floating point values. Some additional "meta-features" such as the RTE sub-task annotation provided by the corpora or LFG parsing status (fragmentary true/false) are also recorded.

The only more involved processing at this stage concerns the detection of *connected clusters* in the match graph. As the matching strategy proceeds node by node and edge by edge, the match graph is typically not connected. We assume that the more and the larger connected parts are found, the more similar are hypothesis and text and thus the more likely entailment holds. An algorithm for the detection of connected clusters in the match graph we implemented starts from nodes without incoming edges. It follows outgoing edges as long as possible. Information about the number of different clusters and about their average size is represented by three different features.

Table 5.2 lists all 42 features we compute, together with a short description and classification as to whether the feature relates to text, hypothesis, or match graph. Also, the feature type is listed – integer, floating point, boolean, or enumeration type. We deliberately accept a certain redundancy in the feature space as we want to be able to explore which information is relevant for entailment decision and how the the different types of information interact. This feature list must not be regarded as being definite. It is rather intended as a basis for future experimentation and refinement.

Most features have been explained before, except for the features on the bottom of the table prefixed with `assigned`. These provide the number of frames and roles as originally assigned by Shalmaneser/Detour before integration with LFG. A simple overlap measure serves to compute the respective matching frames. These features do not depend on the LFG analysis and thus can serve as a fall-back in cases where LFG analysis fails.

5.4.2. Statistical Model

The entailment decision is a two-ways classification problem, for which we have experimented with different classifiers. In order to train a statistical model for textual entailment, we have generated feature vectors for RTE development sets, where entailment values are given. We experimented with different training and test sets and also with various machine learning systems and the attribute selection module of Weka (Witten and Frank, 2005). A first observation was that many learners (evaluators) selected features that seem intuitively important. However, also unintuitive features such as the number of predicates in the hypothesis are high-valued features. This might either indicate that the analysis component or matching algorithm has difficulties with long (short) hypotheses, or this apparent bias is due to idiosyncrasies in the development set. In fact, the results of feature selection and machine learning in general have to be take with some care, given that the development and test sets considered consist only of a couple of hundred sentences. An in-depth discussion of experiments with different

Feature	Comment	t	h	m	Type
matchnode_pred	LFG predicate matches			x	Int.
matchnode_pred_coref	– for structural co-reference			x	Int.
matchnode_pred_pro	– for pronouns			x	Int.
matchnode_wn_related	– using WordNet			x	Int.
matchnode_heuristic_inf	– using inferences			x	Int.
matchedge_syn	All LFG features			x	Int.
matched_feat_subcgf	Only subcategorized functions			x	Int.
matched_feat_modgf	Only modifying functions			x	Int.
matched_feat_gf	Subcat. plus modifying functions			x	Int.
modal_context_unmatched	Nodes with incompatible modality			x	Int.
matchnode_frame	Matched frames (identity)			x	Int.
matchnode_frame_related	Matching frames via frame relations			x	Int.
matchnode_detour_related	Matching frames via Detour			x	Int.
matchnode_detour_relto_h	Impact of Detour on h		x	x	Float.
matchedge_role	Matched roles (total)			x	Int.
matchedge_role_filler	Matched roles (with matching fillers)			x	Int.
lexids_h	Length of hypothesis	x			Int.
lexids_t	Length of text		x		Int.
rel_size_h_relto_t_lexids	Length of hypothesis w.r.t. text	x	x		Float.
preds_m_relto_h	Matched LFG predicates relative to h		x	x	Float.
frames_h/fes_h	Frames/FEs in hypothesis		x		Int.
frames_t/fes_t	Frames/FEs in text	x			Int.
frames_m/fes_m	Frames/FEs in match graph			x	Int.
frames_h_relto_t	Frames in hypothesis relative to text	x	x		Float.
fes_h_relto_t	– for roles	x	x		Float.
frames_m_relto_h	Frames of hypothesis that matched		x	x	Float.
fes_m_relto_h	– for roles		x	x	Float.
clusters_no	Number of connected clusters			x	Int.
clusters_avgsize	Average size of connected clusters			x	Float.
clusters_avgsize_relro_h	– relative to size of hypothesis		x	x	Float.
fragmentary	True if one LFG parse is fragmentary	x	x		Bool.
rte_task	RTE subtask (IE|IR|QA|SUM)	x	x		Enum.
rte_entails	Entailed/Non entailed	x	x		Bool.
Below: frames as assigned by Shalmaneser/Detour; intended for cases, where LFG parse fails					
assigned_frames_h	Frames in hypothesis		x		Int.
assigned_fes_h	– for FEs		x		Int.
assigned_frames_t	Frames in text	x			Int.
assigned_fes_t	– for FEs	x			Int.
assigned_frames_h_relto_t	Ratio of frames in hyp. and text	x	x		Float.
assigned_fes_h_relto_t	– for FEs	x	x		Float.
assigned_frames_m	Matched frames (identity)			x	Int.
assigned_fes_m	– for FEs			x	Int.
assigned_frames_m_relto_h	Matched frames relative to hyp.		x	x	Float.
assigned_fes_m_relto_h	– for FEs		x	x	Float.

Table 5.2.: All features of the SALSA RTE system.

machine learning settings will be given in the next chapter, where we also provide the results the SALSA RTE system achieved in the RTE-2 and RTE-3 challenges.

5.5. Summary of this Chapter

In this chapter, we have presented our own approach to textual entailment and its implementation in the SALSA RTE System. At the heart of the system is a linguistic analysis of text and hypothesis, which combines LFG grammatical, frame semantic, and ontological information. The analyses are represented as tripartite graphs, where the respective levels of analysis are connected via projections.

In a second stage, interaction between the different layers is established in order to enrich and normalize information on the different layers. We mainly exploit the availability of semantic information within LFG f-structures here.

While the linguistic analysis and further enrichment is performed for text and hypothesis in isolation, in the subsequent matching phase, the similarities of both are detected and marked. The basic assumption of this kind of "entailment reasoning" is that (directed) overlap of (meaning representations of) hypothesis and text correlates with entailment – the more of the hypothesis is covered by the text, the more evidence we have that entailment holds. We represent the similarities of text and hypothesis in a third structure we call match graph. It consists of matching nodes and edges plus additional information about different match types that licensed the respective node or edge matches.

Finally, in a feature extraction phase, we generate feature vectors from the information in the match graph, including relative measures with respect to text and hypothesis. These feature vectors are then used to train a statistical model for the textual entailment decision.

6. Evaluation

In this chapter, we evaluate the SALSA RTE system. We illustrate how the system operates, inspect errors, and discuss results the system achieved in RTE challenges. One of the main goals of this chapter is to study the contribution of predicate-argument structure in terms of frame semantics for checking textual entailment. To this end, we first evaluate the system against a shallow baseline. One result of this comparative evaluation is that the full system mostly does not perform better than the baseline and that machine learning results we obtain under different conditions are unstable. As it is in general difficult to derive reliable conclusions about the precise contribution of single parts of complex system architectures (especially in a machine learning setting), we then present a number of experiments that allow to assess in particular the performance and impact of components related to the the layer of frame semantics.

Throughout this chapter, the SALSA RTE system we refer to is the system as described in the previous chapter. Only for the RTE-2 challenge, a "prototype" variant of the system with a preliminary interface between LFG and Shalmaneser/Detour was used.

This chapter is structured as follows. Section 6.1 presents the primal evaluation of the system – its performance on RTE corpora. We describe the system settings used in the RTE-2 and RTE-3 challenges and discuss the results it obtained. In Section 6.2, we will mainly provide a qualitative error analysis. We will also shortly present a few examples to illustrate successful system behavior. Section 6.3 will present a detailed analysis of the system's performance against a "shallow" word-overlap baseline. One of the driving question is how reliable system results are under different machine learning conditions. In Section 6.4, we will discuss why the system does not fully confirm the intuition that an inclusion of frame semantics leads to a significantly better system performance. To this end, we will present a manual annotation of an RTE test set and provide a comparison of the automatic frame semantic annotation against this gold standard. The gold standard is too small for training and testing of the complete SALSA RTE system. We therefore assess the potential impact of frame semantics on the task of recognizing textual entailment by way of a simple algorithm, which simulates the system in a rule-based setting. In Section 6.5, we will summarize and conclude this chapter.

6.1. Performance of the SALSA System on RTE Corpora

The SALSA RTE system participated in the second and third RTE challenge (Bar-Haim et al., 2006; Giampiccolo et al., 2007a). We present the settings and results in Section 6.1.1 and Section 6.1.2, respectively. In Section 6.1.3, we discuss the results.

	Feature	
1	preds_m_relto_h	Predicate matches relative to hypothesis
2	assigned_frames_m_relto_h	Matched frames (Shalmaneser/Detour without LFG) relative to hypothesis
3	assigned_fes_m_relto_h	Matched roles (Shalmaneser/Detour without LFG) relative to hypothesis
4	cluster_avgsize_relto_h	Match graph size relative to hypothesis

Table 6.1.: Feature set for RTE-2 submission.

6.1.1. RTE-2

For the RTE-2 challenge, a prototype of the SALSA RTE system was used, where the interface between LFG and frame semantics was still preliminary. We manually chose a small and intuitively plausible feature set for the submission, which led to constant results on a number of classifiers in pilot experiments. The feature set is displayed in Table 6.1. Two system runs we submitted that use different classifiers from the Weka toolkit (Witten and Frank, 2005). A simple conjunctive rule classifier was used for run 1 and the state-of-the-art LogitBoost[1] classifier from Weka's meta classifiers for run 2. These relatively different classifiers were chosen to level out the impact of the classifier on overall performance. Training was done on the RTE-2 development corpus only; results are measured on the respective test corpus.

The conjunctive rule classifier of run 1 generates a single rule measuring predicate and frame matches relative to the hypothesis (features 2 and 1 from Table 6.1):

$$\begin{aligned}&(\texttt{assigned\_frames\_m\_relto\_h} \leq 0.954546) \text{ and}\\ &(\texttt{preds\_m\_relto\_h} \leq 0.485294)\\ &\Rightarrow \texttt{rte\_entails} = 0\end{aligned}$$

This rule defines that a sentence pair is taken as a true entailment if roughly all frames and half of the LFG predicates of the hypothesis are matched in the text. This rule confirms the intuition that coverage of all frames occurring in the hypothesis is a prerequisite for entailment as this ensures that hypothesis and text deal with the same topic. The lower restriction on predicate overlap can be explained by the characteristic of textual entailment that *some* additional material in the hypothesis is acceptable. Note that the frames considered by this rule are the ones originally assigned by Shalmaneser/Detour and not those integrated by LFG projection (see the lower part of Table 5.2). This indicates that the preliminary interface between LFG and frame semantics in this first prototype implementation indeed did not work very well – too few frames were projected.

The LogitBoost classifier used in run 2 selects features 1, 2 and 4 from Table 6.1 in its iteration steps. This means, that the average size of the connected clusters in the

[1] LogitBoost performs *additive logistic regression* using the classifier *DecisionStump*.

	Overall accuracy	IE	IR	QA	SUM
Run 1	59.0	49.5	59.5	54.5	72.5
Run 2	57.8	48.5	58.5	57.0	67.0

Table 6.2.: RTE-2 results.

match graph (feature 4) is taken into account in addition to the features used in run 1. Thus, a combination of grammatical, semantic and ontological information is considered. Feature 3 counting matched semantic roles was ignored by both learners.

Table 6.2 lists the results on the RTE-2 test data. Run 1 performs slightly better than run 2, overall and on most subtasks. Accuracy of both runs differs considerably across subtasks, ranging, e.g., for run 1 between 49.5% for IE and 72.5% for SUM. The SALSA RTE system performed in the middle ranks of the competing systems.

6.1.2. RTE-3

For the RTE-3 challenge, we implemented a "shallow" module measuring lexical overlap on content words (see Section 6.3 for details) to be used in combination with the full system.

A combined set consisting of the RTE-2 development and test set as well as the RTE-3 development set was used for training. The LogitBoost learner was used again as it had produced good and stable results within several tests we performed. We therefore decided to keep the learner constant and experiment with different features for the RTE-3 submission. To this end, we used the feature (attribute) selection component of the Weka tools.

Feature Selection for RTE-3 Submission

In order to derive the feature set for RTE-3 in a more principled way, we ran several of Weka's attribute selection strategies on different available RTE corpora. The initial set of features are all features displayed in Table 5.2. Table 6.3 displays the five best ranked features of a typical feature set, computed by Weka's ChiSquaredAttributeEval[2] component on the RTE-2 test set with 10-fold cross-validation.

The best ranked feature (a) basically measures word overlap on content words. This confirms the common observation that this is a good indicator for textual entailment as implemented in the available corpora. The remaining high-valued features cover different layers of analysis. Feature (b) is from the grammatical layer, feature (d) is from the realm of structural semantics. The average size of connected clusters (c, e) includes LFG, frame semantic, and ontological information. Features measuring the average size

[2]The component is described as to *evaluate the worth of an attribute by computing the value of the chi-squared statistic with respect to the class.*

	Ranking	Feature	
a	41.052	`preds_m_relto_h`	Predicate matches relative to hypothesis
b	23.583	`matched_feat_gf`	Matching grammatical functions
c	21.111	`clusters_avgsize_rel_h`	Size of connected clusters in match graph relative to size of hypothesis
d	20.909	`matchnode_modal_context`	Number of matching nodes with incompatible modality
e	20.35	`clusters_avgsize`	Size of connected clusters in match graph

Table 6.3.: Typical feature selection result.

	Feature	
a	`preds_m_relto_h`	Predicate matches relative to hypothesis
b	`matched_feat_gf`	Matching grammatical functions
e	`cluster_avgsize`	Size of connected clusters in match graph

Table 6.4.: Feature set for RTE-3 submission.

of connected matching parts of hypothesis and text occur twice, absolute (e) and relative to the hypothesis (c). As variant (e) has been ranked higher than (c) by most of Weka's feature selection modules, we included (e) in the feature set for the RTE-3 submission, together with the two most stable features. The set used in the submission is displayed in Table 6.4.

RTE-3 Results

For the RTE-3 challenge, the selected features were trained separately first, then a "meta classifier" was used to make the final entailment decision. This more elaborate training architecture will be discussed in the next section. We submitted two runs, one with (run 1) and one without (run 2) addition of the "shallow" lexical overlap component. Both runs achieved almost identical results, as can be seen in Table 6.5. Variance across subtasks is high again. The result are slightly better than for RTE-2, which confirms to the general trend. The system performed in the middle ranges compared to competing systems again.

6.1.3. Discussion of Results

The SALSA RTE system has achieved satisfactory results in the RTE-2 and RTE-3 challenges, implementing an approach of modeling textual entailment via "structural and semantic overlap" between text and hypothesis. Frame semantic information seems to contribute to the entailment decision. For example, the conjunctive rule of run 1

	Overall Accuracy	IE	IR	QA	SUM
Run 1	62.5	51.0	68.0	74.0	57.0
Run 2	62.6	50.0	69.0	72.5	59.0

Table 6.5.: RTE-3 results.

(RTE-2) takes a high degree of semantic similarity based on frames, joint with medium degree overlap at the syntactic predicate level to model entailment.

Per-task analysis reveals a huge divergency of accuracy across RTE subtasks. At the same time, the "difficulty" of subtasks is distributed unevenly between RTE-2 and RTE-3, the only obvious constant being that IE is most difficult. This may be explained by the special characteristics of IE hypotheses. They are typically very short factive statements like "John Lennon is dead" providing little context. Many approaches report problems with IE examples. Overall, the unsystematic distribution of difficulty across subtasks in different challenges indicates imbalances in the corpora.

A result which diverges from observations reported for related approaches is that the combination of the full, "deep" system with a shallow overlap measure does not increase overall performance. We will provide a systematic comparative evaluation of the full system and the shallow component in different settings in Section 6.3.

So far, the intuitive appeal of frame semantic analysis has not been confirmed by remarkable increase in system performance. In the feature selection result used as feature set for the RTE-3 submission, frame semantic information is referred to only indirectly via average cluster size. At the same time, grammatical functions score well, which might indicate that (semantic) role information is a relevant entailment indicator. Other features which also occur in high ranks for different Weka feature selection modules are absolute LFG predicate match, as well as match of syntactic edges (`matchnode_pred`, `matchedge_syn`). Frame-related features are rarely selected. Again, from a machine learning view, the size of the development corpus is very small. Features that do not occur frequently and in the majority of sentence pairs are neglected by the machine learning systems. We therefore cannot draw final conclusions on the effectiveness of certain features from these experiments on RTE corpora.

6.2. Inspection of System Behavior

In this section, we will mainly be concerned with a qualitative analysis of errors made by the system (Section 6.2.2). Before that, we will shortly illustrate positive system behavior by presenting examples from RTE corpora for true positive and negative classification (Section 6.2.1). We conclude this section by pointing to open questions in Section 6.2.3.

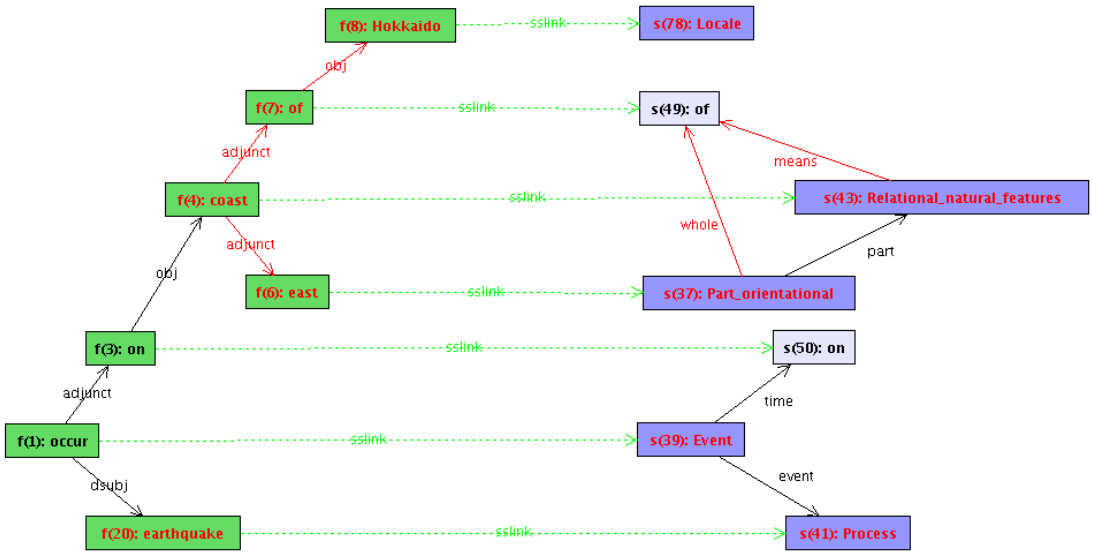


Figure 6.1.: Analysis of (6.2).

6.2.1. Positive System Behavior

True Positives

A typical true positive entailment decision is made by the system for (6.1)-(6.2), triggered by high semantic overlap between hypothesis and text in terms of matching predicates, frames, and f-structure.

(6.1) T: An earthquake has hit the east coast of Hokkaido, Japan, with a magnitude of 7.0 Mw.

(6.2) H: An earthquake occurred on the east coast of Hokkaido, Japan.

An analysis of (6.2) is displayed in Figure 6.1, where red parts are matches with (6.1). For example the grammatical analyses of *east cost of hokkaido* in hypothesis and text match completely. Likewise, the frame PROCESS matches the identical frame in the text. The main predicate *occur* –as expected– does not match the predicate *hit* in the text. Thanks to the frame analysis, the system manages to establish a match on the semantic layer – the frame EVENT evoked by *occur* matches the frame IMPACT evoked by *hit*. This is due to a heuristic "Detour" match, based on WordNet relatedness.

Other true positive examples are (6.3)-(6.4) and (6.5)-(6.6). The examples show that we obtain good results for texts and hypotheses of different lengths (absolute and relative), while some approaches have difficulties, e.g., with longer hypotheses (MacCartney and Manning, 2007).

(6.3) T: The system of government purchases of food under the U.N. Oil-for-Food Program was alleged to have many abuses.

(6.4) H: A government purchases food.

(6.5) T: In one of the latest attacks, a US soldier on patrol was killed by a single shot from a sniper in northern Baghdad, the military said yesterday.

(6.6) H: A sniper killed a U.S. soldier on patrol in Baghdad with a single shot.

True Negatives

True negatives in general are cases where overlap is low on different layers. Treatment of modality has proven quite effective for explicitly handling *dissimilarity*. 27% of correct entailment rejections involve mismatches of modality, while only 11.9% of all sentences contain modal contexts. The sentence pairs below are true negatives that involve what we subsume under modality. In (6.7), *can* is the LFG main predicate such that the complete text is embedded under `dia` modality. In (6.9), the modality mismatched is caused by unmatched will-future.

