SemEval-2010 Task 10: Linking Events and Their Participants in Discourse

Josef Ruppenhofer and **Caroline Sporleder**
Computational Linguistics
Saarland University
{josefr,csporled}@coli.uni-sb.de

Roser Morante
CNTS
University of Antwerp
Roser.Morante@ua.ac.be

Collin Baker
ICSI
Berkeley, CA 94704
collin@icsi.berkeley.edu

Martha Palmer
Department of Linguistics
University of Colorado at Boulder
martha.palmer@colorado.edu

Abstract

We describe the SemEval-2010 shared task on "Linking Events and Their Participants in Discourse". This task is an extension to the classical semantic role labeling task. While semantic role labeling is traditionally viewed as a sentence-internal task, local semantic argument structures clearly interact with each other in a larger context, e.g., by sharing references to specific discourse entities or events. In the shared task we looked at one particular aspect of cross-sentence links between argument structures, namely linking locally uninstantiated roles to their co-referents in the wider discourse context (if such co-referents exist). This task is potentially beneficial for a number of NLP applications, such as information extraction, question answering or text summarization.

1 Introduction

Semantic role labeling (SRL) has been defined as a sentence-level natural-language processing task in which semantic roles are assigned to the syntactic arguments of a predicate (Gildea and Jurafsky, 2002). Semantic roles describe the function of the participants in an event. Identifying the semantic roles of the predicates in a text allows knowing who did what to whom when where how, etc.

However, semantic role labeling as it is currently defined misses a lot of information due to the fact that it is viewed as a sentence-internal task. Hence, relations between different local semantic argument structures are disregarded. This view of SRL as a sentence-internal task is partly due to the fact that large-scale manual annotation projects such as FrameNet[1] and PropBank[2] typically present their annotations lexicographically by lemma rather than by source text.

It is clear that there is an interplay between local argument structure and the surrounding discourse (Fillmore, 1977). In early work, Palmer et al. (1986) discussed filling null complements from context by using knowledge about individual predicates and tendencies of referential chaining across sentences. But so far there have been few attempts to find links between argument structures across clause and sentence boundaries explicitly on the basis of semantic relations between the predicates involved. Two notable exceptions are Fillmore and Baker (2001) and Burchardt et al. (2005). Fillmore and Baker (2001) analyse a short newspaper article and discuss how frame semantics could benefit discourse processing but without making concrete suggestions of how to model this. Burchardt et al. (2005) provide a detailed analysis of the links between the local semantic argument structures in a short text; however their system is not fully implemented either.

With the shared task, we aimed to make a first step towards taking SRL beyond the domain of individual sentences by linking local semantic argument structures to the wider discourse context. The task addresses the problem of finding fillers for roles which are neither instantiated as direct dependents of our target predicates nor displaced through long-distance dependency or coinstantiation constructions. Often a referent for an uninstantiated role can be found in the wider context, i.e. in preceding or following sentences. An example is given in (1), where the CHARGES role

[1] http://framenet.icsi.berkeley.edu/
[2] http://verbs.colorado.edu/~mpalmer/projects/ace.html

(ARG2 in PropBank) of *cleared* is left empty but can be linked to *murder* in the previous sentence.

(1) In a lengthy court case the defendant was tried for murder. In the end, he was cleared.

Another very rich example is provided by (2), where, for instance, the experiencer and the object of jealousy are not overtly expressed as dependents of the noun *jealousy* but can be inferred to be Watson and the speaker, Holmes, respectively.

(2) Watson won't allow that I know anything of art but that is mere jealousy because our views upon the subject differ.

This paper is organized as follows. In Section 2 we define how the concept of Null Instantiation is understood in the task. Section 3 describes the tasks to be performed, and Section 4, how they are evaluated. Section 5 presents the participant systems, and Section 6, their results. Finally, in Section 7, we put forward some conclusions.

2 Null Instantiations

The theory of null complementation used here is the one adopted by FrameNet, which derives from the work of Fillmore (1986).[3] Briefly, omissions of core arguments of predicates are categorized along two dimensions, the licensor and the interpretation they receive. The idea of a licensor refers to the fact that either a particular lexical item or a particular grammatical construction must be present for the omission of a frame element (FE) to occur. For instance, the omission of the agent in (3) is licensed by the passive construction.

(3) No doubt, mistakes were made $0^{Protagonist}$.