(6.7) T: The goal of preserving indigenous culture **can** hardly be achieved by a handful of researchers and curators at museums of ethnology and folk culture.

(6.8) H: Indigenous folk art is preserved.

(6.9) T: Even today, within the deepest recesses of our mind, lies a primordial fear that **will** not allow us to enter the sea without thinking about the possibility of being attacked by a shark.

(6.10) H: A shark attacked a human being.

We see potential for future improvement of system performance by including more dissimilarity measures like the modality feature just illustrated.

6.2.2. Error Analysis

The most obvious source of errors are basic analysis components – the word sense disambiguation system and the parsers. Typical problems are missing or inappropriate frame and role assignments. We will postpone an evaluation of these components until Section 6.4, and first discuss more general shortcomings of our approach.

The current system has a bias towards positive judgment. A probable explanation is that we have put to work many high-frequency features that measure *similarity* (e.g. predicate and frame overlap), but only few and low-frequency features that identify *dissimilarity* like modality mismatch. Therefore, the learners have a tendency to classify too few examples as not entailed, e.g., on the RTE-2 test corpus, we observe 29.5% false positives as opposed to 12.75% false negatives. We will discuss examples of both types errors below, focusing on false positives, though.

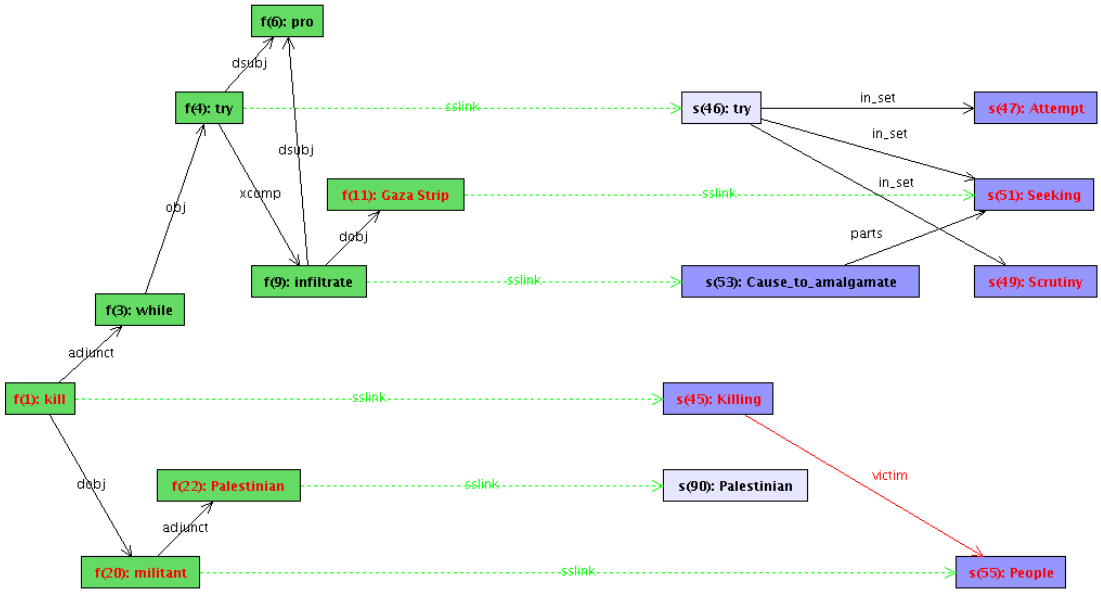


Figure 6.2.: Analysis of (6.12).

False Positives

Two common problems leading to false positive judgments are related to the level of semantic analysis we use and to the matching strategy we apply, respectively.

Granularity and relevance. Our approach is predominantly based on the coarse-grained level of frame semantic description. In certain cases, the approximate semantic techniques and generalizations we define are too optimistic and do not capture information that would be needed to reject entailment. Consider the false positive (6.11)-(6.12). The analysis of the hypothesis is displayed in Figure 6.2.

(6.11) T: Israeli helicopters fired two missiles in separate attacks on a Palestinian refugee camp on Wednesday, killing four people in a stepped-up campaign the army says is aimed at rooting out militants in the Gaza Strip.

(6.12) H: Palestinian militants were killed while trying to infiltrate the Gaza Strip. (FALSE)

Several problems can be observed in this sentence pair. The frame PEOPLE is assigned to *(four) people* as well as *(Palestinian) militants*, both filling the VICTIM role of the respective KILLING frame. While frame analysis successfully identifies comparable (semantically similar) material, we do not capture the information that *people* does not entail *militants*. In this case, this semantic normalization is too coarse-grained and alone

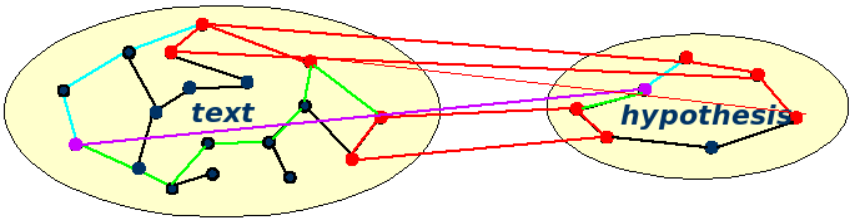


Figure 6.3.: Outliers from within matching clusters.

insufficient for distinguishing between positive and negative entailment. The role *fillers* and also the frame evoking elements have to be further analyzed and taken into account, too. This could be done using, e.g., string comparison or WordNet-based measures to assess the semantic compatibility of the respective parts of hypothesis and text. While our architecture checks WordNet compatibility of predicates in the matching phase, this information is not set in relation to matching or non-matching frames. As such correlations (frame matches vs. non-matching predicates) are not automatically detected by statistical models, it might be necessary to consider larger structures like a frame and its arguments in a more explicit top-down fashion to identify mismatching role fillers in larger configurations.

Another problem which can be seen in the analysis in Figure 6.2 is that the verb *try* in the hypothesis could not be uniquely disambiguated by the Detour weighting algorithm, and therefore has been assigned three different frames.[3] One of them has a counterpart in the text via a frame relation and the two others were related to frames in the text via a WordNet/Detour match. This leads to a total of five out of six frames being matched for this pair, which falsely indicates high semantic similarity. It is evident that frames and matches dot not all have the same relevance. Multiple frames assigned to an ambiguous word or a general frame for named entities like PEOPLE should be less relevant than, e.g., the frame assigned to the main verb if it is a full verb.

Distance of matching nodes. A related problem occurring in a number of cases are nodes in the match graph that are closely connected, e.g., in the hypothesis graph, but match with far distant parts of the text graph as in (6.13)-(6.14).

(6.13) T: Some 420 people have been **hanged in Singapore** since 1991, mostly for drug trafficking, an Amnesty International 2004 report said. That gives the country of **4.4 million people** the highest execution rate in the world relative to population.

(6.14) H: **4.4 million people** were **executed in Singapore**. (FALSE)

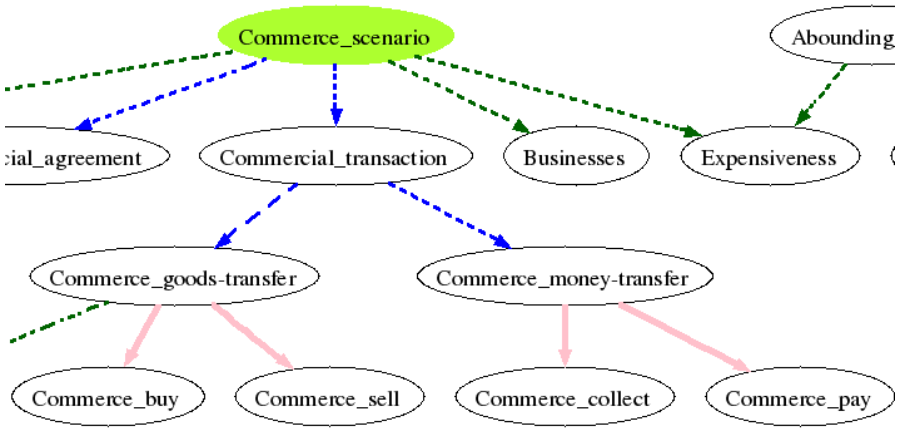


Figure 6.4.: Frame relations (screenshot of FrameGrapher).

In this example, the frames EXECUTION and PEOPLE are projected by very close f-structure nodes in the hypothesis (verb and direct object). In the text, however, the respective nodes belong to two different sentences and are thus far apart. Figure 6.3 schematically shows this configuration. The pink edge represents an "outlier" match.

In order to tackle this problem, we could introduce *weights* that reflect the relative distance of matching node pairs in the text and hypothesis graphs, measured in terms of f-structure or frame structure path distance. In Figure 6.3, the pink match thus could be identified as an outlier and receive a low weight. The distance to the closest matching neighbors is much shorter in the hypothesis than in the text, as can be seen in the relative lengths of the green and blue edges. Introducing respective weights should help to define further criteria for entailment rejection.

False Negatives

Frequent cases of false negatives involve similar, yet non-identical frames in text and hypothesis. Often, the information needed to establish a link between intuitively compatible frames is not available. In (6.15)-(6.16), the semantically compatible frames COMMERCE_SELL and EXPENSIVENESS have been assigned, but can not be matched.

(6.15) T: The first Barbie doll was sold for $3.

(6.16) H: The first Barbie cost 3 dollars. (TRUE)

Both frames are related via a complex path of FrameNet frame relations. The path can be seen in Figure 6.4, which is a screenshot of the FrameNet Grapher[4]. Different colors indicate different relations – COMMERCE_SELL is a perspectivization of COMMERCE_GOODS_TRANSFER, which is a subframe of COMMERCIAL_TRANSACTION.

COMMERCIAL_TRANSACTION in turn is a subframe of COMMERCE_SCENARIO while EXPENSIVENESS uses COMMERCE_SCENARIO. It is an open question however, how different configurations of frame relations can automatically be used to derive frame similarity measures. Moreover, the coverage of frame relations in terms of annotated instances is currently relatively low. We will come back to these problems in Chapter 7.

6.2.3. A Final Remark

Inspection of examples above and the experiments for the RTE-2 challenge indicate that features from different levels of analysis are used by the SALSA RTE system for modeling textual entailment. However, it is difficult to tell what the contribution of the single levels is. For instance, many positive examples such as (6.3)-(6.4) presented throughout this section exhibit high surface overlap while negative ones like (6.11)-(6.12) often do not. This raises the question of how the semantic analysis relates to more shallow measures like predicate overlap. Below, we will systematically compare performance of the full system against a shallow baseline and then present some experiments that identify contributions and shortcomings of different system components.

6.3. Evaluation Against a "Shallow" Baseline

It has been observed for many RTE systems that a combination of separately trained features can lead to an overall improvement in system performance, in particular if features from a more "informed" component and "shallow" ones are combined (e.g. Hickl et al., 2006a; Bos and Markert, 2006). Below, we provide a systematic analysis of the SALSA system's behavior on different training and test sets and with different feature combinations. In Section 6.3.1, we describe the experimental setup. Results are presented and discussed in Section 6.3.2.

6.3.1. Experimental Setup

We implemented an extended interface to automatically run Weka with different configurations. The extended interface supports testing with voting strategies and using a "meta learner", after training individual features or feature sets separately. The latter technique is used by several approaches using "multi-layer" matching strategies (e.g., Haghighi, Ng, and Manning, 2005). All figures we present in the following are computed with the LogitBoost classifier, which is used to train individual feature sets and is also used as meta learner.

[3] With the latest FrameNet and WordNet releases, the frame ATTEMPT is uniquely assigned to `try#v#1` by the Detour system.

[4] `http://framenet.icsi.berkeley.edu/FrameGrapher/`

Lexical Overlap Component

As a baseline to compare our full system against, we used a simple system that approximates textual entailment in terms of lexical overlap between text and hypothesis. This shallow system was also used within run 1 submitted to the RTE-3 challenge. It measures the relative number of words in the hypothesis that also occur in the text. Both text and hypothesis are tagged and lemmatized using Tree Tagger (Schmid, 1994), taking only nouns, non-auxiliary verbs, adjectives and adverbs into account. Training a decision tree on the relative word-overlap as single feature yields a system which performs comparable to other word-overlap based systems, achieving an accuracy of 60.6 % if trained and tested on the RTE-2 development and test set, respectively (using Weka's J48 classifier) and 57.5 % using Weka's LogitBoost classifier.

Feature Combinations

In the experiments, we investigated the behavior of our systems on various combinations of three different sets of text hypothesis pairs, namely the RTE-2 development and test set and the RTE-3 development set (see Table 6.6). We tested four different feature configurations:

(I) All 42 features generated by the system

(II) Three selected features of the system (run 2 of RTE-3 submission):

- `preds_m_relto_h` (overlap of LFG predicates)
- `matched_feat_gf` (matching of grammatical functions)
- `cluster_avgsize` (average size of connected clusters of the match graph)

(III) The features from II plus lexical overlap (run 1 of RTE-3 submission)

(IV) Lexical overlap alone

As a shorthand notation for the different conditions, we will use triples of feature-set, training-set, and test-set, e.g., `I-D2T2-D3` means all 42 features trained on the combined development and test set of RTE-2, tested on the development set of RTE-3.

6.3.2. Discussion of Results

The outcome of the experiments is shown in Table 6.6. Three central results are: (i) the shallow overlap feature most of the time outperforms the more informed features;[5] (ii) the combination of informed and shallow features only has a moderate effect on accuracy, and (iii) almost all features behave quite differently on different training and test sets.

To substantiate these results, we discuss three particular topics below – feature variance, corpus variance, and task variance.

[5]The complete featureset I always performs worse than the selected in features II. This indicates that the classifier does not manage to detect the best features.

test → ↓ train	D2				T2				D3			
	I	II	III	IV	I	II	III	IV	I	II	III	IV
D2					56.25	57.25	**58.625**	57.5	57.875	61.125	66.375	**66.625**
T2	56.375	58.75	60.625	**61.625**					57.5	60.875	63.75	**64.625**
D3	53.875	61.25	**61.75**	**61.75**	56.625	**58.75**	57.25	57.25				
D2T2									58.5	64.25	65.875	**66.375**
D2D3					58	58.625	**60**	58.5				
T2D3	56.75	**61.25**	60.875	60.875								

Table 6.6.: Performance of different feature combinations on different training and test sets using LogitBoost as learner on all feature sets and as "meta learner" for III, the combined set of II and IV.

Feature Variance

Variance among individual features and feature sets is large. Feature set II contains the most reliable and stable features of the complete feature set. We tested how this "more informed" feature set (II) compares to the shallow word overlap feature (IV) and whether their combination (III) increases accuracy. As can be seen from Table 6.6, in most of the cases, IV performs best (best accuracy for each configuration is printed in boldface). The combination of the "deep" features with the "shallow" one usually leads to an improvement, e.g. in the D2-D3 configuration, where feature set II alone achieves 61.125, while the combination with IV boosts the performance by 5% (III). Still, the combination mostly does not perform better than IV alone. There are only few exceptions, where the inclusion of the word overlap feature lowers the performance, e.g. from 61.25 (II-T2D3-D2) to 60.875 (III-T2D3-D2). It would be interesting to test the system on data, where lexical overlap is not such a good entailment indicator.

Combinations of features often perform lower than the best individual feature in the set. For instance, in configuration D2T2-D3, feature III achieves 65.875, compared to 66.375 for IV alone. In contrast to results reported by related approaches, we generally could not observe a positive effect for the combination of features in a meta feature. Apart from the size of the training data, feature dependence might be an explanation for this.

Corpus Variance

Testing on T2 (RTE-2-test) seems to be the hardest task. No configuration achieves an accuracy of more than 60%. In contrast, the overall best performance is 66.625% accuracy (IV-D2-D3). One explanation for this observation is imbalance of the "difficulty" of datasets. An indicator for the difficulty of a test set is the average lexical overlap of text and hypothesis. We measured the average word overlap among entailed and not entailed pairs for different sets and also computed difference in proportion between entailed and not entailed examples. It amounts to 0.05 on T2 and 0.13 on D3 (see also Figure 6.7), thus explaining why D3 is "easier" as compared to T2.

Using a larger training set should lead to a better performance. The bottom part of table 6.6 shows results computed with a combined training set consisting on the RTE-2 training and test set. However, a stable positive effect can not be observed. For most feature sets the gain in performance is very small, e.g. from 57.875 (I-D2-D3) and 57.5 (I-T2-D3) to 58.5 (I-D2T2-D3). On some feature sets, the performance even decreases, e.g. from 61.625 (IV-T2-D2) and 61.75 (IV-D3-D2) to 60.875 (IV-T2D3-D2). The largest effect occurs for feature set II. Its performance increased from 61.125 (II-D2-D3) and 60.875 (II-T2-D3) to 64.25 (II-D2T2-D3). These unstable effects can be explained by the small size of the datasets. In terms of machine learning, extending a training set by factor 2 (from 800 to 1.600 items as we did) does not make a qualitative difference. The improvement observed by (e.g. Hickl et al., 2006a) was achieved by going to 10.000 training items. It would be very interesting to see how the performance of our system would develop on a much larger training set. Unfortunately, the only currently

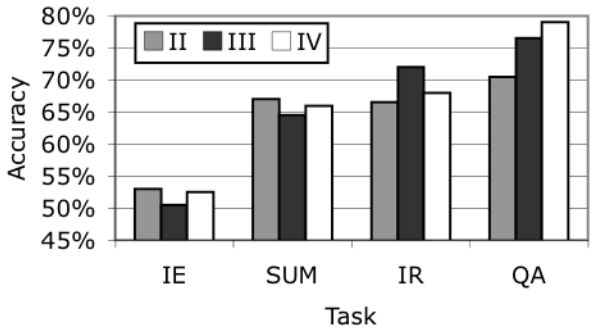


Figure 6.5.: Per task accuracy (D2T2-D3).

Task	Entailed	Not Entailed	Δ
IE	0.41	0.35	0.06
SUM	0.39	0.22	0.17
IR	0.29	0.23	0.06
QA	0.44	0.20	0.24
	0.38	0.25	0.13

Table 6.7.: Average word overlap per task for D3.

available textual entailment corpora are the relatively small ones provided within the RTE challenges.

Task Variance

Figure 6.5 shows a per task analysis for the feature sets II, III and IV (D2T2-D3). The system performs best on the Question Answering (QA) task, where it achieves almost 80% accuracy. This differs from the experience with the RTE-2 submission, where the system performed best in the summarization (SUM) task. A somewhat unexpected result is that the more "informed" feature set II performs better in SUM than the shallow feature set IV, while it is the other way round in QA. The per task analysis also confirms the observation that the combination of (deep and shallow) feature sets (III) behaves heterogeneously in terms of accuracy – on information extraction (IE) and SUM it performs worse than both II and IV; on information retrieval (IR), it outperforms II and IV; and on QA, its accuracy is between that of II and IV. Given the general variance discussed above, however, we think that these observations do not allow general conclusions.

The large variability of the shallow overlap feature (IV), which ranges between 52.5% (IE) and 79% (QA), can partly be explained if we compare the average word overlap measures for positive and negative pairs among the individual tasks (Table 6.7). The

difference (Δ) between positive and negative examples measures the discriminative power of word overlap in deciding entailment. For IE, SUM, and QA, this difference correlates with the accuracy achieved on the respective tasks (cf. Figure 6.5). Note however, that the difference (Δ) between positive and negative examples in IE and IR is identical while the accuracy of the word overlap feature differs drastically on both tasks. Their absolute overlap values differ, though.

6.4. Experiments with a Manual Gold Standard

Experiments performed in the context of the RTE-2 and RTE-3 submissions were ambivalent concerning the impact of frame semantic information on entailment decisions derived by the SALSA RTE system. However, the comparative evaluation reported above clearly shows that "informed" features of the SALSA RTE system overall do not perform better than a shallow word overlap baseline. In other words, the intuitively expected positive impact of using frame semantics for modeling textual entailment has not yet been confirmed by increased system performance. We identified the factors detailed below as potential limit to the success of applying frame semantics in automatic entailment checking.

FrameNet coverage FrameNet is incomplete in several respects. One factor limiting its success in entailment checking could be that its coverage is simply not good enough to provide a solid semantic annotation of the given corpora. A related study in the context of annotation of a German newspaper corpus (Burchardt et al., 2006a), reports that one third of the verb occurrences could not be annotated with existing FrameNet frames (largely due to the incompleteness of the frame inventory, not to cross-lingual differences).

Quality of automatic linguistic analysis The SALSA RTE system strongly relies on parsers providing syntactic analyses (LFG and Collins) and on Shalmaneser and the Detour system providing frame semantic annotations. One reason for low overall performance could be that the quality of the annotations produced by the systems is not good enough to significantly support RTE inferences. Another problem could be the quality of the interface between LFG and Shalmaneser/Detour.

Entailment modeling If appropriate annotations are provided, it might be the case that the available information is not optimally accessed and modeled by the matching algorithm and feature space we used. The algorithm was deliberately designed to be robust against partial or missing analyses. The feature space is relatively redundant as we wanted to test which features are considered relevant by the machine learners. Yet, the overall bias towards positive entailment judgement might indicate that this approach is too liberal.

In the following, we study these different factors. To this end, we manually annotated the RTE-2 challenge test set with frame semantic information. This corpus called FATE

serves as gold standard for the experiments concerning frame semantics. In Section 6.4.1, we give a short description of the FATE corpus and on this basis in Section 6.4.2 evaluate the impact of the FrameNet resource coverage. Section 6.4.3 will be concerned with the linguistic analysis components. We will provide performance statistics for syntactic parsers. We also compare the automatic annotation by Shalmaneser/Detour with the gold standard to get an indication of the quality of the automatic frame semantic analysis and evaluate the LFG/Shalmaneser/Detour interface. Finally, in Section 6.4.4 we will assess the potential of using frame semantic annotation for checking textual entailment by processing the gold standard and the automatic annotation with a simple, controlled algorithm.

6.4.1. The FATE Corpus

This section presents the FATE ("Frame Annotation for Textual Entailment") corpus we built as gold standard. The FATE annotation has been carried out on the test set of the RTE-2 challenge. Frames and roles were manually annotated on top of a syntactic structure produced by the Collins parser (Collins, 1999) – the same syntactic analysis the Shalmaneser/Detour systems take as input. The annotation basically follows the annotation guidelines used for the creation of the German SALSA corpus (Burchardt et al., 2006a).

Below, we will shortly describe central features of the corpus which deviate from the SALSA annotation and which are relevant for our evaluation purposes. More details on FATE can be found in Burchardt and Pennacchiotti (to appear). The corpus will be available at the SALSA project page.[6]

We created FATE for studying the usefulness of frame annotation for textual entailment and therefore focused the annotation on those parts of the sentences which are central for this task. In particular, we only annotate *relevant FEEs* and do so only within *relevant spans* of the texts.

Relevant FEEs

Intuitively, not all frames that can be annotated to a given text or hypothesis are candidates for playing a central role in an entailment decision. For example, an event-denoting frame evoked by the main verb is more likely relevant than a frame like CALENDRIC_UNIT evoked by a noun like *Monday*. Starting from the idea is that textual entailment inferences should mainly by supported by properties and descriptions of central facts, we introduced the notion of *relevant FEEs*. With this notion, we refer to words that evoke frames which are somehow relevant to the overall situation(s) described. Consider the following example.

(6.17) T: **Authorities** in Brazil **say** that **more** than 200 **people** are being **held hostage** in a **prison** in the **country's remote**, Amazonian **jungle state** of Rondonia.

[6] www.coli.uni-saarland.de/projects/salsa

(6.18) H: **Authorities** in Brazil **hold** 200 **people** as **hostage**.

A complete annotation of (6.17)-(6.18), e.g., in the style of FrameNet's full text annotation[7], would select all words displayed in boldface as FEEs. In the FATE annotation, we only annotate the underlined, relevant FEEs (words considered not relevant as FEE are still available as arguments of frames). Indeed, these are the only words which evoke frames describing the central situations: *hostage* evokes the KIDNAPPING frame, *say* evokes STATEMENT.

In a pilot annotation for a small, randomly selected set of 15 sentences, we achieved a very high inter-annotator agreement concerning the question of what are relevant FEEs. We elaborated operational characterizations of relevance, which have been used to guide the main annotation. For example, a word counts as relevant only if the frame it evokes also realizes one or more roles in the given text. Note that texts and hypothesis are annotated separately in random order – the relevance of FEEs is judged for each sentence independently.

Relevant Spans

For many pairs in the RTE datasets only parts of the texts contribute to the inferential process that allows to derive the hypothesis from the text. A typical example is (6.19)-(6.20) below, where only the sections in bold face are really important.

(6.19) T: Soon after the EZLN had returned to Chiapas, Congress approved a different version of the COCOPA Law, which did not include the autonomy clauses, claiming they were in contradiction with some constitutional rights (private property and secret voting); this was seen as a betrayal by **the EZLN and other political groups.**

(6.20) H: **EZLN is a political group.**

In FATE, we annotated only the specific sections within the text/hypothesis pairs that contain relevant material for the task of textual entailment recognition. The annotators were provided with a markup of the these relevant spans. The spans are derived using the ARTE annotation, which provides alignment annotations for the positive pairs in the RTE-2 test set (Garoufi, 2007). As the ARTE annotation does not provide alignment annotation for negative text/hypothesis pairs, it cannot be used to provide an automatic markup of relevant spans for these pairs. Accordingly, we conducted a full annotation for negative examples in FATE. Indeed, in negative entailment pairs, the concept of relevant span is often not applicable to the texts. For example, if a hypothesis contains information which is not contained in the text, there is no markable available in the text.

FATE Annotation Statistics

The inter-annotator agreement in the FATE annotation was quite good, e.g., 88% for frame assignment. The sporadic cases of disagreement on frames usually involve the

[7] See `http://framenet.icsi.berkeley.edu/index.php?Itemid=84`

Leadership	196	Attempt	40
Statement	152	Being_employed	38
Killing	92	Causation	36
People_by_vocation	90	Death	35
Change_position_on_a_scale	85	Intentionally_create	35
Attack	73	Businesses	34
Finish_competition	68	Education_teaching	33
Being_located	51	Hostile_encounter	31
Event	50	Protecting	31
Military	49	Activity_start	29
Surpassing	46	Becoming_aware	29
Using	46	Means	29
Cause_change_of_position_on_a_scale	45	Cause_harm	28
Aggregate	43	Locale_by_use	26
Medical_conditions	42	Behind_the_scenes	25

Table 6.8.: Most frequently annotated frames in the RTE-2 test set.

choice of different but highly similar frames (e.g. Risky_situation vs. Run_risk) or an Unknown frame (see below) used by one annotator instead of the correct one present in the FrameNet hierarchy. The final corpus contains 4488 annotated frames and 9512 annotated roles, with an average of 5.6 frames per pair, and 2.1 roles per frame.

Table 6.8 reports the most frequent frames occurring in FATE, and their number of occurrences. This list gives an impression of the semantic domains characterizing the RTE-2 corpus, mostly referring to killing, disaster and competition events.

6.4.2. FrameNet Coverage

FrameNet's coverage on the RTE corpus is rather good and therefore unlikely to be an important factor for performance issues of the SALSA RTE system. In order to assess the impact of FrameNet's coverage, the annotators of FATE marked cases where no suitable FrameNet was available with a pseudo frame (or role) called Unknown. The annotation contains 373 Unknown frame instances, accounting only for the 8% of the total frames (unknown roles are 1% of the total roles). Thus, FrameNet's coverage is at 92%. This number differs considerably from figures reported for SALSA's German corpus annotation cited above. One possible reason for this discrepancy might by the strategy of annotating only *relevant* frames in FATE. Also, the annotators of FATE were allowed to annotate frames that looked appropriate in a rather flexible way, while the SALSA annotation for German followed a stricter annotation guideline.

	LFG						Collins
	coverage	full	frag.	full t+h	frag. t or h	single t or h	coverage
dev	98.0	87.7	10.3	77.9	21.3	0.8	99.1
test	99.1	85.4	13.6	74.3	23.7	2.0	99.0

Table 6.9.: Parsing results on RTE-2 corpora for texts (t) and hypothesis (h) (in %).

6.4.3. Quality of Automatic Linguistic Analysis

In this section, we first explore the quantitative performance of the parsers. We then present an experiment carried out to assess the accuracy of the Shalmaneser/Detour systems against the FATE corpus as a gold standard. Finally, we evaluate the interface between LFG and Shalmaneser/Detour.

Quantitative Performance of Parsers

Syntactic parsing. Providing a full qualitative evaluation of syntactic parsing is beyond the scope of this thesis. We refer to the literature such as Kaplan et al. (2004), where a comparative evaluation of the LFG parser and the Collins parser (used by Shalmaneser) is provided. The accuracy values reported for both parsers are comparable (both achieve *F-scores* in the high 70s).

Quantitative performance of the LFG and the Collins parser on RTE data is quite good. We ran both parsers on the datasets of the first three RTE challenges. Both have almost full coverage. Table 6.9 shows parsing results on the RTE-2 development and test corpus, which are comparable to results obtained on other sets.

The LFG parser provides much richer representations than the comparably shallow Collins parser (cf. Kaplan et al., 2004). In its overall coverage figures given in Table 6.9, fragmentary parses (see Section 3.2.2) are included. More than 10% of the sentences receive only partial LFG parses. For textual entailment, text/hypothesis *pairs* have to be parsed. Table 6.9 also displays how fragmentary parses are distributed among sentence pairs – about 75% of the pairs are fully parsed (*full t+h*); in about 22% of the cases, either text or hypothesis are fragmentary (*frag. t or h*). Only few texts or hypotheses remain without a parsed counterpart (*single t or h*).

We do not perform a compositional semantics construction within LFG. So, fragmentary parses are unproblematic as long as the relevant predicates and their arguments reside within parsed fragments.[8] If relevant parts are fragmented, we might miss information about some grammatical relations or certain modifiers in the analysis. Yet, the frames and roles can (mostly) still be projected (see below). Also, the subsequent processing stages are robust in that they do not rely on fully connected parses.

[8] If this is the case, fragmentation has minor impact only – the overall properties of the sentence such as the number of predicates change and peripheral parts do not receive a full grammatical analysis. Both should not have a strong impact on the entailment decision.

Semantic parsing. We ran the semantic parsers Shalmaneser and Detour on all RTE corpora to assess their quantitative performance. The number of frames assigned by Shalmaneser, e.g., on the RTE-2 corpora, averages at 2.4 frames and 3.8 roles per sentence. The number of frames is roughly comparable to the number of manually annotated frames in FATE (see Section 6.4.1). The number of roles is a little lower than in the manual annotation. If we take into account, that the manual frame annotation in FATE provides only a partial annotation (relevant frames within relevant spans on positive pairs), the number of frames assigned by Shalmaneser is too sparse. Combined with the Detour system, the number goes up to 3.8 frames and 8.4 roles per sentence. Below, we will determine the quality of the automatic frame assignment.

Quality of Automatic Frame Semantic Analysis

The following experiments assess the quality of the automatic frame semantic analysis with the help of the gold standard FATE corpus.

Experimental setup. Both semantic parsers were run – Shalmaneser alone (SHA) and Shalmaneser boosted with the Detour (DET). The special frame UNKNOWN was removed from FATE before running the experiments. System performance is evaluated on three different tasks – frame, role, and filler assignment. In frame assignment, we count the number of frame labels assigned by the system that occur in the gold standard annotation of the respective sentence (precision), and the number of frame labels in the gold standard annotation that can also be found in the system annotation (recall). We further distinguish two different conditions – *strict* and *relaxed* match. Strict match means identity of labels in automatic annotation and gold standard. Relaxed match also accepts pairs of labels which are related via the FrameNet hierarchy or by the WordNet inheritance relation using the algorithm described in Section 5.3.5. Role assignment is evaluated only on the frames that are correctly annotated by the system. We compute precision as the number of roles assigned by the system that are in the gold standard, and recall as the number of roles in the gold standard that have been assigned by the system. Similarly, we compute accuracy of filler assignment on the set of correctly hypothesized roles only, as the percentage of fillers which have identical syntactic head lemmas in automatic annotation and gold standard.

The relaxed frame match has been adopted to account for the fact that a certain variance is normal in the task of assigning semantic annotations, as is shown in the inter-annotator agreement values on frame assignment reported, e.g., in Burchardt et al. (2006a). The relaxed frame match is intended as "fairer" condition for the systems. For example consider the following text:

(6.21) Cars exported by Japan decreased.[9]

Here, in the gold standard annotation, *exported* relates to the frame EXPORT, while Shalmaneser assigns the frame SENDING, which is still a plausible annotation.

[9] From RTE-2 test set.

System	Frame assignment				Role assignment		Filler assignment
	Strict		Relaxed				
	Precision	Recall	Precision	Recall	Precision	Recall	Accuracy
SHA	0.35	0.40	0.48	0.55	0.54	0.37	0.77
DET	0.19	0.55	0.30	0.85	0.52	0.36	0.75

Table 6.10.: Systems' performance over the gold standard.

Analysis of results. Comparison of the systems' results with the gold standard annotation shows that the systems' performance is low and needs to be improved. Table 6.10 shows the results for the different tasks. Results on frame assignment indicate that the systems have some difficulties in assigning exactly the gold standard frame. Yet, the large improvement from strict to relaxed match (e.g., Precision +13% and Recall +15% for SHA), shows that the systems often assign a frame which is similar to the gold standard frame. The results on role assignment leave room for improvement. Filler assignment results are relatively good, showing that in most cases, a correctly hypothesized role is filled with the correct syntactic constituent.

Results for frame, role, and filler assignment are lower than results of Shalmaneser reported in Erk and Pado (2006). This is probably due to two facts. First, Shalmaneser has been trained and tested on different kinds of corpora; sentences in the RTE dataset are typically longer and less "prototypical" if compared to the FrameNet sample corpus Shalmaneser has been trained on. Second, the FATE annotation was performed using a liberal annotation style. This aspect has been addressed by the use of the Detour system, which implements a more flexible frame assignment. However, the precision of the Detour on the FrameNet data presented in Section 4.2 was also higher than that on the RTE gold standard corpus. Again, as there are no standard measures for frame similarity, the figures give only an estimate of the quality of the automatic annotation.

Quality of LFG/Shalmaneser Interface

The interface between Shalmaneser/Detour and LFG is a piece of software that establishes the semantic projection from LFG f-structure nodes into frames and FEs (see Section 5.2.1). Looking at its accuracy, we have to distinguish between the projection of frames and the projection of FEs. In the current implementation, 97% of the frames and 74% of the roles assigned by Shalmaneser/Detour can successfully be projected from LFG some f-structure node.

As frames are typically evoked by single words, a high accuracy can be achieved in the interface. About 90% of the frames can be treated by a straightforward algorithm based on head lemma information. It has been improved to 97% by implementing heuristics rules that use full forms and substring-matching in cases where the lemma has not been found. Projecting FEs has proved more difficult as the fillers are often named entities,

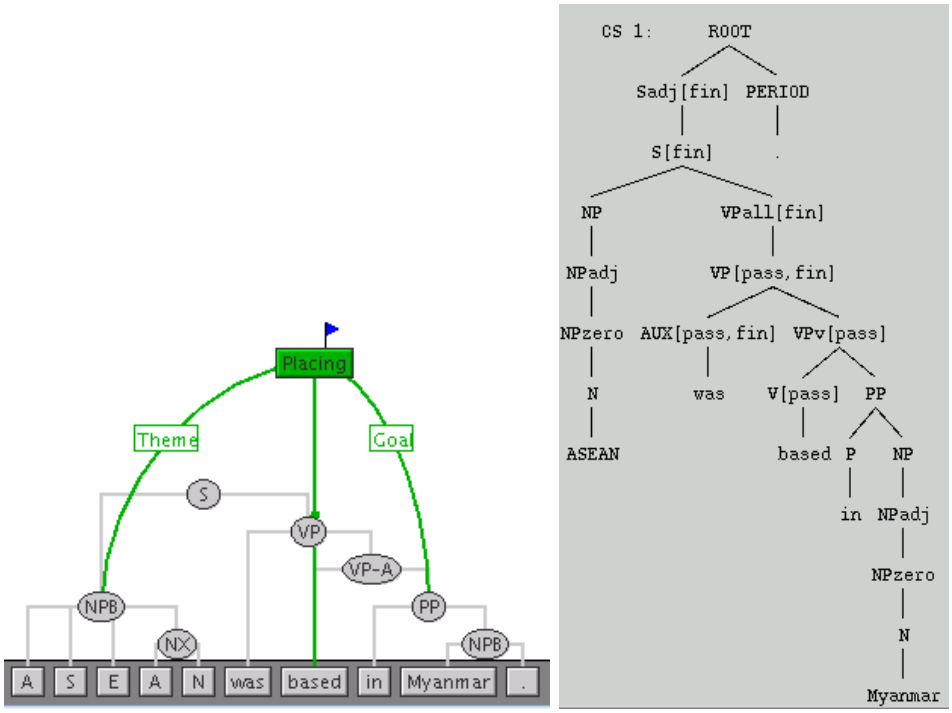


Figure 6.6.: Different tokenization in Collins and LFG parse.

pronouns, or complex phrases, where differences across parsers are ubiquitous. The initial accuracy for lemma-based role projection was at only 64%. The heuristic method mentioned above improved the accuracy to currently 74%. Still, further improvement of the interface is possible. Systematic cases like the different tokenization of *ASEAN* in the example displayed in Figure 6.6 should be relatively easy to handle, but more effort will be needed for a special treatment of less systematic differences.

Intermediate Result

We have shown above that FrameNet's coverage on the RTE corpora is rather good and therefore most probably not an issue impeding the performance of frame-based modeling of textual entailment.

Poor evaluation results of the linguistic components are more likely explanations for the limited performance of the SALSA RTE system. While syntactic parsing seems to work sufficiently well, the quality of the automatic frame semantic analysis is not satisfactory. Precision of frame assignment between 20% and 50% in different conditions and roughly 50% recall is rather low. In particular, recall for role assignment of about 35% is a problem since role (mis-)match is intuitively more important for the entailment decision than frame match. Unfortunately, this sparseness of role assignments is amplified

by problems occuring in the interface that establishes the semantic role projection from the grammatical layer.

6.4.4. Entailment Modeling

The suboptimal performance of automatic linguistic analysis is a probable explanation for the fact that our entailment architecture does not perform as good as we had expected. The question remains whether it would perform better if the frame assignment component had gold standard quality.

The most straightforward experiment is to re-run the SALSA RTE system on the gold standard corpus. Low performance would most likely indicate a lack in modeling entailment. However, due to its annotation strategy the gold standard does not lend itself for being processed by the system. An exploratory system run on four fold cross validation on the gold standard confirmed the expectation that the corpus has an artificial bias. Accuracy went up to 75%. The reason is that only relevant frames were annotated in FATE and in particular only relevant spans in positive cases. Therefore, negative pairs tend to have a considerably higher absolute number of frames. This bias affects all features in the system based on frame number counts. After disabling the respective features if possible, the system still had a better performance than that on the automatically annotated data. But as the gold standard has only the size of a RTE test set, and we cannot expect significant results from training and testing the full system. A large variance across folds confirms this expectation.

As an alternative, we now present an experiment based on a simple algorithm which extracts frame-based statistical information from the positive and negative examples of the annotated corpus. It also captures the overlap of frame structures between text and hypothesis in an entailment pair. In contrast to the full system, the algorithm focuses on frame semantic annotation and has no fall-back alternative such as grammatical functions.

Experimental Setup

We studied if basic frame-based information extracted from the gold standard helps in discriminating positive and negative entailment in the RTE-2 test set. We investigated three basic types of information:

Frame overlap This is the percentage of frames in the hypothesis which have a matching frame in the text. By match, we mean an exact match here, i.e., both frames must have the same label.

Role overlap Role overlap measures the presence of identical role label annotations in hypothesis and text irrespective of the respective fillers. We count role overlap only on those frames of the hypothesis which have a matching frame in the text.

Filler overlap This is the percentage of matching role fillers, counted on matched roles. We consider two types of filler matching – strict matching where the filler strings

Information type	Annotation	Average overlap		Difference
		On positives	On negatives	
Frame overlap	Gold standard	42.1%	28.1%	+14.0%
	Shalmaneser	39.6%	29.6%	+10%
Role overlap	Gold standard	56.0%	47.3%	+8.7%
	Shalmaneser	33.2%	29.7%	+3.5%
	Baseline lexical overlap	76.6%	66.1%	+10.5%

Table 6.11.: Average frame and role overlap over positive and negative pairs in the gold standard and in Shalmaneser output and lexical baseline.