The omission is a constructional omission because it can apply to any predicate with an appropriate semantics that allows it to combine with the passive construction. On the other hand, the omission in (4) is lexically specific: the verb *arrive* allows the Goal to be unspecified but the verb *reach*, also a member of the Arriving frame, does not.

(4) We arrived 0^{Goal} at 8pm.

[3] Palmer et al.'s (1986) treatment of uninstantiated 'essential roles' is very similar (see also Palmer (1990)).

The above two examples also illustrate the second major dimension of variation. Whereas, in (3) the protagonist making the mistake is only existentially bound within the discourse (instance of indefinite null instantiation, INI), the Goal location in (4) is an entity that must be accessible to speaker and hearer from the discourse or its context (definite null instantiation, DNI). Finally, note that the licensing construction or lexical item fully and reliably determines the interpretation. Whereas missing by-phrases have always an indefinite interpretation, whenever *arrive* omits the Goal lexically, the Goal has to be interpreted as definite, as it is in (4).

The import of this classification to the task here is that we will concentrate on cases of DNI, be they licensed lexically or constructionally.

3 Description of the Task

3.1 Tasks

We originally intended to offer the participants a choice of two different tasks: a **full task**, in which the test set was only annotated with gold standard word senses (i.e., frames) for the target words and the participants had to perform role recognition/labeling and null instantiation linking, and a **NI only** task, in which the test set was already annotated with gold standard semantic argument structures and the participants only had to recognize definite null instantiations and find links to antecedents in the wider context (NI linking).

However, it turned out that the basic semantic role labeling task was already quite challenging for our data set. Previous shared tasks have shown that frame-semantic SRL of running text is a hard problem (Baker et al., 2007), partly due to the fact that running text is bound to contain many frames for which no or little annotated training data are available. In our case the difficulty was increased because our data came from a new genre and domain (i.e., crime fiction, see Section 3.2). Hence, we decided to add standard SRL, i.e., role recognition and labeling, as a third task (**SRL only**). This task did not involve NI linking.

3.2 Data

The participants were allowed to make use of a variety of data sources. We provided a training set annotated with semantic argument structure and null instantiation information. The annotations were originally made using FrameNet-style and

later mapped semi-automatically to PropBank annotations, so that participants could choose which framework they wanted to work in. The data formats we used were TIGER/SALSA XML (Erk and Padó, 2004) (FrameNet-style) and a modified CoNLL-format (PropBank-style). As it turned out, all participants chose to work on FrameNet-style annotations, so we will not describe the PropBank annotation in this paper (see Ruppenhofer et al. (2009) for more details).

FrameNet-style annotation of full text is extremely time-consuming. Since we also had to annotate null instantiations and co-reference chains (for evaluation purposes, see Section 4), we could only make available a limited amount of data. Hence, we allowed participants to make use of additional data, in particular the FrameNet and PropBank releases.[4] We envisaged that the participants would want to use these additional data sets to train SRL systems for the full task and to learn something about typical fillers for different roles in order to solve the NI linking task. The annotated data sets we made available were meant to provide additional information, e.g., about the typical distance between an NI and its filler and about how to distinguish DNIs and INIs.

We annotated texts from two of Arthur Conan Doyle's fiction works. The text that served as training data was taken from "The Adventure of Wisteria Lodge". Of this lengthy, two-part story we annotated the second part, titled "The Tiger of San Pedro". The test set was made up of the last two chapters of "The Hound of the Baskervilles". We chose fiction rather than news because we believe that fiction texts with a linear narrative generally contain more context-resolvable NIs. They also tend to be longer and have a simpler structure than news texts, which typically revisit the same facts repeatedly at different levels of detail (in the so-called 'inverted pyramid' structure) and which mix event reports with commentary and evaluation, thus sequencing material that is understood as running in parallel. Fiction texts should lend themselves more readily to a first attempt at integrating discourse structure into semantic role labeling. We chose Conan Doyle's work because most of his books are not subject to copyright anymore, which allows us to freely release the annotated data. Note, however, that this choice of data means that our texts come from a different domain and genre than many of the examples in FrameNet and PropBank as well as making use of a somewhat older variety of English.[5]

Table 1 provides basic statistics of the data sets. The training data had 3.1 frames per sentence and the test data 3.2, which is lower than the 8.8 frames per sentence in the test data of the 2007 SemEval task on Frame Semantic Structure Extraction.[6] We think this is mainly the result of switching to a domain different from the bulk of what FrameNet has made available in the way of full-text annotation. In doing so, we encountered many new frames and lexical units for which we could not ourselves create the necessary frames and provide lexicographic annotations. The statistics also show that null-instantiation is relatively common: in the training data, about 18.7% of all FEs are omitted, and in the test set, about 18.4%. Of the DNIs, 80.9% had an antecedent in the training data, and 74.2% in the test data.