Annotation	Match type	Overlap		Difference
		On positives	On negatives	
Gold standard	Strict	12.7%	9.8%	+2.9%
	Levenshtein	20.0%	14.8%	+5.2%
Shalmaneser	Strict	14.8%	13.6%	+1.2%
	Levenshtein	18.8%	18.7%	+0.1%

Table 6.12.: Average role filler overlap over positive and negative pairs in the gold standard and in Shalmaneser output.

are identical and *Levenshtein* matching when the Levenshtein distance between the two fillers is above a given threshold. Levenshtein matching is adopted to approximate a lexical overlap strategy.

In order to find out if the three frame-based types of overlap information offers a relevant contribution for checking entailment, we compare their discriminative power against the simple lexical overlap baseline we already used as baseline in the experiments reported in Section 6.3. We compute frame, role and filler overlap on the gold standard and the output of Shalmaneser on the same corpus.

Results

Results for the simple algorithm are displayed in Table 6.11 and 6.12. Results for *frame* overlap (Table 6.11) show that indeed the gold standard offers a notable discriminative power of 14% between positive and negative examples. Compared to the baseline of lexical overlap, frame overlap shows a slightly higher discriminative power (+10.5% vs. +14%). Incidentally, we also tested the effect of including some frame relations in a limited and controlled manner (allowing "Inheritance", "Subframe", and "Perspective_on" relations with a maximal path length of four), but the figures did not change notably.

The discriminative power of *role* overlap is lower. On the gold standard, it is at 8.7%, slightly lower than the lexical baseline. The results for strict *role filler* mismatch displayed in Table 6.12 is almost neglectable. On the gold standard, it is at 2.9% difference. The use the shallow Levenshtein measure improves the results to 5.2%.

Discussion of Results

Results for frame and role overlap are not unexpected, as both positive and negative pairs should by and large talk about the same situation and type of participants. In theory, the most significant indication as to whether entailment holds should be observed on the level of filler match and mismatch. Yet, the results for filler mismatch do not confirm this intuition. The slight improvement from strict to Levenshtein match indicates that further elaborating the comparison of role fillers is an urgent and promising challenge for future research.

Regarding automatic analysis, results show that frame overlap computed using Shalmaneser's analysis does not have a discriminative power higher than the baseline (+10% vs. +10.5% difference). This is due to the fact that the output of Shalmaneser is noisy. Similar results are obtained by the Shalmaneser data on role and filler match, revealing an informational gap existing between the information produced by a gold standard annotation and an automatic parser.

Overall,the goal of the experiment was to find out (i) whether processing high quality frame information as provided by a gold standard would be effective within a simple algorithm for modeling entailment, and (ii) whether straightforward modeling of this semantic information in an overlap-based architecture is sufficient.

Results show that this way of modeling frame semantics does not make optimal use of the information contained in the dataset. Strong evidence to discriminate between negative and positive examples cannot be found in this straightforward manner. In the light of these experiments, it is not surprising that the full SALSA RTE architecture, working on automatic frame semantic analyses, does not perform significantly better than the lexical baseline. Future work will be necessary on the one hand to derive more elaborate processing models for frame semantic information and on the other hand to improve the quality of the automatic frame semantic analyses.

6.5. Summary and Discussion

In this chapter, we have evaluated the performance of the SALSA RTE system. We also addressed the general question to what extent frame semantic information modeling predicate-argument structure contributes to the automatic recognition of textual entailment.

Results achieved by the system in RTE challenges were promising and inspection of example analyses has shown that the grammatical and frame semantic knowledge has neatly been integrated. Still, in experiments to identify factors that are most relevant for entailment, features from the level of frame semantics did not occur in high ranks. Instead, the shallow level of predicate overlap scored best, followed by the level of

grammatical functions. While this can be taken as evidence that the level of predicate-argument structure should also be relevant, the only high-ranked feature containing frame semantic information is average match cluster size, representing a combination of grammatical, semantic and ontological information.

The fact that features from different "deeper" and more "shallow" levels were ranked highly in feature selection gave rise to the assumption that they are largely independent. We systematically evaluated the performance of the full, "informed" system against and in combination with a "shallow" lexical overlap measure. From these experiments, it is evident that beating the lexical overlap baseline is still difficult for the more informed features. In fact, both seem to model more or less the same information. Another result is that the variance in accuracy is high across different training and test conditions for all features. This confirms observations of other RTE groups and can be best explained with the limited size of the corpora and with idiosyncrasies. It indicates that conclusions derived from measuring system accuracy on the given corpora have to be taken with care.

In order to determine the performance and contribution of those systems components that relate to frame semantic modeling, we designed a gold-standard corpus and performed a number of experiments with it. Three main results are these.

First, the *coverage* of the Berkeley FrameNet lexicon is unexpectedly good for the RTE data. Thus, it is unlikely that coverage is a relevant factor limiting the applicability of frame semantics in real settings.

Second, performance of the linguistic analysis components, especially the *quality of frame semantic analysis* produced by state of the art shallow semantic parsers is comparably low on RTE data. This is mirrored in the limited discriminative power of Shalmaneser annotations over positive and negative entailment examples, especially when compared to a lexical overlap baseline. Also, the accuracy of the SALSA system's interface between LFG f-structure and frame semantics needs to be improved.

Third, the task of *entailment modelling* has emerged as the major factor limiting the applicability of frame semantics in RTE. We proved that simple overlap strategies at the frame semantic level –such as computing frame, role, and filler overlap– do not perform well. While overlap strategies have been applied with some success at the lexical and syntactic level in other studies, we showed that this is not the case for frame semantics. Any system for RTE seeking to use FrameNet or similar role-semantic resources, should model such knowledge in more sophisticated ways, where the richness of frame semantics can be made explicit and exploitable. We believe that research related to the FrameNet hierarchy, e.g., defining frame similarity measures, is most promising for the near future. We will give pointers in this direction in Chapter 7.

Part III.

Further Directions and Conclusions

7. Towards Integration of FrameNet's Hierarchy

In previous chapters, we argued that further achievements of frame-based inference architectures, e.g., for checking textual entailment, depend on *automatic* access of information coded in FrameNet's hierarchy[1]. By human inspection, this information can easily be accessed, interpreting different types of frame relations in suitable ways. For example, it is possible to establish links between (7.1) and both (7.2) and (7.3).

(7.1) John bought a car from Mary for 10.000$. (COMMERCE_BUY)

(7.2) John got a car. (GETTING)

(7.3) Mary got 10.000$. (GETTING)

The relation between (7.1)-(7.2) is straightforward – COMMERCE_BUY inherits from GETTING and the argument positions map correctly. Establishing a link between (7.1)-(7.3) with correct argument fillers is possible, yet more intricate. The shortest path between both frames consists of five steps, requiring moving up and down in the FrameNet hierarchy along three different relation types. As concerns automation, we are not aware of substantial research on computational similarity measures for frames or frame semantic annotations. This is especially surprising in light of the rich literature available for related resources like WordNet.

At the same time, larger frame "scenarios" are predestined to be used for gathering information provided within multi-sentence text fragments. This is increasingly important for modeling textual entailment as the trend goes towards longer, more realistic texts such as (7.4).

(7.4) T: In the first trial in the world in connection with the terrorist attacks of 11 September 2001, **the Higher Regional Court of Hamburg** has passed down the maximum sentence. **Mounir al Motassadeq** will spend 15 years in prison. The 28-year-old Moroccan was found guilty as an **accessory to murder** in more than 3000 cases. (TRIAL, SENTENCING,PRISON, ASSISTANCE ...)

(7.5) H: **Mounir al Motassadeq** faced a trial at **the Higher Regional Court of Hamburg** for **accessory to murder**. (TRIAL)

[1] As in the case of, e.g., WordNet, the term *hierarchy* is used in a wider sense. In fact, not all relations are hierarchical.

In order to establish entailment in (7.4)-(7.5), it is necessary to combine information provided within four different frame instances in (7.4) to cover the single frame instance in (7.5) and all its role fillers (the respective frames are underlined and the role fillers are printed in boldface). Although the idea of using frame semantics for discourse related tasks is not new (Fillmore, 1977; Fillmore and Baker, 2001), little research has been conducted in this area so far.[2]

In this chapter, we want to explore the potential of using FrameNet's hierarchy. In Section 7.1, we will report issues that arise if one tries to make use of the FrameNet hierarchy. The aim of this section is mainly to give pointers for future research. In Section 7.2, we will present a case study of how to use frame semantics for providing loosely connected meaning representations of larger discourse. In Section 7.3, we will shortly wrap up central points.

7.1. Issues Related to FrameNet's Hierarchy

For checking textual entailment, as for many other natural language processing tasks, it is necessary to measure *semantic similarity*, e.g., between known classes of entities and a new entity occurring in a text. Already in the case of nouns –which often refer to relatively concrete objects– it is not easy to define one comprehensive measure. This is proven by the vast number of such measures which have been defined for WordNet. A main difficulty is to find appropriate ways of delimiting concepts or concept clusters in the face of different levels of granularity of sense distinctions that can be applied.

Modeling semantic similarity for events is a challenge as they form a heterogeneous class. In FrameNet, we find frames that have different ontological status in that they describe, e.g., complex events (COMMERCIAL_TRANSACTION), sub-events hereof (COMMERCE_GOODS-TRANSFER), linguistic perspectives on events (COMMERCE_BUY), or partial aspects of events (RECIPROCALITY). Additional complexity is introduced by the fact that semantic roles (and possibly fillers) have to be taken into account as well. This overall complexity surfaces in FrameNet in the existence of many types of frame relations, which are to be interpreted in different ways.

Below, we discuss issues of automatically interpreting FrameNet's hierarchy (Section 7.1.1). In Section 7.1.2, we give reason for some structural issues, namely a tension between linguistic and cognitive modeling that can be observed in FrameNet.

7.1.1. Impact of Different Relation Types

FrameNet defines a variety of frame relations to specify how frames relate to other frames. To see why different relation types can be an issue for natural language processing, consider Figure 7.1, which displays a number of frames in the context of the COMMERCIAL_TRANSACTION scenario. Definitions of frames relevant for our discussion are given in Table 7.1.

[2] Semantic role information in general is not used much for discourse processing. An exception is Ponzetto and Strube (2006), who use PropBank roles for anaphora resolution

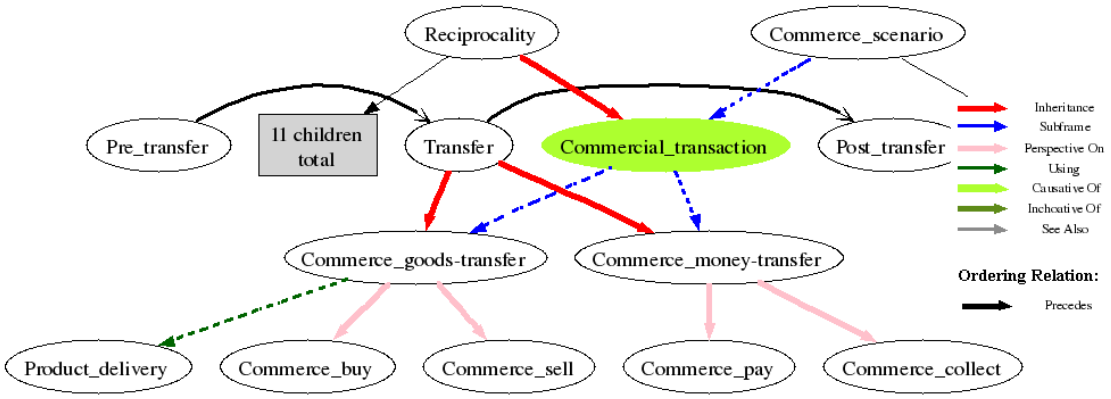


Figure 7.1.: Frame relations of COMMERCIAL_TRANSACTION (screenshot of FrameGrapher).

Just to illustrate potential problems, we try out a naive way of determining the similarity of two frame-annotated sentences by following frame relations in a straightforward fashion. This approach is comparable to taking path length as indicator for the similarity of different synsets in WordNet. If we do so without awareness of different frame relation types, we arrive at contradictory results. Consider the two sentences (7.6) and (7.7). They instantiate the frames COMMERCE_BUY and COMMERCE_PAY, respectively and convey roughly the same meaning.

(7.6) $[\text{John}]_{\text{BUYER}}$ bought $[\text{a car}]_{\text{GOODS}}$ $[\text{from Mary}]_{\text{SELLER}}$ $[\text{for 10.000\$}]_{\text{MONEY}}$.
(COMMERCE_BUY)

(7.7) $[\text{John}]_{\text{BUYER}}$ paid $[\text{Mary}]_{\text{SELLER}}$ $[\text{10.000\$}]_{\text{MONEY}}$ $[\text{for a car}]_{\text{GOODS}}$.
(COMMERCE_PAY)

The frames instantiated by the example sentences can be found on the bottom of Figure 7.1. Both frames are connected via Perspectivization links to COMMERCE_GOODS-TRANSFER and COMMERCE_MONEY-TRANSFER, respectively. On this level, we can follow two different types of links upward – subframe links to COMMERCIAL_TRANSACTION (CT) and inheritance links to TRANSFER (TR). The role fillers that we end up following the different paths are displayed in Table 7.2. For example, for (7.6) the SELLER role of the super-frame COMMERCIAL_TRANSACTION would be instantiated with *Mary*.

If we consider both sentences, we can reach four different super-frame pairs (CT-CT, CT-TR, TR-CT, TR-TR). Unsurprisingly, if we choose the commercial transaction scenario for both sentences (CT-CT), their frame and role analysis is identical (see Table 7.2 on the left). With this choice, we can confirm the intuition that both sentences are similar in meaning.

Frame	Definition and core roles
TRANSFER	This frame involves a Donor transferring a Theme to a Recipient. Core roles: DONOR, RECIPIENT, THEME, TRANSFERORS
COMMERCIAL_TRANSACTION	These are words that describe basic commercial transactions involving a Buyer and a Seller who exchange Money and Goods. [...] Core roles: BUYER, GOODS, MONEY, SELLER
COMMERCE_GOODS-TRANSFER	The subframe of the Commercial_transaction in which the Seller gives the Goods to the Buyer (in exchange for the Money). (Non-lexical frame) Core roles: BUYER, EXCHANGERS, GOODS, MONEY, SELLER
COMMERCE_MONEY-TRANSFER	The subframe of the Commercial_transaction frame which involves the transfer of Money from the Buyer to the Seller (in exchange for the Goods). (Non-lexical frame) Core roles: BUYER, EXCHANGERS, GOODS, MONEY, SELLER
COMMERCE_BUY	These are words describing a basic commercial transaction involving a buyer and a seller exchanging money and goods, taking the perspective of the buyer. Core roles: BUYER, GOODS
COMMERCE_PAY	This frame involves agents paying MONEY for GOODS. [...] Core roles: BUYER, GOODS, MONEY, RATE, SELLER
GETTING	A Recipient starts off without the Theme in their possession, and then comes to possess it. Although the Source from which the Theme came is logically necessary, the Recipient and its changing relationship to the Theme is profiled. Core roles: RECIPIENT, THEME

Table 7.1.: Frame definitions and core roles.

However, all other possibilities would not permit to confirm this intuition. In two cases (CR-TR, TR-CR), the frames are more or less unrelated. This neither indicates similarity nor dissimilarity of the sentences. In the last case (TR-TR), the frames are identical but the role analyses are incompatible as can be seen in Table 7.2 on the right. This choice strongly suggest that both sentences are dissimilar in meaning. While this is true regarding the transfer aspect ("who gets what"), it is evidently not true for both sentences in general.

An open research question is how to (automatically) (i) find appropriate paths through the hierarchy in a given constellation and (ii) how to interpret them. The problem is manifest. Coming back to our initial example ((7.1) through (7.3), repeated below), the link between COMMERCE_BUY from (7.8) and GETTING from (7.10) can be established

	Commercial_transaction		Transfer	
	(7.6)	(7.7)	(7.6)	(7.7)
Mary	SELLER	SELLER	DONOR	RECIPIENT
John	BUYER	BUYER	RECIPIENT	DONOR
(for) a car	GOODS	GOODS	THEME	
(for) 10.000$	MONEY	MONEY		THEME

Table 7.2.: Super-frames for (7.6) and (7.7).

as follows (see also Figure 7.1).

(7.8) John bought a car from Mary for 10.000$. (Commerce_buy)

(7.9) [John]$_{\text{Recipient}}$ got [a car]$_{\text{Theme}}$. (Getting)

(7.10) [Mary]$_{\text{Recipient}}$ got [10.000$]$_{\text{Theme}}$. (Getting)

Commerce_buy perspectivizes Commerce_goods_transfer, which is a subframe of Commercial_transaction. The latter has another subframe Commerce_money_transfer, which is perspectivized in Commerce_collect (*"Mary charged 10.000$"*). This frame finally inherits from Getting (not displayed in Figure 7.1) such that the role fillers are instantiated as in (7.10). Comparing this comparably long path moving up and down the hierarchy with the direct inheritance between Commerce_buy from (7.8) and Getting as instantiated by (7.9), it is by no means obvious how to arrive at the knowledge that the "semantic distance" between (7.8) and both (7.9) and (7.10) is roughly the same.

From inspection of examples, one can speculate that some order of precedence of the different types of relations might be involved. For example, it seems that the subframe relation "blocks" inheritance under certain conditions. In the general case, i.e., for frames which are not part of a scenario, connectedness via the inheritance relation plus role compatibility should provide reliable evidence that the situations described are compatible. The Uses relation and in particular Perspectivization seem to behave like inheritance in this respect. In the case of Causation, it is much easier to provide an interpretation in a computational way. For example, an indicator to establish entailment can be that the "consequent frame" is the resulting state of the action mentioned in the "antecedent frame".

Another open question concerns the impact of the directness of relations when traversing the hierarchy. While, e.g., brother terms in the WordNet hierarchy like *rocket* and *skibob* are typically relatively dissimilar, hyponyms and hypernyms are usually more similar. The example above indicates that this might be different for certain relations in FrameNet. Still, as has been shown for other hierarchies (e.g. in Čulo, 2006, for the case of GermaNet), the contrastive features between classes do not apply to all classes and levels of the hierarchy in the same way. Within FrameNet, e.g, the two frames

COOKING_CREATION and DUPLICATION, which inherit from INTENTIONALLY_CREATE are comparably dissimilar. In contrast, the frames INHIBIT_MOVEMENT and ARREST inherit from INTENTIONALLY_AFFECT and are much more compatible.

As we cannot provide a comprehensive treatment of frame relations within the scope of this thesis, we have to leave the discussion at that.

7.1.2. Tension between Cognitive and Linguistic Modeling

FrameNet can be seen as being an "interface" between lexical semantics (capturing predicate meaning) and knowledge representation (capturing information about typical situations). As a consequence, sometimes a divergence can be observed between the often defeasible character of natural language semantics one the hand and the typically stricter information contained in idealized, abstract models of the world on the other hand.

This divergence is mirrored in the question whether inheritance describes a conceptual relation or a relation which is tied more closely to natural language. In Fillmore et al. (2001), inheritance was regarded more on a conceptual level, e.g., to describe so-called *profiling*.[3] An example discussed is the profiling of the place of departure in a DEPARTING frame, which inherits from the more general MOTION frame. Moreover, multiple inheritance ("frame blending") had originally been proposed to relate frames by way of a "semantic decomposition". For example, JUDGMENT_COMMUNICATION was conceived as a blend of both JUDGMENT and COMMUNICATION (now: STATEMENT). This had been expressed via multiple inheritance.

However, in the current FrameNet database, both inheritance relations have been replaced by the "weaker" Uses relation, which is harder to utilize in natural language processing for its unclear semantics. Ruppenhofer et al. (2006) give as one reason for changing from inheritance to Uses that the (overtly realizable) role inventory of the frames is partly incompatible – while, e.g., JUDGMENT_COMMUNICATION can realize an EVALUEE and a REASON, STATEMENT realizes the same content as a single MESSAGE, as can be seen in (7.11) and (7.12).

(7.11) You accused [me]$_{\text{EVALUEE}}$ [of bluffing]$_{\text{REASON}}$. (JUDGMENT_COMMUNICATION)

(7.12) You said [I was bluffing]$_{\text{MESSAGE}}$. (STATEMENT)

This example illustrates different rationales for establishing frame relations. One is driven by conceptual considerations of whether the situations described by one frame are more specific than the ones described by other frames and the other one is driven by observations about possible and impossible linguistic realization patterns of semantic roles.

[3] In Narayanan, Baker, Fillmore, and Petruck (2003), FrameNet members state to have borrowed the term *profiling* from Langacker (1987), who distinguishes between *profile* and *base*. The profile refers to the meaning of the word under consideration itself while the base refers to encyclopedic knowledge needed to understand it.

"Lexicographic" hierarchy (current)	Hierarchy determined purely by roles	Incorporated role is made explicit
INSTANCE	SOLE_INSTANCE	INSTANCE
INSTANCE	ITEM	INSTANCE
TYPE	TYPE	TYPE
INSTANCE_PROP		INSTANCE_PROP
↑ Uses	↑ Inherits_from	↑ Inherits_from
SOLE_INSTANCE	INSTANCE	SOLE_INSTANCE
ITEM	INSTANCE	ITEM
TYPE	TYPE	TYPE
	INSTANCE_PROP	+ITEM_PROP: sole

Table 7.3.: Towards a more uniform hierarchy.

From the perspective of practical natural language processing, an open question is whether it is possible to (automatically) arrive at a FrameNet hierarchy that is easier to utilize. As concerns the Uses relation, it is conceivable how a simpler, stricter version of FrameNet could look like. We sketch one idea for re-structuring the hierarchy by means of the two frames INSTANCE and SOLE_INSTANCE, that currently stand in a Uses relation. The frame INSTANCE describes "transparent nouns which denote instances of types of entities or events" and SOLE_INSTANCE describes adjectives which describe items as being the only instance of the given type. Example annotations are found in (7.13) and (7.14).

(7.13) [This algorithm]$_{\text{INSTANCE}}$ is a [typical]$_{\text{INSTANCE\_PROP}}$ instance [of bottom up clustering]$_{\text{TYPE}}$. (INSTANCE)

(7.14) [Steven]$_{\text{ITEM}}$ is the sole [survivor]$_{\text{TYPE}}$ in his family.[4] (SOLE_INSTANCE)

Both frames with their full role inventory are displayed in Table 7.3 in the left column. The role SOLE_INSTANCE.ITEM inherits from INSTANCE.INSTANCE, TYPE is inherited without renaming. The linguistic reason why SOLE_INSTANCE cannot inherit from INSTANCE is that the instance property (of being sole) is absorbed (or incorporated) by the adjectives evoking it. Consequently, the role inventory of SOLE_INSTANCE lacks a role corresponding to INSTANCE_PROP. As inheritance would require that all roles of the super-frame (INSTANCE) are also available at the sub-frame (SOLE_INSTANCE), it is not applicable here. Yet, SOLE_INSTANCE is still construed as being a related to (and probably more specific as) INSTANCE. To capture this, the Uses relation has been annotated.

[4]The annotation is from FrameNet, we would have included *in his family* in the TYPE as well, as this would result in a unique description.

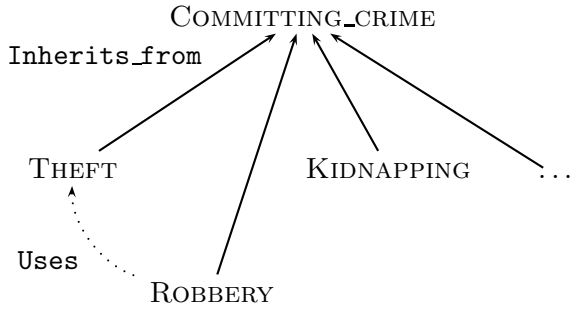


Figure 7.2.: Frames inheriting from COMMITTING_CRIME.

Again, the tension here is that while it is conceptually arguable that SOLE_INSTANCE is more specific than INSTANCE, the respective linguistic role inventories do not allow to establish an inheritance relation between both.

Still, for automatic natural language processing, it would be desirable to have a dense annotation of FrameNet's inheritance relation, as is the case in WordNet. moreover, as non-monotonic inheritance structures tend to complicate models and processing architectures, it would be preferably to use proper inheritance if possible. Two possibilities for re-structuring FrameNet accordingly are displayed in Table 7.3 in the middle and right column. The solution in the middle may seem strange as it swaps both frames, defining INSTANCE as subframe of SOLE_INSTANCE. This solution implements the idea that the hierarchy is based solely on the available role inventories.[5] An advantage would be that this solution could be implemented fully automatically. However, as it leads to a conceptually inadequate hierarchy –not every INSTANCE is a SOLE_INSTANCE– this is not a convincing option. Our preferred solution is the one displayed in the right column of Table 7.3. Here, the incorporated role INSTANCE_PROP is made explicit within SOLE_INSTANCE. It is added (re-named to ITEM_PROP) and annotated with an appropriate semantic type. More research would be needed to check whether this method is applicable in general.

FrameNet itself in some cases addresses comparable problem caused by incorporated roles using the special role type *core unexpressed*, which introduces non-monotonicity. We will illustrate it with an example. Table 7.4 shows the roles of the four frames displayed in Figure 7.2 and contrasts the respective role types (core, non-core, extrathematic, core unexpressed, see Section 3.3.4).

The top row displays the roles that are identical throughout all four frames. The middle row shows the roles that are added monotonically by one of the daughter frames or that change their role type during inheritance. The last row shows a case of an incorporated role – the crime itself, which is a core role of COMMITTING_CRIME, is not applicable at either THEFT or ROBBERY (although KIDNAPPING lists an EVENT role, we

[5] We disregard the question as to whether the role inheritance of INSTANCE from ITEM is justifiable.

		Comm._crime	Theft	Robbery	Kidnapping
Identical roles	frequency	nc	nc	nc	nc
	manner	nc	nc	nc	nc
	means	nc	nc	nc	nc
	perpetrator	c	c	c	c
	place	nc	nc	nc	nc
	purpose	nc	nc	nc	nc
	reason	nc	nc	nc	nc
	time	nc	nc	nc	nc
Monotonic role extension/ Change of coreness	instrument	ext	nc		
	source		c	c	nc
	goal				nc
	victim		c	c	c
	goods		c	nc	
	degree				nc
	duration				nc
	manner				nc
	means				nc
Role deletion	crime	c-u			nc (Event)

Table 7.4.: Frame element inheritance (c=core, c-u=core-unexpressed, nc=non-core, ext=extrathematic).

did not find any annotated sentences providing an example realization).

Technically, inheritance of the crime role is blocked by annotating this role as core-unexpressed on the level of Committing_crime. A disadvantage of this solution is that the inheritance relation is occasionally turned into a non-monotonic relation. This is not only theoretically unattractive, but it might also lead to computational issues if standard methods cannot be applied. Again, we would prefer to make the incorporated roles explicit in the daughter frames. Incidentally, in German, the crime role can easily be realized within a Theft frame as in (7.15).

(7.15) Gestern stahlen [drei Maskierte]$_{\text{perpetrator}}$ [bei einem Überfall im Landesmuseum]$_{\text{crime}}$ [drei wertvolle Gemälde]$_{\text{goods}}$. (Theft)

Table 7.4 also shows that the roles of Theft and Robbery, which stand in a Uses relation (see Figure 7.2), are identical with the exception that the latter frame does not have an instrument role. At the same time, one can find attestations like *he robbed it with a chain gun* in the WWW. If the missing instrument role is the only reason why Robbery does not properly inherit from Theft, we would suggest to adapt these frames in order to arrive at a more uniform inheritance structure. One solution

could be to annotate the instrument as extrathematic (as has been done in the case of COMMITTING_CRIME, from which ROBBERY inherits).

All in all, the current FrameNet hierarchy is not only incomplete, it is also idiosyncratic in places. Originally designed for lexicographic research, it should partly be re-structured to become more committing and thus easier to be used within automatic natural language processing. Ideally, the completeness problem could be addressed, e.g., by focusing on a thorough annotation of inheritance relations. It needs to be determined, to what extent this process could be performed (semi-)automatically.

Abstracting away from these issues that occur when two frames or single sentence analyses are compared, the FrameNet hierarchy bears further potential for the analysis of larger discourse. In the next section, we will demonstrate how this potential can practically be used.

7.2. Building Text Meaning Representations from Contextually Related Frames

A natural advantage of FrameNet lies in its capability of describing larger scenarios, thus potentially covering the meaning of multiple sentences ("discourse"). The relevance of frame semantics for text linguistics has already been pointed out by Fillmore in an early work on frame semantics (Fillmore, 1977, p.4):

> Successful text analysis has got to provide an understanding of the development of an image or scene [...]: The first part of the text activates an image or scene [...]; later parts of the text fill in more and more information [...].

Apart from a more programmatic discussion in Fillmore and Baker (2001), we are not aware of any substantial research in this area. In this section, we investigate frame annotations *in context* as a partial text meaning representation on a fairly concrete level. We present a case study where we investigate different types of relations between frames when assigned to contiguous portions of text –contextual relations from syntactic parsing and frame relations from FrameNet– and show how specific patterns of such relations support the inference of co-referential links between frames. We discuss possibilities of using learning techniques to induce links between frames.

We proceed as follows. Section 7.2.1 presents an outline of our investigations into frame-based text meaning representation. We discuss how to connect frame annotations to obtain an interlinked (yet partial) semantic representation. In Section 7.2.2, we then present a worked-out example that illustrates how specific configurations of lexico-semantic and contextual relations can license the induction of co-referential links between frames. We discuss how this process can be generalized and automated in Section 7.2.3.

7.2.1. Frame Semantics for Partial Text Meaning Representation

As frame semantic descriptions are focusing on open class categories (verbs, nouns, adjectives), full text annotations are necessarily partial. By applying frames to contiguous

portions of text –due to the lack of constructional "glue"– we obtain partially connected predicate-argument structures in a network of potential frame-to-frame relations. In order to construct a more densely connected frame-based text meaning representation, we need to infer additional links between frames and frame elements. For this we can exploit contextual relations between frames and frame elements as given by syntactic parsing, e.g., structural embedding or adjacency relations between neighboring frames. When trying to induce contextually linked frames, we have to distinguish two levels:

1. The level of frame *instances*, where we can infer co-reference of events or role fillers, and

2. the level of *types*, where we can infer intrinsic relations between frames and roles that are not yet included in the FrameNet graph.

At the instance level, we can establish co-referential links between, e.g., a filled role of one frame instance with an unfilled role of another frame instance provided we find sufficient supporting evidence. Two roles can be linked, e.g., if –at the type level– the respective frames stand in a subframe relation and in addition, the frame instances are contextually related, e.g., by functional-syntactic embedding or else by way of a discourse relation.

At the type level, we can induce relations between frames or roles on the basis of, e.g., recurrent anaphoric linking patterns observed in texts. The induction of meaning relations at the type level is more involved and requires use of annotated corpora and learning techniques. In both cases, the induction of co-reference relations between frames can only be heuristic.

7.2.2. Frames in Context – A Case Study

In this section, we present a case study that establishes systematic patterns of lexical-semantic and contextual relations that support the induction of co-referential relations between frames and roles. As an example we chose a short news wire text[6] that pertains to the "scenario frame" CRIMINAL_PROCESS:

(7.16) In the first trial in the world in connection with the terrorist attacks of 11 September 2001, the Higher Regional Court of Hamburg has passed down the maximum sentence. Mounir al Motassadeq will spend 15 years in prison. The 28-year-old Moroccan was found guilty as an accessory to murder in more than 3000 cases.

Table 7.5 lists all target predicates, frames and roles that are relevant for the example. Role fillers that correspond to local constituents are displayed in the right column in italic (non-bold) type font, e.g., TRIAL.CASE is filled by the constituent *terrorist attacks*. Roles that cannot be associated with any constituents are not displayed or left unfilled (e.g., ATTACK.VICTIM). Frame element fillers and co-references between frame elements that

[6] http://www.germnews.de/archive/dn/2003/02/19.html

Target	Frame	Frame element	Filler (*given* vs. ***induced***)	
trial	TRIAL	CASE	*terrorist attacks*	(1)
		CHARGE	***accessory to murder***	(2)
		COURT	***Higher Regional Court***	(3)
		DEFENDANT	***28-year-old Moroccan***	(4)
attacks	ATTACK	ASSAILANT	*terrorist*	(5)
		VICTIM		(6)
		TIME (non-core)	*11 September 2001*	(7)
sentence	SENTENCING	CONVICT	***Mounir al Motassadeq***	(8)
		COURT	*Higher Regional Court*	(9)
		TYPE	*maximum sentence*	(10)
prison	PRISON	INMATES	*Mounir al Motassadeq*	(11)
		DURATION (exth.)	*15 years*	(12)
found guilty	VERDICT	CASE	***terrorist attacks***	(13)
		CHARGE	*accessory to murder*	(14)
		DEFENDANT	*28-year-old Moroccan*	(15)
		FINDING	*guilty*	(16)
accessory	ASSISTANCE	CO-AGENT		(17)
		FOCAL_ENTITY	*murder*	(18)
		HELPER	***28-year-old Moroccan***	(19)
murder	KILLING	KILLER		(20)
		VICTIM	*more than 3000 cases*	(21)

Table 7.5.: Frame annotations with given/inferred frame element linkings.

can be induced on the basis of frame relations, contextual relations or bridging inferences are displayed in boldface. For example, *Higher Regional Court* that originally only fills the role SENTENCING.COURT can be induced to be filler of the role TRIAL.COURT as well.

Frame Relations

The frames evoked in the example pertain to the following frame relations. Both SENTENCING and TRIAL are subframes of CRIMINAL_PROCESS. VERDICT in turn is a subframe of TRIAL. These frame relations are displayed (among other things) in Figure 7.3 by straight lines. Additionally, ASSISTANCE inherits from INTENTIONALLY_ACT.

Contextual Relations

The example features different types of contextual relations between frames and roles such as functional syntactic embedding, surface order, discourse relations, or co-reference. In Figure 7.3, central relations are displayed by dashed lines. For example, SENTENCING

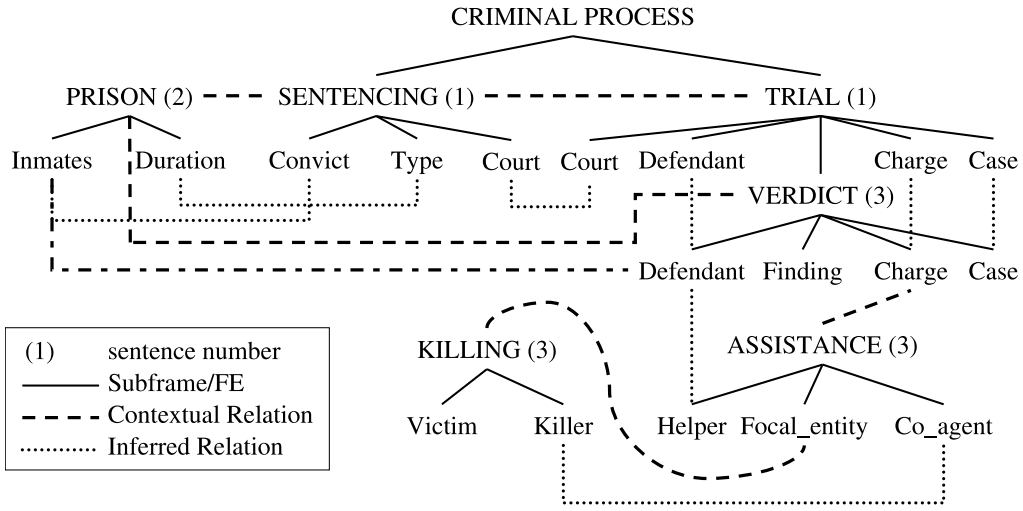


Figure 7.3.: Frame relations, contextual relations, and inferred relations.

and TRIAL are syntactically related by functional (adjunct) embedding; the KILLING frame is embedded within the FOCAL_ENTITY role of ASSISTANCE; the sentence projecting PRISON follows, and stands in a discourse relation (elaboration) to the sentence projecting SENTENCING. Finally, the referents corresponding to the roles PRISON.INMATES and VERDICT.DEFENDANT can be recognized as co-referent.

Inferred Relations

Based on these lexico-semantic and contextual relations, we can infer further semantic relations between roles and frames, such as co-referential binding of unfilled roles. Figure 7.3 shows the inferred relations that we identified in (7.16) by dotted lines. Closer study of the inferred relations reveals a number of underlying patterns of justifications, which we will exemplify in turn. In the majority of cases, we can infer role bindings on the basis of (a variety of) patterns of lexical semantic and contextual relations between frames and roles. In some cases, further lexical semantic knowledge is required, which is not yet encoded in FrameNet, such as "semantic control" between frames and elements. We will later discuss an example which motivates that additional semantic information, such as referential and temporal properties, needs to be considered for inducing role bindings.

Inference on instance level. Figure 7.4 schematically illustrates a pattern in which we induce an instance link between role fillers like TRIAL.COURT (r_1) and SENTENCING.COURT (r_2) (see (3) and (9) in Table 7.5). In the figure, frame (and role) *types*

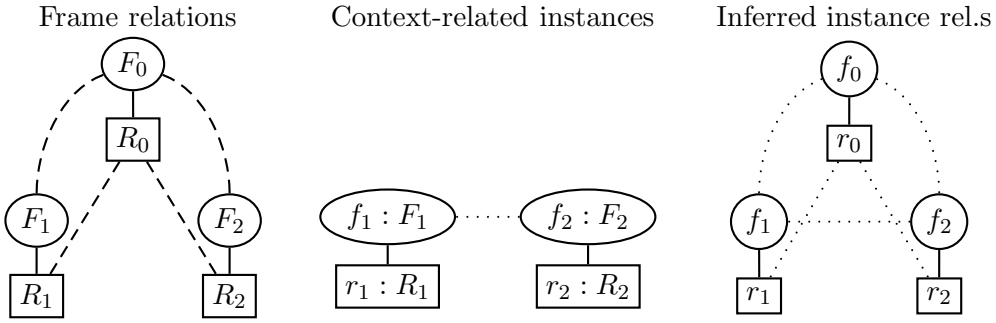


Figure 7.4.: Inferring instance relations.

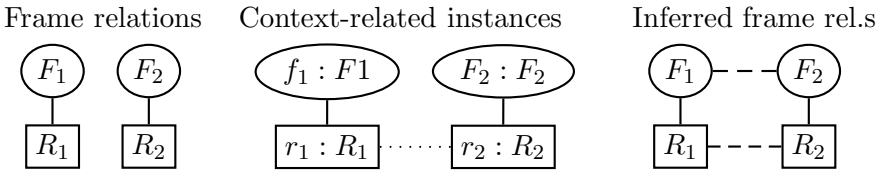


Figure 7.5.: Inducing frame relations.

are printed in upper case, *instances* in lower case. $f_1 : F_1$ means that f_1 is an instance of frame F_1. In this concrete example, SENTENCING (F_2) and TRIAL (F_1) are subframes of CRIMINAL_PROCESS (F_0) and the roles R_1 and R_2 both inherit from CRIMINAL_PROCESS.COURT (R_0). These type level inheritance and subframe relations are displayed by dashed lines (left). In addition, the given frame instances (f_1, f_2) stand in a functional (adjunct) embedding relation. This contextual relation is displayed by dotted lines (middle).

On this basis, we assume that both frame instances are subframes of *the same* CRIMINAL_PROCESS scenario instance (f_0). In other words, a "larger" cluster of frames is instantiated by the observed instances. This allows linking of the roles (r_1) and (r_2) (right) on the basis of type level information. Other examples of role identifications that follow this pattern are (1)-(13), (2)-(14), (4)-(15) in Table 7.5, which are based on the subframe relation between TRIAL and VERDICT.

Inference on type level. Figure 7.5 illustrates a case where role identification is induced on the type level. In this concrete pattern, the support is provided by a contextual co-reference relation. The frames PRISON (F_1) and VERDICT (F_2) are unrelated in FrameNet (left). In the text, the referents of the roles PRISON.INMATES (r_1) and VERDICT.DEFENDANT (r_2) are marked co-referent by means of a definite description (middle). We therefore induce a role identification at the type level by assumption of an "anonymous" frame-to-frame relation that can be further specified, e.g., as a causation relation or a subframe relation within some scenario. Of course, such heuristic inductions

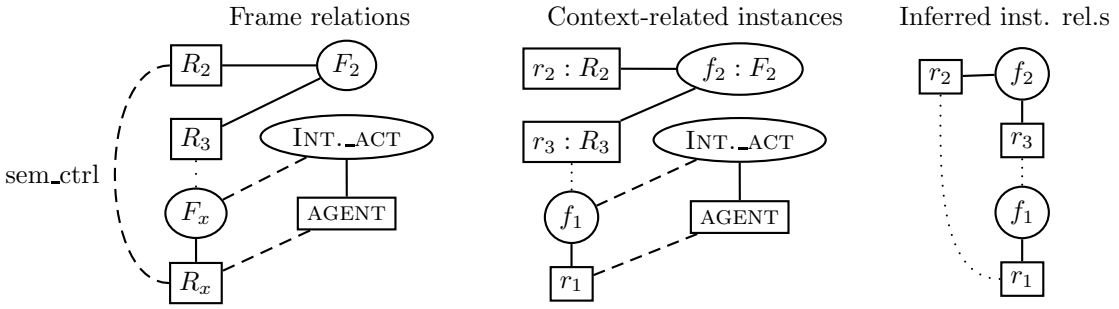


Figure 7.6.: Inferring instance relations (via "semantic control").

cannot be based on a single observed instance, but require recurrent observations of justifications. One application of this technique might be the semi-automatic extension of FrameNet's inheritance structure on the basis of annotated data such as the FrameNet corpus. A requirement on the corpus is that either more than one frame is annotated per sentence or that frame annotations are carried out on subsequent sentences.