To ensure a high quality of the annotations, both data sets were annotated by more than one person and then adjudicated. The training set was annotated independently by two experienced annotators and then adjudicated by the same two people. The test set was annotated by three annotators and then adjudicated by the two experienced annotators. Throughout the annotation and adjudication process, we discussed difficult cases and also maintained a wiki. Additionally, we created a software tool that checked the consistency of our annotations against the frame, frame element and FE-relation specifications of FrameNet and alerted annotators to problems with their annotations. The average agreement (F-score) for frame assignment for pairs of annotators on the two chapters in the test set ranges from 0.7385 to 0.7870. The agreement of individual annotators with the adjudicated gold standard ranges from 0.666 to 0.798. Given that the gold standard for the two chapters features 228 and 229 different frame types, respectively, this level of agreement seems quite good.

[4]For FrameNet we provided an intermediate release, FrameNet 1.4 alpha, which contained more frames and lexical units than release 1.3.

[5]While PropBank provides annotations for the Penn Treebank and is thus news-based, the lexicographic annotations in FrameNet are extracted from the BNC, a balanced corpus. The FrameNet full-text annotations, however, only cover three domains: news, travel guides, and nuclear proliferation reports.

[6]The statistics in Table 1 and all our discussion of the data includes only instances of semantic frames and ignores the instances of the Coreference, Support, and Relativization frames, which we labeled on the data as auxiliary information.

data set	sentences	tokens	frame inst.	frame types	overt FEs	DNIs (resolved)	INIs
train	438	7,941	1,370	317	2,526	303 (245)	277
test	525	9,131	1,703	452	3,141	349 (259)	361

Table 1: Statistics for the provided data sets

For the annotation of NIs and their links to the surrounding discourse we created new guidelines as this was a novel annotation task. We adopted ideas from the annotation of co-reference information, linking locally unrealized roles to all mentions of the referents in the surrounding discourse, where available. We marked only identity relations but not part-whole or bridging relations between referents. The set of unrealized roles under consideration includes only the core arguments but not adjuncts (peripheral or extra-thematic roles in FrameNet's terminology). Possible antecedents are not restricted to noun phrases but include all constituents that can be (local) role fillers for some predicate plus complete sentences (which can sometimes fill roles such as MESSAGE).

4 Evaluation

As noted above, we allowed participants to address three different tasks: SRL only, NI only, full task. For role recognition and labeling we used a standard evaluation set-up, i.e., accuracy for role labeling and precision, recall, F-Score for role recognition.

The NI linkings were evaluated slightly differently. In the gold standard, we identified referents for null instantiations in the discourse context. In some cases, more than one referent might be appropriate, e.g., because the omitted argument refers to an entity that is mentioned multiple times in the context. In this case, a system is given credit if the NI is linked to any of these expressions. To achieve this we create equivalence sets for the referents of NIs (by annotating coreference chains). If the NI is linked to any item in the equivalence set, the link is counted as a true positive. We can then define **NI linking precision** as the number of all true positive links divided by the number of links made by a system, and **NI linking recall** as the number of true positive links divided by the number of links between an NI and its equivalence set in the gold standard. **NI linking F-Score** is then the harmonic mean between NI linking precision and recall.

Since it may sometimes be difficult to determine the correct extent of the filler of an NI, we score an automatic annotation as correct if it includes the head of the gold standard filler in the predicted filler. However, in order to not favor systems which link NIs to very large spans of text to maximize the likelihood of linking to a correct referent, we introduce a second evaluation measure, which computes the overlap (Dice coefficient) between the words in the predicted filler (P) of an NI and the words in the gold standard one (G):

$$\textbf{NI linking overlap} = \frac{2|P \cap G|}{|P| + |G|} \qquad (5)$$

Example (6) illustrates this point. The verb *won* in the second sentence evokes the Finish_competition frame whose COMPETITION role is omitted. From the context it is clear that the competition role is semantically filled by *their first TV debate* (head: *debate*) and *last night's debate* (head: *debate*) in the previous sentences. These two expressions form the equivalence set for the COMPETITION role in the last sentence. Any system that would predict a linkage to a filler that covers the head of either of these two expressions would score a true positive for this NI. However, a system that linked to *last night's debate* would have an NI linking overlap of 1 (i.e., 2*3/(3+3)) while a system linking the whole second sentence *Last night's debate was eagerly anticipated* to the NI would have an overlap of 0.67 (i.e., 2*3/(6+3))

(6) US presidential rivals Republican John McCain and Democrat Barack Obama have yesterday evening attacked each other over foreign policy and the economy, in [their first TV debate]$_{Competition}$. [Last night's debate]$_{Competition}$ was eagerly anticipated. Two national flash polls suggest that [Obama]$_{Competitor}$ won$_{Finish_competition}$ $0^{Competition}$.