Lexical meaning postulates: "semantic control". In some cases, patterns of frame and contextual relations are not sufficient to induce role-identification. We found that sometimes further lexical semantic information is required, in particular what we call *semantic control*, as a kind of meaning postulate. For some frames, it is part of their inherent lexical meaning that a given role must be co-referent with the agent- or patient-like role of an embedded frame. For example, the defendant in a verdict is (found to be) the actor in the event that constitutes the charge of the verdict, as illustrated by the examples below.

(7.17) **He** was convicted for jumping out of a bush naked and making lewd suggestions to a jogger [...].[7] (VERDICT, SELF_MOTION)

(7.18) Defendant Stephanie Mohr has appealed **her** conviction for violating 18. U.S.C. 242.[8] (VERDICT, COMPLIANCE)

(7.19) **She** was found guilty for lying to federal investigators.[9] (VERDICT, PREVARICATION)

This is schematically represented in Figure 7.6 (left). VERDICT (F_2) features semantic control, in that VERDICT.DEFENDANT (R_2) is marked identical to the agent of some frame F_x embedded within its CHARGE role (R_3) (dashed line). Agenthood is formally

[7] From `news.ninemsn.com.au`
[8] From `www.usdoj.gov`
[9] From `www.womensnewsdaily.com`

represented by inheritance from INTENTIONALLY_ACT.AGENT.[10] In our running example (middle), VERDICT.CHARGE (r_3) embeds ASSISTANCE (f_1). Furthermore, ASSISTANCE.HELPER (r_1) inherits from INTENTIONALLY_ ACT.AGENT. We can thus conclude that the filler of VERDICT.DEFENDANT (r_2) is identical to the ASSISTANCE.HELPER (r_1) (right) ((15)-(19), Table 7.5). Other examples that involve semantic control are (17)-(20) and (8)-(11) (the latter assuming a causative relation between SENTENCING and PRISON).

7.2.3. Acquisition of Role-linking Patterns

We have identified various patterns of lexico-semantic and contextual relations that support the induction of co-reference relations between frames and roles. Frame relations proved essential for linking contextually related (neighboring) frame instances. Different types of contextual relations could be observed to support role identification – syntactic and semantic embedding, anaphoricity, and connectedness by discourse relations or surface linearization. More data needs to be investigated to determine the weight of the individual factors. In future work, one could apply statistical methods for acquiring role-linking patterns from analyzed (annotated) text samples of a restricted domain, like CRIMINAL_PROCESS. The aim is to learn weighted role-linking patterns that can be formalized as probabilistic inference rules. For experiments along these lines, see Liakata and Pulman (2004); Lin and Pantel (2001).

We have provided an abstract definition of semantic control in terms of the agent role marked by inheritance from the perspectivizing frame INTENTIONALLY_ACT (the frame INTENTIONALLY_AFFECT additionally defines a patient role). This will facilitate the acquisition of lexical semantic control relations, yet it relies on the full specification of such inheritance relations in the FrameNet data (for the chosen domain). Based on inferred or given role-linkings and subframe relations, one could also learn more involved patterns of "bridging" inferences between frames. For example, in (7.20), *Baragiola* locally fills the role SENTENCING.CONVICT and via an elliptical construction also the role of ESCAPE.ESCAPEE.

(7.20) [Baragiola]$_{\text{CONVICT/ESCAPEE}}$ had previously been convicted of murder in Italy, but had escaped in 1980 and obtained Swiss citizenship. (SENTENCING/ESCAPE)

Assuming that we have already learned a role-linking between SENTENCING.CONVICT and PRISON.INMATES, we can heuristically infer the existence of a PRISON frame, with PRISON.INMATES referentially bound to the ESCAPE.ESCAPEE. In turn, one could try to find more instances of the pattern PRISON-ESCAPE as additional support for the heuristic inference. So, this approach could be used to "bootstrap" densely connected frame scenarios from given annotation.

Finally, examples like (7.21) go beyond the expressivity of the frame semantic representations we used above. In order to treat them, we need to enrich the representations

[10] In fact, not all frames constituting charges in the examples above are (yet) annotated to inherit from INTENTIONALLY_ACT.

with more "deep" semantic information to control the induction of role identification. We need to model referential properties, such as the introduction of new discourse referents (*a new trial*), and event modification by locational or temporal adjuncts. The former will be crucial to define "blocking" factors for role identification rules, the latter will provide deeper semantic characterizations of contextual relations between frames, such as temporal sequence.

(7.21) Mounir El Motassadeq (born April 3, 1974) is a Moroccan. In February 2003 he was convicted [...]. As of April 2004 he is the only person to have been convicted in direct relation to the September 11, 2001 attacks. The verdict and sentence were set aside on appeal [...]. A new trial is expected in mid-2004. (From Wikipedia)

7.3. Summary of this Chapter

This chapter was concerned with the access and use of information encoded in FrameNet's hierarchy. We started from the observation that there exists neither literature on similarity measures for frames or frame annotations nor on the use of frame annotation for discourse modeling – two issues which are relevant for further achievements in using frame semantics for detecting textual entailment.

We then explicated issues that arise from the multitude of existing frame relations, which makes the FrameNet structure much harder to interpret than, e.g., WordNet's inheritance structure for nouns. We gave reasons for why this is the case and sketched first ideas for re-structuring FrameNet in order to adapt it to the needs of automatic processing.

As concerns the use of frames for discourse analysis, we provided a worked-out example that illustrates how actual frame analyses in context can be made more dense by relating frames and roles of "neighboring" frames. The basis for this are on the one hand frames scenarios and on the other hand contextual relations like syntactic embedding or anaphoric reference. In the other direction, we have shown how frequently observed patterns in annotated corpora can be used to heuristically induce new frame relations.

More research on the ideas we presented in this chapter is needed. Once there are results available like similarity measures for frames, we can adjust our RTE system accordingly. Hopefully, this will lead to a more significant contribution of the frame semantic information for natural language tasks like RTE in the near future.

8. Conclusions

In this thesis, we presented the SALSA RTE system, a competitive, large-scale, semantics-based natural language inference system. While implementation of such a system would have been out of reach a few years ago, we are far from proclaiming that all problems have been solved.

When we started to prepare this thesis, broad-coverage syntactic parsers had been available for some time while the traditional, logic-based approaches to computational semantics had proven to be inapt for practical applications. This "semantic gap" had partly been filled by "shallow" approaches, that managed to approximate natural language meaning in prototypal applications with some success. These approaches were often based on distributional measures and the WordNet taxonomy. A natural next step was to explore the contribution of more structured lexical-semantic information in the form of predicate-argument structure, for which semantic parsers had become available.

The newly introduced task of *recognizing textual entailment* (RTE) was attractive for this undertaking for a number of reasons: (i) it subsumes a number of typical applications, (ii) it offers a corpus for realistic, empirical evaluation, and (iii) several studies had shown that the level of predicate-argument structure is relevant for modeling many types of variation and paraphrases encountered in the RTE data.

The framework we developed in this thesis is one of the first that explores the contribution of predicate-argument structure in terms of Fillmore's *frame semantics* for broad-scale natural language inference. Textual entailment is approximated via "informed" features that capture the structural and *semantic overlap* between analyses of (entailing) texts and (entailed) hypotheses. In the implementation, this information is provided by state-of-the-art frame semantic parsers, combined with deep grammatical analysis and a WordNet-based component that deals with aspects of semantic modeling not covered by frame semantics. The SALSA RTE system's performance was on a par with related (non-commercial) systems in last years' RTE challenges.

This "pioneering" study can serve to guide future and in-depth research on both frame-based inference systems and lexical-semantic approaches to textual entailment. Our work offers major contributions in the field of applied lexical semantics because frame-semantic information had neither been integrated in a comparable, broad-scale end-to-end system before, nor had the coverage of Berkeley FrameNet and that of existing semantic parsers been evaluated on realistic data. In the rest of this chapter, we will go through four of the most important challenges we have addressed in this thesis and finally sketch some starting points for future research. The challenges we will discuss are:

- Cope with coverage issues of FrameNet.
- Construct frame-based meaning representations.

- Design a suitable inference framework.
- Evaluate its impact on the RTE task.

Cope with coverage issues of FrameNet. It is known that the Berkeley FrameNet database is incomplete in several respects. Two central types of limitations are: (i) missing frames (senses); and (ii) missing LUs (incomplete lexicon). Another limitation are missing frame relations (see below). These issues potentially restrict the usefulness of FrameNet for tasks such as RTE. One result of this thesis is that we could eliminate the first concern and provide an effective solution for the second:

(i) Missing frames are not an issue for RTE. The topics dealt in RTE corpora such as killing, disaster, and competition events are covered well by FrameNet. As we have shown with a manually annotated gold-standard, more than 90% of the information relevant for entailment can be annotated with existing frames. This somewhat unexpected result can only partially explained with the flexibility we granted human annotators for assigning frames that seem by and large plausible and the strategy of annotating only *relevant* material.

(ii) Statistically trained systems like Shalmaneser completely fail in assigning any frame if they encounter a word that is not contained in the FrameNet lexicon. They also fail if too few annotations of an existing LU are available in the FrameNet corpus for training. We addressed these problems by developing the WordNet-based "Detour (to FrameNet)" system, which guarantees high coverage. The detour practically implements a flexible frame assignment strategy. It has almost 90% recall and a precision of about 50%. This means, in 50% of the cases, where statistically trained systems fail, the correct (gold standard) frame can be assigned by the Detour algorithm. If an approximate notion of frame similarity is taken into account in the evaluation against a gold-standard, adding the Detour system improves both precision and recall as compared to Shalmaneser alone.

While coverage issues of automatic frame and role assignment can be alleviated by the combination of Shalmaneser and the Detour system, we found out that their precision is still far from a manually annotated gold standard. Substantial improvement of frame-semantic analysis tools will be a prerequisite for further achievements in frame-based natural language processing.

For future improvement of Detour's precision, it would be helpful to be able to automatically distinguish words/readings for which a frame exists from those that are definitely not covered by FrameNet. First investigations in this direction are reported in Erk (2006) under the heading *outlier detection*. It would also be desirable to have a better external WordNet word-sense disambiguation component. As WordNet has become a kind of standard reference format for lexical resources, it would in general be good to have an annotation of existing FrameNet's LUs with WordNet synsets. This would considerably facilitate the use of FrameNet in applications.

Construct frame-based meaning representations. Recent research on predicate-argument structure has mostly been concerned with automatic role labeling. One outcome of this research is Shalmaneser, the frame semantic parser used in this thesis. The question how of to integrate this kind of knowledge in natural language processing tasks has largely been untouched. Possible reasons are coverage issues (see above), and also limited robustness of analysis tools.

We decided to address this latter point by combining FrameNet information with grammatical information from a state-of-the-art "deep" syntactic LFG framework. LFG f-structure predicates and grammatical functions can serve as fall-back for cases, where frame or role analyses are not available. Another reason for integrating grammatical information in the SALSA RTE system was that surface and syntactic phenomena make up a good part of textual entailment. We are convinced that it is a good strategy to check entailment on the lowest possible level of analysis in hybrid architectures. While semantic descriptions in theory provide nice abstractions from irrelevant surface, it is evident that semantic processing at the same time introduces noise and has to deal with a number of issues such as limited coverage.

In order to represent grammatical, frame semantic, and ontological information in a common framework, we designed a suitable *meaning representation* format and implemented an *interface* integrating information from the primal analysis components. Grammatical and FrameNet information is integrated by a *semantic projection* from LFG f-structure graphs into frame structures. WordNet knowledge was integrated by way of a third projection called *ontological projection*. The resulting tripartite graphs for text and hypothesis represent all given information in an integrated way while keeping the levels of description apart. This architecture is easily extensible as we have shown by including information from the SUMO knowledge ontology. We also demonstrated how this architecture nicely supports refinement and further normalizations by way of interleaving information from different levels. This does not only lead to more dense frame semantic annotations, but also allows to approximately mark phenomena from compositional semantics like modality, as has been shown.

The technical realization of the interface between basic components (LFG and Shalmaneser/Detour) was difficult. Exhaustive fine tuning was not possible for time reasons. An option for future research to circumvent problems on this basic level would be do devise a direct semantic construction for frame semantics from a major grammar framework like LFG or HPSG. Either in a machine learning setting, as proposed by Frank and Erk (2004), or based on the linking patterns provided, e.g., by FrameNet for English or the semantic lexicon of SALSA (Burchardt, Padó, Spohr, Frank, and Heid, 2008) for German.

Design a suitable inference framework. Frame-based reasoning has also not been discussed in the literature so far. Not even similarity measures for frames or frame annotations are currently available. We designed a frame-based inference framework for textual entailment from scratch. Like related approaches, we approximate entailment via semantic similarity of hypothesis and text. In the SALSA RTE system, this was

implemented by computing directed overlap of meaning representations of hypothesis and text, assuming that the more of the hypothesis is covered by the text, the more indication we have that entailment holds.

Starting from the meaning representations we sketched above, we designed a *graph matching* technique that captures different aspects of structural and semantic similarity of representations for text and hypothesis. We deliberately opted for a declarative, verbose, and partly redundant representation format to be able study, which factors have an impact on the entailment decision. An advantage of these "match graphs" is that they are easily human readable, thus supporting error analysis and correction. The matching information is translated into feature vectors, and the entailment decision is made by a well-established machine learning system, trained on RTE corpora.

Our initial expectation that the machine learners would help establish interesting correlations between the different types of information such as matching frames and unmatched role fillers was not fulfilled. In the evaluation, it turned out that more controlled ways of comparing text and hypothesis might be needed to make better use of information from different layers. For a start, one could try to come up with a more controlled, precise, "rule based" definition of graph matching for a restricted number of examples or phenomena. In the architecture, this could be achieved by defining new, more complex features. Doing so, however, seems to rely on a more comprehensive understanding of the notion of textual entailment.

Concrete improvements for the SALSA system we identified concern its bias towards positive entailment judgments. This tendency should be compensated by introducing more negative features that measure, e.g., the distance –semantic or constructional– between material involved in partial match graphs. More generally, in addition to the determination of similarity, we should improve modeling clues for *dissimilarity*. The detection of incompatible modalities has proved rather effective, but can be further extended to lexically induced modalities (e.g., *possibility of, alleged, promise*).

Evaluate the impact of predicate-argument structure on the RTE task. Results obtained by the SALSA RTE system in RTE challenges were comparable to those of related systems. However, the intuitive appeal of frame semantics has not yet manifested itself in a significant improvement of system results. The most indicative features for entailment used by the machine learners stem from the layer of grammatical analysis. One reason for this result is certainly sparsity of training data. As we are working with statistical methods, corpora in the magnitude of 1000 training examples are definitely a challenge in itself. Only features that occur very frequently are considered by the learners.

Results concerning the *potential* of frame annotation to discriminate between positive and negative entailment are still ambivalent. We have shown that even a high-quality manual frame annotation manages to model textual entailment only slightly better than the lexical baseline. Pure frame annotation of text and hypothesis and comparison in a straightforward manner obviously does not suffice for achieving a significant effect. As text and hypothesis should talk about more or less about the same situation in positive

and negative examples, it is evident that the semantic normalization provided by pure frame annotation is not sufficient for distinguishing entailment from non-entailment. Why inclusion of (heuristic) measures for role and filler overlap did not exhibit a better discriminative power in our experiments still has to be determined. Intuitively, it should more often the case that role fillers between text and hypothesis differ in negative pairs than in positive ones.

An overall problem is the absence of similarity measures for frames and frame annotation instances. This problem could naturally be resolved making use of frame relations. Yet, to date, the number of annotated instances is not only relatively sparse, it is also largely open how existing frame relations can be automatically interpreted. System results as well as experiments with the FATE corpus suggest that principled access to frame relations are a necessary requirement for future improvements in frame-based inferencing.

Future Directions

Below, we want to sketch three directions for future research we find promising. The first directly concerns further development of systems similar to the SALSA RTE system. The second addresses a weakness in the setup of the RTE challenge. The third identifies a future topic beyond textual entailment on the way towards intelligent automatic information access.

Focus on hard examples. Lexical (surface) overlap often serves as baseline for checking textual entailment. On the available RTE corpora, accuracy of such a baseline is in the low 60s. Systems beating this baseline are often very complex and have access to additional corpora and large proprietary knowledge resources (see Chapter 2). Many less complex and informed systems, like the SALSA RTE system, are challenged by this baseline. It seems that systems often implement variants of this baseline on different levels of analysis.

To overcome this, one option for future research is to focus on modeling those examples, where surface overlap fails as entailment measure. These "hard" examples are on the one hand positive entailment pairs that exhibit low surface overlap and on the other hand negative pairs with high overlap. The following example, which was already discussed in Section 1.3, is one such example.

(8.1) T: El-Nashar was detained July 14 in Cairo. Britain notified Egyptian authorities that it suspected he may have had links to some of the attackers.

(8.2) H: El-Nashar was arrested in Egypt. (TRUE)

The number of these "hard" examples in current RTE corpora is manageable. For example, the RTE-3 development set contains only 34 negative pairs with full surface overlap (measured as described in Section 6.3.1) and 41 positive pairs with an overlap less than 50%. An interesting research question is whether it is possible to characterize these examples such that they can be automatically identified. Intuitively, only "deep"

semantic systems can tackle these examples. One could then design hybrid architectures, where shallow and deep components cooperate in a more sophisticated way. One could try to identify strengths and weaknesses of different approaches and also of concrete system components, algorithms, and resources used with the help of these examples.

Acquisition of negative examples. From the discussion above, one can also derive the goal to generate more principled entailment corpora, where surface overlap and textual entailment correlate less strongly. This leads to the more basic question on "what is a good negative example?" and on how to generate good negative examples. As we have already discussed in Section 2, negative examples in current RTE corpora correspond to system "failures". This models the general philosophy of the RTE challenge that the task of RTE should be application driven. Consider the examples below.

(8.3) T: More than 2,000 people lost their lives in the devastating Johnstown Flood.

(8.4) H: 100 or more people lost their lives in a ferry sinking. (IR, FALSE)

(8.5) T: Geithner criticized the report's findings on the World Bank, including a commission recommendation that the Bank stop lending to emerging market economies, concentrate its resources on the poorest countries and shift the IFI's concessional assistance from loans to grants.

(8.6) H: The World Bank is criticized for activities. (IR, FALSE)

(8.7) T: The renewed attention to the war came as peace activists, camping near the president's ranch, awaited a performance Sunday evening by peace movement icon Joan Baez.

(8.8) H: Folk singer Joan Baez brought her latest anti-war message Sunday to President Bush's adopted hometown, supporting Iraq war protesters camping out near his ranch. (SUM, FALSE)

The "errors" made by the respective systems are diverse. In (8.3)-(8.4), *flood* and *ferry sinking* have been confused and numbers do not fit. The problem with (8.6) is that is not related to the main message of (8.5), although it is indirectly compatible with what is said. Finally, in (8.7)-(8.8), different sources of information seem to have been mixed up.

We have the impression that these negative examples are not equally well suited for studying the textual entailment as empirical phenomenon, although it is difficult to pinpoint why this is the case. Properties of good negative examples might be prototypicality and the chance of being misjudged. (8.7)-(8.8) is not likely to be falsely judged as entailed. We think that, e.g., borderline cases that are not judged as entailed *in the vast amount of cases* are good candidates for more realistic "distractors". At the same time, we are not in favor of Zaenen et al. (2005)'s idea of selecting artificial examples with certain semantic properties for RTE corpora (cf. the discussion reported in Section 2.4.3). Our point is that it is important to develop a clear idea on what constitutes a good negative example.

Modeling relevance. In an information access scenario, textual entailment can be used as a mans to check semantic consistency of certain statements (hypotheses) with respect to documents they have been derived from (texts). Textual entailment itself does neither provide a clue on how to *find* suitable candidate texts in a data collection nor does it measure how *relevant* a particular text is for supporting the given hypothesis.

These tasks have to be carried out by appropriate systems or system components. But they have to be carried out in intelligent ways in order to meet our original goal of user-friendly information access. Texts which are not textually entailed by a hypothesis might still be highly relevant for a user looking for certain information. To illustrate this, consider using a question answering system to find an answer to the question *Is nuclear power safe?*. A straightforward hypothesis a question answering system could come up with is (8.11).