5 Participating Systems

While a fair number of people expressed an interest in the task and 26 groups or individuals

downloaded the data sets, only three groups submitted results for evaluation. Feedback from the teams that downloaded the data suggests that this was due to coinciding deadlines and to the difficulty and novelty of the task. No group addressed the full task, two groups (GETARUNS++ and SEMAFOR) tackled the NI only task, and also two groups, the SRL only task (CLR and SEMAFOR).

All participating systems were built upon existing systems for semantic processing which were modified for the task. Two of the groups, GETARUNS++ and CLR, employed relatively deep semantic processing, while the third, SEMAFOR, employed a shallower probabilistic system. Different approaches were taken for NI linking. The SEMAFOR group modeled NI linking as a variant of role recognition and labeling by extending the set of potential arguments beyond the locally available arguments to also include noun phrases from the previous sentence. The system then uses, among other information, distributional semantic similarity between the heads of potential arguments and role fillers in the training data. The GETARUNS++ group applied an existing system for deep semantic processing, anaphora resolution and recognition of textual entailment, to the task. The system analyzes the sentences and assigns its own set of labels, which are subsequently mapped to frame semantic categories. For more details of the participating systems please consult the separate system papers.

6 Results and Analysis

6.1 SRL Task

	Argument Recognition			Label
	Prec.	Rec.	F1	Acc.
SHA	0.6332	0.3884	0.4812	0.3471
SEM	0.6528	0.4674	0.5448	0.4184
CLR	0.6702	0.1121	0.1921	0.1093

Table 2: Shalmaneser (SHA), SEMAFOR (SEM) and CLR performance on the SRL task (across both chapters)

The results on the SRL task are shown in Table 2. To get a better sense of how good the performance of the submitted systems was on this task, we applied the Shalmaneser statistical semantic parser (Erk and Padó, 2006) to our test data and report the results. Note, however, that we used a Shalmaneser trained only on FrameNet version 1.3 which is different from the version 1.4 alpha that was used in the task, so its results are lower than what can be expected with release 1.4 alpha.

We observe that although the SEMAFOR and the CLR systems score a higher precision than Shalmaneser for argument recognition, the SEMAFOR system scores considerably higher recall than Shalmaneser, whereas the CLR system scores a much lower recall.

6.2 NI Task

Tackling the resolution of NIs proved to be a difficult problem due to a variety of factors. First, the NI sub-task was completely new and involves several steps of linguistic processing. It also is inherently difficult in that a given FE is not always omitted with the same interpretation. For instance, the Content FE of the Awareness frame evoked by *know* is interpreted as indefinite in the blog headline *More babbling about what it means to know* but as definite in a discourse like *Don't tell me you didn't know!*. Second, prior to this SemEval task there was no full-text training data available that contained annotations with all the kinds of information that is relevant to the task, namely overt FEs, null-instantiated FEs, resolutions of null-instantiations, and coreference. Third, the data we used also represented a switch to a new domain compared to existing FrameNet full-text annotation, which comes from newspapers, travel guides, and the nuclear proliferation domain. Our most frequent frame was Observable_bodyparts, whereas it is Weapons in FrameNet full-text. Fourth, it was not well understood at the beginning of the task that, in certain cases, FrameNet's null-instantiation annotations for a given FE cannot be treated in isolation of the annotations of other FEs. Specifically, null-instantiation annotations interact with the set of relations between core FEs that FrameNet uses in its analyses. As an example, consider the CoreSet relation, which specifies that from a set of core FEs at least one must be instantiated overtly, though more of them can be. As long as one of the FEs in the set is expressed overtly, null-instantiation is not annotated for the other FEs in the set. For instance, in the Statement frame, the two FEs Topic and Message are in one CoreSet and the two FEs Speaker and Medium are in another. If a frame instance occurs with an overt Speaker and

an overt Topic, the Medium and Message FEs are not marked as null-instantiated. Automatic systems that treat each core FE separately, may propose DNI annotations for Medium and Message, resulting in false positives.