(8.9) T1: The nuclear sector is one of the safest industries in the United States [...].[1]

(8.10) T2: The Chernobyl Disaster showed the world the ugly face of nuclear power [...] the next accident is only a question of time.[2]

(8.11) H: Nuclear power is safe.

Possible texts retrieved by the system to support the hypothesis are (8.9) and (8.10). Both texts are relevant with respect to the query, but only (8.9) textually entails (8.11). An intelligent question answering system should therefore also generate hypotheses like *Nuclear power is not safe* or *Nuclear power is possibly safe* and seek for texts like (8.10) entailing these hypotheses as well.

Finally, we would be glad if our framework and results were taken as a basis for future research. We hope that this thesis succeeded in providing a clearer picture of the contribution and shortcomings of components already available or still needed for frame-based natural language processing.

[1] From `www.whitehouse.gov`.
[2] From `www.ippnw-students.org`.

Part IV.

Appendix

A. Complete Architecture of SALSA RTE System

Figure A.1 gives a rather technical overview of the scripts and Makefile commands, as well as the data types involved in the complete system architecture of the SALSA RTE system. A more detailed technical documentation would go beyond the scope of this thesis.

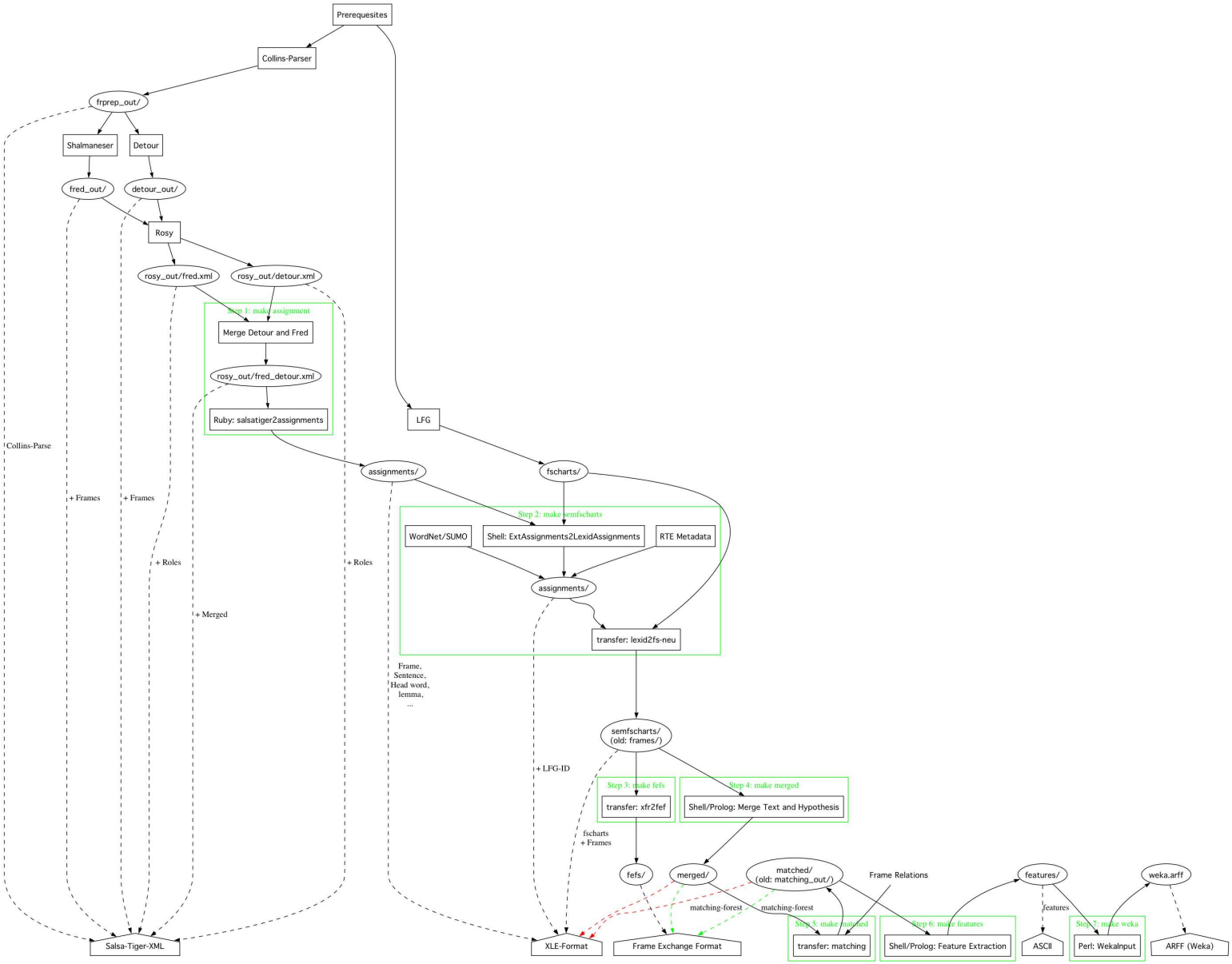


Figure A.1.: SALSA RTE system components, directory names and makefile structure.

B. Re-write Rules for Named Entities

```
%-------------------------------------------------------%
% rule-based assignments of sumo-concepts, based on NER %
%-------------------------------------------------------%

%-------------------------------------------------------%
% names + more specific info from NE-tags in morphology %
%-------------------------------------------------------%

% HUMAN=+ and NTYPE | NSEM | PROPER | PROPER-TYPE name
+human(X,+), nsem_proper_type(X,name)
==> sumo_syn(OntX,'Human')
&& make_ont_proj(X,_,OntX).

+ont(SemX,OntX), +sumo_syn(OntX,'Human')
==> frame(SemX,'PeopleDEF'), 'Person'(SemX,SemX).

% name modifiers become Person roles of the respective People frame
+frame(SemX,'People'), +'S::'(X,SemX),
+name_mod(X,Set), +in_set(Y,Set), +'S::'(Y,SemY),
frame(SemY,'PeopleDEF'), 'Person'(SemY,SemY)
==> 'Person'(SemX,SemY).

nsem_proper_type(X,name),
+phi(Cs,X),
+terminal(Cs,'+Company',_)
==> sumo_syn(OntX,'Corporation')
&& make_ont_proj(X,_,OntX).

+ont(SemX,OntX), +sumo_syn(OntX,'Corporation')
==> frame(SemX,'Businesses').

%-------------------------------------------------------%
% names + more specific info from NE-tags in morphology %
%-------------------------------------------------------%

% NTYPE | NSEM | PROPER | PROPER-TYPE title
nsem_proper_type(X,title)
```

```
==> frame(SemX,'Leadership')
&& make_proj(X,SemX).

%--------%
% dates  %
%--------%

% NTYPE | NSEM | TIME | month or day
+ntype(X,Y), +nsem(Y,Z), +time(Z,month)
==> sumo_syn(OntX,'Month')
&& make_ont_proj(X,_,OntX).

+ntype(X,Y), +nsem(Y,Z), +time(Z,day)
==> sumo_syn(OntX,'Day')
&& make_ont_proj(X,_,OntX).

+ont(SemX,OntX), +'S::'(_X,SemX),
( +sumo_syn(OntX,'Month') | +sumo_syn(OntX,'Day') )
==> frame(SemX,'Calendric_unit').

%-----------------------------------------------------------%
% locations + more specific info from NE-tags in morphology %
%-----------------------------------------------------------%

nsem_proper_type(X,location),
+phi(Cs,X),
+terminal(Cs,'+Place',_)
==> sumo_syn(OntX,'GeographicArea')
&& make_ont_proj(X,_,OntX).

nsem_proper_type(X,location),
+phi(Cs,X),
+terminal(Cs,'+Continent',_)
==> sumo_syn(OntX,'Continent')
&& make_ont_proj(X,_,OntX).

+ont(SemX,OntX),
( +sumo_syn(OntX,'GeographicArea') | +sumo_syn(OntX,'Continent') )
==> frame(SemX,'Locale').

nsem_proper_type(X,location),
+phi(Cs,X),
+terminal(Cs,'+Country',_)
==> sumo_syn(OntX,'Nation')
```

```
&& make_ont_proj(X,_,OntX).

nsem_proper_type(X,location),
+phi(Cs,X),
+terminal(Cs,'+City',_)
==> sumo_syn(OntX,'City')
&& make_ont_proj(X,_,OntX).

+ont(SemX,OntX),
( +sumo_syn(OntX,'Nation') |  +sumo_syn(OntX,'City') )
==> frame(SemX,'Political_locales').
```

C. FEF Export Format

The LFG exchange format used by the XLE re-write system (see Section 5.2.2) is not easily human readable as internal information is intermingled with the representation of the f-structures and projections. For inspection and exchange, we transform into a Prolog-like exchange format we call *FEF* (Frame Exchange Format). It contains only essential information, which has been further normalized, e.g., diverse values for the `tense` predicate are mapped onto `pres` and `past`. Table C.1 shows a FEF representation of the analysis of (5.3). In the table, some of the information related to the verb *leave*,

Normalized f-structure with projection	Frames, roles and ontological information
pred(f(1),leave).	**frame(s(24),'Departing').**
tense(f(1),pres).	**source(s(24),s(28)).**
stmt_type(f(1),declarative).	theme(s(24),s(27)).
mood(f(1),indicative).	frame(s(27),'People').
dsubj(f(1),f(8)).	role(s(27),person).
dobj(f(1),f(3)).	person(s(27),s(41)).
pred(f(3),'Sweden').	person(s(27),s(27)).
proper(f(3),location).	frame(s(28),'Political_locales').
num(f(3),sg).	frame(s(28),'Locale').
pred(f(8),'Larsson').	
proper(f(8),name).	**ont(s(24),s(34)).**
num(f(8),sg).	ont(s(27),s(38)).
mod(f(8),f(11)).	ont(s(28),s(33)).
pred(f(11),'Henrik').	ont(s(41),s(40)).
proper(f(11),name).	
num(f(11),sg).	wn_syn(s(33),'sweden#n#1').
	sumo_syn(s(33),'Nation').
sslink(f(1),s(24)).	sumo_syn(s(33),'GeographicArea').
sslink(f(3),s(28)).	sumo_inst(s(33),'Nation').
sslink(f(8),s(27)).	**wn_syn(s(34),'leave#v#1').**
sslink(f(11),s(41)).	**sumo_syn(s(34),'Leaving').**
	sumo_syn(s(38),'Human').
	sumo_syn(s(40),'Human').

Table C.1.: FEF for *Henrik Larsson leaves Sweden.*

which is represented by the f-structure node `f(1)` is highlighted. For example, its deep object is `f(3)` and its semantic projection points to `s(24)`. In turn, the frame of this semantic projection is DEPARTING and the SOURCE role points to the semantic projection of *Sweden*. The ontological projection of `s(24)` is `s(34)`, which is annotated with the WordNet synset `leave#v#1` and the SUMO class `Leaving`.

Bibliography

Rod Adams, Gabriel Nicolae, Cristina Nicolae, and Sanda Harabagiu. Textual entailment through extended lexical overlap and lexico-semantic matching. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.

Eneko Agirre and Philip Edmonds, editors. *Word Sense Disambiguation: Algorithms and Applications*. Number 33 in Text, Speech and Language Technology. Springer, 2006.

James F. Allen. Towards a general theory of action and time. *Artificial Intelligence*, 23 (2):123–154, 1984.

Ricardo Baeza-Yates and Bertheir Ribeiro-Neto. *Modern Information Retrieval*. Addision-Wesley, 2000.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. The Berkeley FrameNet project. In Christian Boitet and Pete Whitelock, editors, *Proceedings of ACL/COLING*, San Francisco, California, 1998. Morgan Kaufmann Publishers.

Ulrike Baldewein, Katrin Erk, Sebastian Pado, and Detlef Prescher. Semantic role labelling for chunk sequences. In *Proceedings of the CoNLL'04 shared task*, Boston, MA, 2004.

Roy Bar-Haim, Idan Szpektor, and Oren Glickman. Definition and analysis of intermediate entailment levels. In *Definition and Analysis of Intermediate Entailment Levels*, Ann Arbor, Michigan, 2005.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampiccolo, Bernardo Magnini, and Idan Szpektor, editors. *Proceedings of the Second PASCAL Challenges Workshop on Recognising Textual Entailment*, Venice, Italy, 2006.

Roy Bar-Haim, Ido Dagan, Bill Dolan, Lisa Ferro, Danilo Giampicollo, Bernardo Magnini, and Idan Szpektor. The Second PASCAL Recognising Textual Entailment Challenge. Presentation given at the workshop, 2007.

John A. Bateman, Renate Henschel, and Fabio Rinaldi. Generalized Upper Model 2.0: documentation. Technical report, GMD/Institut für Integrierte Publikations- und Informationssysteme, Darmstadt, Germany, 1995.

Patrick Blackburn, Johan Bos, Michael Kohlhase, and Hans de Nivelle. Inference and computational semantics. *Studies in Linguistics and Philosophy, Computing Meaning*, 77(2):11–28, 2001.

Hans C. Boas. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4):445–478, 2005.

Daniel Bobrow, Dick Crouch, Tracy Halloway King, Cleo Condoravdi, Lauri Karttunen, Rowan Nairn, Valeria de Paiva, and Annie Zaenen. Precision-focused textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.

Johan Bos. Towards wide-coverage semantic interpretation. In Harry Bunt, Jeroen Geertzen, and Elias Thijsse, editors, *Proceedings of Sixth International Workshop on Computational Semantics IWCS-6*, Tilburg, the Netherlands, 2005.

Johan Bos and Katja Markert. When logical inference helps determining textual entailment (and when it doe sn't). In *Proceedings of the RTE-2 Workshop*, Venice, Italy, 2006.

Johan Bos, Stephen Clark, Mark Steedman, James R. Curran, and Julia Hockenmaier. Wide-coverage semantic representations from a ccg parser. In *Proceedings of Coling 2004*, pages 1240–1246, Geneva, Switzerland, Aug 23–Aug 27 2004. COLING.

Sabine Brants, Stefanie Dipper, Siliva Hansen, Wolfgang Lezius, and George Smith. The TIGER treebank. In *Proceedings of the Workshop on Treebanks and Linguistic Theories*, Sozopol, 2002.

Joan Bresnan. *Lexical-functional syntax*. Blackwells, London, 2001.

Aljoscha Burchardt and Anette Frank. Approximating Textual Entailment with LFG and FrameNet Frames. In *Proceedings of the RTE-2 Workshop*, 2006.

Aljoscha Burchardt and Marco Pennacchiotti. FATE: a FrameNet-Annotated corpus for Textual Entailment. In *Proceedings of LREC 2008*, to appear.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. A WordNet detour to FrameNet. In Bernhard Fisseni, Hans-Christian Schmitz, Bernhard Schröder, and Petra Wagner, editors, *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, pages 408–421, Frankfurt am Main, 2005a. Lang, Peter.

Aljoscha Burchardt, Anette Frank, and Manfred Pinkal. Building text meaning representations from contextually related frames – a case study. In *Proceedings of the Sixth International Workshop on Computational Semantics, IWCS-6*, Tilburg, The Netherlands, 2005b.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. The SALSA corpus: a german corpus resource for lexical semantics. In *Proceedings of LREC 2006*, Genoa, Italy, 2006a.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Pado, and Manfred Pinkal. Salto – a versatile multi-level annotation tool. In *Proceedings of LREC 2006*, Genoa, Italy, 2006b.

Aljoscha Burchardt, Nils Reiter, Stefan Thater, and Anette Frank. A Semantic Approach to Textual Entailment: System Evaluation and Task Analysis. In *Proceedings of the Third PASCAL Recognising Textual Entailment Challenge Workshop*, 2007.

Aljoscha Burchardt, Sebastian Padó, Dennis Spohr, Anette Frank, and Ulrich Heid. Formalising Multi-layer Corpora in OWL DL – Lexicon Modelling, Querying and Consistency Control. In *Proceedings of the 3rd International Joint Conference on Natural Language Processing, IJCNLP 2008,*, Hyderabad, India, 2008.

Aljoscha Burchardt, Marco Pennacchiotti, Stefan Thater, and Manfred Pinkal. Assessing the impact of frame semantics on textual entailment. *Special Issue of the Journal of Natural Language Engineering on Textual Entailment and Paraphrasing.*, submitted.

John Burger and Lisa Ferro. Generating an entailment corpus from news headlines. In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Eugene Charniak. A maximum-entropy-inspired parser. In *Proceedings of NACL*, pages 132–139, San Francisco, CA, USA, 2000. Morgan Kaufmann Publishers Inc.

Peter Clark, Phil Harrison, John Thompson, William Murray, Jerry Hobbs, and Christiane Fellbaum. On the role of lexical and world knowledge in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.

Stephen Clark, Julia Hockenmaier, and Mark Steedman. Building Deep Dependency Structures Using a Wide-Coverage CCG Parser. In *Proceedings of ACL-02*, Philadelphia, 2002.

Allan M. Collins and M. Ross Quillian. Retrieval time from semantic memory. *Journal of Verbal Learning and Verbal Behavior*, 8:240–248, 1969.

Michael Collins. *Head-driven Statistical Models for Natural Language Parsing.* PhD thesis, University of Pennsylvania, Philadelphia, 1999.

Ann Copestake and Dan Flickinger. An open-source grammar development environment and broad-coverage english grammar using HPSG. In *Proceedings of LREC-2000*, 2000.

Dick Crouch, Cleo Condoravdi, Valeria de Paiva, Reinhard Stolle, and Daniel G. Bobrow. Entailment, intensionality and text understanding. In Graeme Hirst and Sergei Nirenburg, editors, *HLT-NAACL 2003 Workshop: Text Meaning*, Edmonton, Alberta, Canada, May 31 2003. Association for Computational Linguistics.

Richard Crouch. Packed rewriting for mapping semantics to KR. In *Proceedings of the Sixth International Workshop on Computational Semantics*, Tilburg, 2005.

Richard Crouch, Cleo Condoravdi, Reinhard Stolle, Tracy King, John Everett, and Daniel Bobrow. Scalability of Redundancy Detection in Focused Document Collections. In *Proceedings First International Workshop on Scalable Natural Language Understanding, SCANALU-2002*, Heidelberg, Germany, 2002.

Richard Crouch, Lauri Karttunen, and Annie Zaenen. Circumscribing is not excluding: A reply to manning. Palo Alto Research Center, 2006.

Ido Dagan, Oren Glickman, and Bernardo Magnini. The PASCAL Recognising Textual Entailment Challenge. In Joaquin Quiñonero-Candela, Ido Dagan, Bernardo Magnini, and Florence D'Alché-Buc, editors, *Evaluating Predictive Uncertainty, Visual Object Categorization and Textual Entailment*, volume 3944 of *Lecture Notes in Computer Science*, pages 177–190, Heidelberg, Germany, 2006. Springer.

Donald Davidson. The logical form of action sentences. In Nicholas Rescher, editor, *The Logic of Decision and Action*. University of Pittsburgh Press, Pittsburgh, 1967.

Rodolfo Delmonte, Antonella Bristot, Marco Aldo Piccolino Boniforti, and Sara Tonelli. Entailment and anaphora resolution in rte3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.

Donald Dowty. Thematic proto-roles and argument selection. *Language*, 67(3), 1991.

Michael Ellsworth, Katrin Erk, Paul Kingsbury, and Sebastian Pado. PropBank, SALSA and FrameNet: How design determines product. In *Proceedings of the Workshop on Building Lexical Resources From Semantically Annotated Corpora, LREC-2004*, Lisbon, 2004.

Katrin Erk. Frame assignment as word sense disambiguation. In Harry Bunt, Jeroen Geertzen, and Elias Thijsse, editors, *Proceedings of the Sixth International Workshop on Computational Semantics (IWCS)*, Tilburg University, Tilburg, The Netherlands, January 2005.

Katrin Erk. Unknown word sense detection as outlier detection. In *Proceedings of the main conference on Human Language Technology Conference of the North American Chapter of the Association of Computational Linguistics*, Morristown, NJ, USA, 2006. Association for Computational Linguistics.

Katrin Erk and Sebastian Pado. A powerful and versatile XML format for representing role-semantic annotation. In *Proceedings of LREC-2004*, Lisbon, Portugal, 2004.

Katrin Erk and Sebastian Pado. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC 2006*, Genoa, Italy, 2006.

Katrin Erk, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. Towards a resource for lexical semantics: a large german corpus with extensive semantic annotation. In *Proceedings of ACL-03*, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Christiane Fellbaum. English verbs as a semantic net. *Int. Journal of Lexicography*, 3 (4):278–301, 1990.