Therefore, we think that the evaluation that we initially defined was too demanding for a novel task. It would have been better to give separate scores for 1) ability to recognize when a core FE has to be treated as null-instantiated; 2) ability to distinguish INI and DNI; and 3) ability to find antecedents. The systems did have to tackle these steps anyway and an analysis of the system output shows that they did so with different success. The two chapters of our test data contained a total of 710 null instantiations, of which 349 were DNI and 361 INI. The SEMAFOR system recognized 63.4% (450/710) of the cases of NI, while the GETARUNS++ system found only 8.0% (57/710). The distinction between DNI and INI proved very difficult, too. Of the DNI and INI classifications that the SEMAFOR system proposed, 54.7% (246/450) were accurate, for GETARUNS++, the percentage is higher at 64.2% (35/57), but also based on fewer proposed classifications. A simple majority-class baseline gives a 50.8% accuracy. Interestingly, the SEMAFOR system labeled many more INIs than DNIs, thus often misclassifying DNIs as INI. The GETARUNS++ system applied both labels about equally often.

7 Conclusion

In this paper we described the SemEval-2010 shared task on "Linking Events and Their Participants in Discourse". The task is novel, in that it tackles a semantic cross-clausal phenomenon that has not been treated before in a task, namely, linking locally uninstantiated roles to their coreferents at the text level. In that sense the task represents a first step towards taking SRL beyond the sentence level. A new corpus of fiction texts has been annotated for the task with several types of semantic information: semantic argument structure, coreference chains and NIs. The results scored by the systems in the NI task and the feedback from participant teams shows that the task was more difficult than initially estimated and that the evaluation should have focused on more specific aspects of the NI phenomenon, rather than on the completeness of the task. Future work will focus on modeling the task taking this into account.

Acknowledgements

Josef Ruppenhofer and Caroline Sporleder are supported by the German Research Foundation DFG (under grant PI 154/9-3 and the Cluster of Excellence Multimodal Computing and Interaction (MMCI), respectively). Roser Morante's research is funded by the GOA project BIOGRAPH of the University of Antwerp. We would like to thank Jinho Choi, Markus Dräger, Lisa Fuchs, Philip John Gorinski, Russell Lee-Goldman, Ines Rehbein, and Corinna Schorr for their help with preparing the data and/or implementing software for the task. Thanks also to Katrin Erk and Carlo Strapparava for their support during the task organization period.

References

C. Baker, M. Ellsworth, and K. Erk. 2007. SemEval-2007 Task 19: Frame semantic structure extraction. In *Proceedings of SemEval-07*.

A. Burchardt, A. Frank, and M. Pinkal. 2005. Building text meaning representations from contextually related frames – A case study. In *Proceedings of IWCS-6*.

Katrin Erk and Sebastian Padó. 2004. A powerful and versatile XML format for representing role-semantic annotation. In *Proceedings of LREC-2004*.

K. Erk and S. Padó. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC-06*.

C.J. Fillmore and C.F. Baker. 2001. Frame semantics for text understanding. In *Proc. of the NAACL-01 Workshop on WordNet and Other Lexical Resources*.

C.J. Fillmore. 1977. Scenes-and-frames semantics, linguistic structures processing. In Antonio Zampolli, editor, *Fundamental Studies in Computer Science, No. 59*, pages 55–88. North Holland Publishing.

C.J. Fillmore. 1986. Pragmatically controlled zero anaphora. In *Proceedings of the Twelfth Annual Meeting of the Berkeley Liguistics Society*.

D. Gildea and D. Jurafsky. 2002. Automatic labeling of semantic roles. *Computational Linguistics*, 28(3):245–288.

M. Palmer, D. Dahl, R. Passonneau, L. Hirschman, M. Linebarger, and J. Dowding. 1986. Recovering implicit information. In *Proceedings of ACL-1986*.

M. Palmer. 1990. *Semantic Processing for Finite Domains*. CUP, Cambridge, England.

Josef Ruppenhofer, Caroline Sporleder, Roser Morante, Collin Baker, and Martha Palmer. 2009. Semeval-2010 task 10: Linking events and their participants in discourse. In *The NAACL-HLT 2009 Workshop on Semantic Evaluations: Recent Achievements and Future Directions (SEW-09)*.