Christiane Fellbaum, editor. *WordNet. An electronic lexical database*. Language, Speech, and Communication. MIT Press, Cambridge, MA, 1998.

Charles J. Fillmore. Frame semantics and the nature of language. In *Annals of the New York Academy of Sciences: Conference on the Origin and Development of Language and Speech*, volume 280, pages 20–32, 1976.

Charles J. Fillmore. Scenes-and-frames semantics. In Antonio Zampolli, editor, *Linguistic Structures Processing*, volume 5 of *Fundamental Studies in Computational Science*, Amsterdam, 1977. North Holland.

Charles J. Fillmore. Frame semantics. In The Linguistic Society of Korea, editor, *Linguistics in the Morning Calm*. Hanshin, Seoul, 1982.

Charles J. Fillmore. Frames and the semantics of understanding. *Quaderni di Semantica*, 6(2):222–254, 1985.

Charles J. Fillmore. The case for case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–210. Holt, Rinehart, and Winston, New York, 1968.

Charles J. Fillmore and B. T. Sue Atkins. FrameNet and lexicographic relevance. In *Proceedings of LREC-1998*, Granada, Spain, 1998.

Charles J. Fillmore and Collin F. Baker. Frame semantics for text understanding. In *Proceedings of WordNet and Other Lexical Resources Workshop, NAACL*, 2001.

Charles J. Fillmore, Charles Wooters, and Collin F. Baker. Building a large lexical databank which provides deep semantics. In *Proceedings of the Pacific Asian Conference on Language, Information and Computation*, HongKong, 2001.

Gerhard Fliedner. *Linguistically Informed Question Answering*. PhD thesis, Saarland University, 2007.

Anette Frank and Katrin Erk. Towards an LFG syntax-semantics interface for frame semantics annotation. In Alexander Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing*, pages 1–12. Springer Verlag, Heidelberg, 2004. Lecture Notes in Computer Science, Vol. 2945.

Anette Frank, Hans-Ulrich Krieger, Feiyu Xu, Hans Uszkoreit, Berthold Crysmann, Brigitte Jörg, and Ulrich Schäfer. Question answering from structured knowledge sources. *Journal of Applied Logic, Special Issue on Questions and Answers: Theoretical and Applied Perspectives*, 5(1):20–48, 3 2006.

Aldo Gangemi, Nicola Guarino, Claudio Masolo, Alessandro Oltramari, and Luc Schneider. Sweetening ontologies with dolce. In *Proceedings of the 13th International Conference on Knowledge Engineering and Knowledge Management*, Siguenza, Spain, 2002.

Konstantina Garoufi. Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. *M.Sc. thesis*, Saarland University, 2007.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan, editors. *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, 2007a. Association for Computational Linguistics.

Danilo Giampiccolo, Bernardo Magnini, Ido Dagan, and Bill Dolan. The third pascal recognizing textual entailment challenge. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing* Giampiccolo et al. (2007a), pages 1–9.

Daniel Gildea and Julia Hockenmaier. Identifying semantic roles using combinatory categorial grammar. In *Proceedings of the 2003 conference on Empirical methods in natural language processing*, pages 57–64, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Daniel Gildea and Daniel Jurafsky. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288, 2002.

Jeffrey Steven Gruber. *Studies in Lexical Relations.* PhD thesis, MIT, 1965.

Aria D. Haghighi, Andrew Y. Ng, and Christopher D. Manning. Robust textual inference via graph matching. In *HLT '05: Proceedings of the conference on Human Language Technology and Empirical Methods in Natural Language Processing*, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

Per-Kristian Halvorsen and Ronald M. Kaplan. Projections and semantic description in lexical-functional grammar. In *FGCS*, pages 1116–1122, 1988.

Birgit Hamp and Helmut Feldweg. Germanet - a lexical-semantic net for german. In *Proceedings of* ACL *workshop Automatic Information Extraction and Building of Lexical Semantic Resources for* NLP *Applications, Madrid.*, 1997.

Sanda M. Harabagiu, Dan I. Moldovan, Marius Pasca, Rada Mihalcea, Mihai Surdeanu, Razvan C. Bunescu, Roxana Girju, Vasile Rus, and Paul Morarescu. Falcon: Boosting knowledge for answer engines. In *TREC*, 2000.

Jesus Herrera, Anselmo Penas, Alvaro Rodrigo, and Felisa Verdejo. UNED at PASCAL RTE-2 Challenge. In *Proceedings of the RTE-2 Workshop*, Venice, Italy, 2006.

Andrew Hickl and Jeremy Bensley. A discourse commitment-based framework for recognizing textual entailment. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.

Andrew Hickl, Jeremy Bensley, John Williams, Kirk Roberts, Bryan Rink, and Ying Shi. Recognizing Textual Entailment with LCC's Groundhog System. In *Proceedings of the RTE-2 Workshop*, 2006a.

Andrew Hickl, Jeremy Bensley, John Williams, Kirk Roberts, Bryan Rink, and Ying Shi. Recognizing Textual Entailment with LCC's Groundhog System. Presentation held at the second RTE workshop., 2006b.

Ray Jackendoff. *Semantic Interpretation in Generative Grammar*. MIT Press, Cambridge, Mass., 1972.

Ray Jackendoff. *Semantic Structures*. The MIT Press, Cambridge, MA, 1990.

Hans Kamp and Uwe Reyle. *From Discourse to Logic: Introduction to Model-theoretic Semantics, Logic and Discourse Representation Theory*. Kluwer Academic Publishers, 1993.

Ron Kaplan, Stefan Riezler, Tracy H King, John T Maxwell III, Alex Vasserman, and Richard Crouch. Speed and accuracy in shallow and deep stochastic parsing. In Daniel Marcu Susan Dumais and Salim Roukos, editors, *Proceedings of HLT-NAACL 2004*, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Adam Kilgarriff and J. Rosenzweig. Framework and results for English SENSEVAL. *Computers and the Humanities*, 34(1-2), 2000.

Paul Kingsbury, Martha Palmer, and Mitch Marcus. Adding semantic annotation to the Penn TreeBank. In *Proceedings of HLT*, San Diego, 2002.

Karin Kipper, Hoa Trang Dang, and Martha Palmer. Class-based construction of a verb lexicon. In *AAAI/IAAI*, pages 691–696, 2000.

George Lakoff. Linguistics and natural logic. In D. Davidson and G. Harman, editors, *Semantics of Natural Language*, pages 545–665. Reidel, Dordrecht, 1972.

Ronald W. Langacker. *Foundations of cognitive grammar: Theoretical Prerequisites*. Stanford University Press, Stanford, CA, 1987. Vol 1, 1987(Hardcover), 1999(Paperback).

Douglas B. Lenat. CYC: A large-scale investment in knowledge infrastructure. *Communications of the ACM*, 38(11):33–38, 1995.

Beth Levin. *English Verb Classes and Alternations*. The University of Chicago Press, Chicago, 1993.

Maria Liakata and Stephen Pulman. Learning Theories from Texts. In *Proceedings of COLING 2004*, Geneva, Switzerland, 2004.

Dekang Lin. An information-theoretic definition of similarity. In Jude W. Shavlik, editor, *Proceedings of the Fifteenth International Conference on Machine Learning (ICML 1998), Madison, Wisconson, USA, July 24-27, 1998*, pages 296–304. Morgan Kaufmann, 1998.

Dekang Lin and Patrick Pantel. DIRT - Discovery of inference rules from text. In *Proceedings of the ACM Conference on Knowledge Discovery and Data Mining (KDD-01)*, San Francisco, CA, 2001.

Ken Litkowski. Componential analysis for recognizing textual entailment. In *Proceedings of the RTE-2 Workshop*, 2006.

John B. Lowe, Collin F. Baker, and Charles J. Fillmore. A frame-semantic approach to semantic annotation. In *Proceedings of ACL SIGLEX Workshop on Tagging Text with Lexical Semantics*, Washington, D.C., 1997. ACL.

Bill MacCartney and Christopher D. Manning. Natural logic for textual inference. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.

Inderjeet Mani. *Automatic summarization.*, volume 3 of *Natural language processing series*, pages 221 – 223. John Benjamins, Amsterdam, 2001.

Christopher D. Manning. Local textual inference: It's hard to circumscribe, but you know it when you see it - and nlp needs it. `http://nlp.stanford.edu/~manning/papers/LocalTextualInference.pdf`, 2006.

Claudio Masolo, Stefano Borgo, Aldo Gangemi, Nicola Guarino, Alessandro Oltramari, and Luc Schneider. The wonderweb library of foundational ontologies. Technical report, ISTC-CNR/Institute of Cognitive Sciences and Technologies, Rome, 2003.

Diana McCarthy, Rob Koeling, Julie Weeds, and John Carroll. Using automatically acquired predominant senses for word sense disambiguation. In Rada Mihalcea and Phil Edmonds, editors, *Senseval-3: Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text*, Barcelona, Spain, July 2004. Association for Computational Linguistics.

Rada Mihalcea and Phil Edmonds, editors. *Proceedings of Senseval-3: The Third International Workshop on the Evaluati on of Systems for the Semantic Analysis of Text*, Barcelona, Spain, 2004.

George A. Miller, Richard Beckwith, Christiane Fellbaum, Derek Gross, and Katherine J. Miller. Introduction to wordnet: an on-line lexical database. In *International Journal of Lexicography*, number 3 in 4, pages 235–244, 1990.

Behrang Mohit and Srini Narayanan. Semantic extraction with wide-coverage lexical resources. In Marti Hearst and Mari Ostendorf, editors, *HLT-NAACL 2003: Short Papers*, pages 64–66, Edmonton, Alberta, Canada, May 27 - June 1 2003. Association for Computational Linguistics.

Dan Moldovan and Adrian Novischi. Lexical chains for question answering. In *Proceedings of the 19th international conference on Computational linguistics*, Morristown, NJ, USA, 2002. Association for Computational Linguistics.

Dan Moldovan, Sanda Harabagiu, M. Paşca, R. Mihalcea, R. Goodrum, R. Gîrji, and V. Rus. LASSO: A Tool for Surfing the Answer Net. *Proceedings of TREC*, 8, 1999.

Dan Moldovan, Christine Clark, Sanda Harabagiu, and Steve Maiorano. Cogex: a logic prover for question answering. In *NAACL '03: Proceedings of the 2003 Conference of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Dan I. Moldovan and Vasile Rus. Logic form transformation of wordnet and its applicability to question answering. In *Proceedings of ACL-2001*, Morristown, NJ, USA, 2001. Association for Computational Linguistics.

Richard Montague. The proper treatment of quantification in ordinary English. In J. Hintikka, J. Moravcsik, and P. Suppes, editors, *Approaches to Natural Language*. Reidel, Dordrecht, 1973.

Chirstoph Monz and Maarten de Rijke. Light-weight entailment checking for computational semantics. In Patrick Blackburn and Michael Kohlhase, editors, *Workshop Proceedings ICoS-3*, pages 59–72, 2001.

Christof Monz and Maarten de Rijke. Deductions with meaning. In Michael Moortgat, editor, *Logical Aspects in Computational Linguistics*, LNAI. Springer, Berlin et al., 1999.

Thomas S. Morton. Coreference for NLP applications. In *Proceedings of ACL-2000*, pages 173–180, Morristown, NJ, USA, 2000. Association for Computational Linguistics.

Rowan Nairn, Cleo Condoravdi, and Lauri Karttunen. Computing relative polarity for textual inference. In *Proceedings of ICoS-5*, Buxton, UK, 2006.

Srini Narayanan and Sanda Harabagiu. Answering questions using advanced semantics and probabilistic inference. In Sanda Harabagiu and Finley Lacatusu, editors, *HLT-NAACL 2004: Workshop on Pragmatics of Question Answering*, pages 10–16, Boston, Massachusetts, USA, May 2 - May 7 2004. Association for Computational Linguistics.

Srini Narayanan, Collin F. Baker, Charles J. Fillmore, and Miriam R. L. Petruck. Framenet meets the semantic web: Lexical semantics for the web. In Dieter Fensel, Katia P. Sycara, and John Mylopoulos, editors, *International Semantic Web Conference*, volume 2870 of *Lecture Notes in Computer Science*. Springer, 2003.

Ian Niles. Mapping WordNet to the SUMO Ontology. *Proceedings of the IEEE International Knowledge Engineering conference*, 2003.

Ian Niles and Adam Pease. Linking lexicons and ontologies: Mapping WordNet to the Suggested Upper Merged Ontology. In *Proceedings of IKE '03*, Las Vegas, Nevada, 2003.

Ian Niles and Adam Pease. Towards a standard upper ontology. In *FOIS '01: Proceedings of the international conference on Formal Ontology in Information Systems*, pages 2–9, New York, NY, USA, 2001. ACM Press.

Ian Niles and Allan Terry. The milo: A general-purpose, mid-level ontology. In Hamid R. Arabnia, editor, *Proceedings of the International Conference on Information and Knowledge Engineering. IKE'04*, pages 15–19, Las Vegas, USA, 2004. CSREA Press.

K. Ohara, S. Fujii, T. Ohori, R. Suzuki, H. Saito, and S. Ishizak. The Japanese FrameNet project: An introduction. In *Proceedings of the Workshop on Building Lexical Resources from Semantically Annotated Corpora at LREC 2004*, 2004.

Alessandro Oltramari, Aldo Gangemi, Nicola Guarino, and Claudio Masolo. Restructuring wordnet's top-level: The ontoclean approach. In *Proceedings of the LREC 2002 (OntoLex workshop)*, Las Palmas, Spain, 2002.

Sebastian Pado and Katrin Erk. To cause or not to cause: Cross-lingual semantic matching for paraphrase modelling. In *Proceedings of the Cross-Language Knowledge Induction Workshop*, Cluj-Napoca, Romania, 2005.

Martha Palmer, Daniel Gildea, and Paul Kingsbury. The proposition bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106, March 2005.

Siddharth Patwardhan, Satanjeev Banerjee, and Ted Pedersen. SenseRelate::TargetWord — a generalized framework for word sense disambiguation. In *Proceedings of the ACL Interactive Poster and Demonstration Sessions*, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Marco Pennacchiotti. *Methods and Resources for Recognizing Textual Entailment*. PhD thesis, University of Roma Tor Vergata, 2007.

Miriam R. L. Petruck. Frame semantics. In Jef Verschueren, Jan-Ola Östman, Jan Blommaert, and Chris Bulcaen, editors, *Handbook of Pragmatics*. Benjamins, Philadelphia, 1996.

Miriam R.L. Petruck, Charles J. Fillmore, Collin F. Baker, Michael Ellsworth, and Josef Ruppenhofer. Reframing FrameNet data. In *Proceedings of The 11th EURALEX International Congress*, Lorient, France, 2004.

Manfred Pinkal. *Logic and Lexicon: the semantics of the indefinite*, volume 56 of *Studies in linguistics and philosophy*. Kluwer, Dordrecht, 1995.

Carl Pollard and Ivan A. Sag. *Head-Driven Phrase Structure Grammar*. University of Chicago Press and CSLI Publications, Chicago, Illinois, 1994.

Simone Paolo Ponzetto and Michael Strube. Semantic role labeling for coreference resolution. In *Companion Volume of the Proceedings of the 11th Meeting of the EACL*, Trento, Italy, 2006.

Sameer Pradhan, Wayne Ward, Kadri Hacioglu, James H. Martin, and Daniel Jurafsky. Semantic role labeling using different syntactic views. In *Proceedings of ACL-2005*, Morristown, NJ, USA, 2005. Association for Computational Linguistics.

Long Qiu, Min-Yen Kan, and Tat-Seng Chua. Paraphrase recognition via dissimilarity significance classification. In *Proceedings of the 2006 Conference on Empirical Methods in Natural Language Processing*, Sydney, Australia, July 2006. Association for Computational Linguistics.

Nils Reiter. Towards a linking of framenet and sumo. Master's thesis, Saarland University, September 2007.

Nils Reiter. Detour - a CPAN module. Student project, 2006.

Philip Resnik. Using information content to evaluate semantic similarity in a taxonomy. In *Proc. of the 14th IJCAI*, Montreal, Canada, 1995.

Stefan Riezler, Tracy H. King, Ronald M. Kaplan, Richard. Crouch, John T. Maxwell, and Mark Johnson. Parsing the Wall Street Journal using a Lexical-Functional Grammar and Discriminative Estimation Techniques. In *Proceedings of the ACL'02*, Philadelphia, PA., 2002.

Josef Ruppenhofer, Michael Ellsworth, Miriam R. L. Petruck, and Christopher R. Johnson. Framenet: Theory and practice. Available at `http://framenet.icsi.berkeley.edu/`, 2006.

Roger C. Schank. Identification of conceptualizations underlying natural language. In Roger C. Schank and Kenneth M. Colby, editors, *Computer Models of Thought and Language*, pages 187–247. W. H. Freeman, San Francisco, CA, 1973.

Jan Scheffczyk, Adam Pease, and Michael Ellsworth. Linking FrameNet to the Suggested Upper Merged Ontology. In *Proceedings of Formal Ontology in Information Systems*, 2006.

Helmut Schmid. Probabilistic part-of-speech tagging using decision trees. In *Proceedings of International Conference on New Methods in Language Processing*, Manchester, UK, 1994.

Lei Shi and Rada Mihalcea. Putting pieces together: Combining FrameNet, VerbNet and WordNet for robust semantic parsing. In Alexander F. Gelbukh, editor, *Computational Linguistics and Intelligent Text Processing, 6th International Conference*, Lecture Notes in Computer Science, pages 100–111. Springer, 2005.

Carlos Subirats and Miriam Petruck. Surprise! Spanish FrameNet! In *Proceedings of the Workshop on Frame Semantics at the XVII. International Co ngress of Linguists*, Prague, 2003.

Beth Sundheim, editor. *Proceedings of the Third Message Understanding Conference*, San Mateo, CA, 1991. Morgan Kaufmann.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. Using predicate-argument structures for information extraction. In *Proceedings of ACL-2003*, Morristown, NJ, USA, 2003. Association for Computational Linguistics.

Alfred Tarski. *Logic, Semantics, Metamathematics*. Hackett, second edition, 1983. Translated by J. H. Woodger.

Marta Tatu and Dan Moldovan. A logic-based semantic approach to recognizing textual entailment. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, Sydney, Australia, July 2006. Association for Computational Linguistics.

Marta Tatu and Dan Moldovan. COGEX at RTE 3. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.

Marta Tatu, Brandon Iles, John Slavick, Adrian Novischi, and Dan Moldovan. COGEX at the Second Recognizing Textual Entailment Challenge. In *Proceedings of the RTE-2 Workshop*, Venice, Italy, 2006.

Olga Uryupina. *Knowledge acquisition for coreference resolution*. PhD thesis, Saarland University, 2007.

Lucy Vanderwende and William B. Dolan. What syntax can contribute in the entailment task. In Joaquin Quinonero Candela, Ido Dagan, Bernardo Magnini, and Florence d'Alche Buc, editors, *MLCW*, volume 3944 of *Lecture Notes in Computer Science*, pages 205–216. Springer, 2005.

Oliver Čulo. Methoden zur manuellen und automatischen Verknüpfung semantischer Verbklassifikationen. Diploma thesis, Universität des Saarlandes, 2006.

Ellen M. Voorhees. The trec-8 question answering track report. In *Proceedings of TREC*, 1999.

Terry Winograd. On primitives, prototypes, and other semantic anomalies. In *Proceedings of the 1978 workshop on Theoretical issues in natural language processing*, pages 25–32, 1978.

Ian H. Witten and Eibe Frank. *Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, San Francisco, 2 edition, 2005.

The XTAG Research Group. A Lexicalized Tree Adjoining Grammar for English. Technical Report IRCS 01-03, University of Pennsylvania, 2001.

Zdenek Žabokrtský. Automatic functor assignment in the prague dependency treebank. In *Proceedings of TSD'00*, 2000.

Annie Zaenen, Lauri Karttunen, and Richard Crouch. Local textual inference: Can it be defined or circumscribed? In *Proceedings of the ACL Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, pages 31–36, Ann Arbor, Michigan, June 2005. Association for Computational Linguistics.

Fabio Massimo Zanzotto, Alessandro Moschitti, Marco Pennacchiotti, and Maria Teresa Pazienza. Learning textual entailment from examples. In *Proceedings of the RTE-2 Workshop*, Venice, Italy, 2006.

Fabio Massimo Zanzotto, Marco Pennacchiotti, and Alessandro Moschitti. Shallow semantic in fast textual entailment rule learners. In *Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing*, Prague, June 2007. Association for Computational Linguistics.