Automatic induction of FrameNet lexical units

Marco Pennacchiotti(†), **Diego De Cao**(‡), **Roberto Basili**(‡), **Danilo Croce**(‡), **Michael Roth**(†)

(†) Computational Linguistics
Saarland University
Saarbrücken, Germany
{pennacchiotti,mroth}@coli.uni-sb.de

(‡) DISP
University of Roma Tor Vergata
Roma, Italy
{decao,basili,croce}@info.uniroma2.it

Abstract

Most attempts to integrate FrameNet in NLP systems have so far failed because of its limited coverage. In this paper, we investigate the applicability of distributional and WordNet-based models on the task of *lexical unit induction*, i.e. the expansion of FrameNet with new lexical units. Experimental results show that our distributional and WordNet-based models achieve good level of accuracy and coverage, especially when combined.

1 Introduction

Most inference-based NLP tasks require a large amount of semantic knowledge at the predicate-argument level. This type of knowledge allows to identify meaning-preserving transformations, such as active/passive, verb alternations and nominalizations, which are crucial in several linguistic inferences. Recently, the integration of NLP systems with manually-built resources at the predicate argument-level, such as FrameNet (Baker et al., 1998) and PropBank (Palmer et al., 2005) has received growing interest. For example, Shen and Lapata (2007) show the potential improvement that FrameNet can bring on the performance of a Question Answering (QA) system. Similarly, several other studies (e.g. (Bar-Haim et al., 2005; Garoufi, 2007)) indicate that frame semantics plays a central role in Recognizing Textual Entailment (RTE). Unfortunately, most attempts to integrate FrameNet or similar resources in QA and RTE systems have so far failed, as reviewed respectively in (Shen and Lapata, 2007) and (Burchardt and Frank, 2006). These studies indicate *limited coverage* as the main reason of insuccess. Indeed, the FrameNet database only contains 10,000 lexical units (LUs), far less than the 210,000 entries in WordNet 3.0. Also, frames are based on more complex information than word senses, so that their manual development is much more demanding (Burchardt et al., 2006; Subirats and Petruck, 2003).

Therefore, there is nowadays a pressing need to adopt learning approaches to extend the coverage of the FrameNet lexicon by automatically acquiring new LUs, a task we call **LU induction**, as recently proposed at SemEval-2007 (Baker et al., 2007). Unfortunately, research in this area is still somehow limited and fragmentary. The aim of our study is to pioneer in this field by proposing two unsupervised models for LU induction, one based on distributional techniques and one using WordNet as a support; and a combined model which mixes the two. The goal is to investigate to what extent distributional and WordNet-based models can be used to induce frame semantic knowledge in order to safely extend FrameNet, thus limiting the high costs of manual annotation.

In Section 2 we introduce the LU induction task and present related work. In Sections 3, 4 and 5 we present our distributional, WordNet-based and combined models. Then, in Section 6 we report experimental results and comparative evaluations. Finally, in Section 7 we draw final conclusions and outline future work.

2 Task Definition and Related Work

As defined in (Fillmore, 1985), a frame is a conceptual structure modeling a prototypical situation, evoked in texts through the occurrence of its lexical units. A *lexical unit* (LU) is a predicate that linguistically expresses the situation of the frame. Lexical units of the same frame share semantic arguments. For example the frame KILLING has lexical units such as *assassin, assassinate, blood-bath, fatal, murderer, kill, suicide* that share semantic arguments such as KILLER, INSTRUMENT, CAUSE, VICTIM. Building on this frame-semantic model, the Berkeley FrameNet project (Baker et al., 1998) has been developing a frame-semantic lexicon for the core vocabulary of English since 1997. The current FrameNet release contains 795 frames and about 10,000 LUs. Part of FrameNet is also a corpus of 135,000 annotated example sentences from the British National Corpus (BNC).

LU induction is a fairly new task. Formally, it can be defined as the task of assigning a generic lexical unit not yet present in the FrameNet database (hereafter called *unknown LU*) to the correct frame(s). As the number of frames is very large (about 800) the task is intuitively hard to solve. A further complexity regards multiple assignments. Lexical units are sometimes ambiguous and can then be mapped to more than one frame (for example the word *tea* could map both to FOOD and SOCIAL_EVENT). Also, even unambiguous words can be assigned to more than one frame – e.g. *child* maps to both KINSHIP and PEOPLE_BY_AGE.

LU induction is relevant to many NLP tasks, such as the semi-automatic creation of new FrameNets, and semantic role labelling. LU induction has been integrated at SemEval-2007 as part of the Frame Semantic Structure Extraction shared task (Baker et al., 2007), where systems are requested to assign the correct frame to a given LU, even when the LU is not yet present in FrameNet. Johansson and Nugues (2007) approach the task as a machine learning problem: a Support Vector Machine trained on existing LUs is applied to assign unknown LUs to the correct frame, using features derived from the WordNet hierarchy. Tested on the FrameNet gold standard, the method achieves an accuracy of 0.78, at the cost of a low coverage of 31% (i.e. many LUs are not assigned). Johansson and Nugues (2007) also experiment with a simple model based on standard WordNet similarity measures (Pedersen et al., 2004), achieving lower performance. Burchardt and colleagues (2005) present Detour, a rule-based system using words in a WordNet relation with the unknown LU to find the correct frame. The system achieves an accuracy of 0.39 and a coverage of 87%. Unfortunately this algorithm requires the LU to be previously disambiguated, either by hand or using contextual information.

In a departure from previous work, our first model leverages distributional properties to induce LUs, instead of relying on pre-existing lexical resources as WordNet. This guarantees two main advantages. First, it can predict a frame for any unknown LU, while WordNet based approaches can be applied only to words having a WordNet entry. Second, it allows to induce LUs in languages for which WordNet is not available or has limited coverage. Our second WordNet-based model uses sense information to characterize the frame membership for unknown LU, by adopting a semantic similarity measure which is sensitive to *all* the known LUs of a frame.

3 Distributional model

The basic idea behind the distributional approach is to induce new LUs by modelling existing frames and unknown LUs in a semantic space, where they are represented as distributional co-occurrence vectors computed over a corpus.

Semantic spaces are widely used in NLP for representing the meaning of words or other lexical entities. They have been successfully applied in several tasks, such as information retrieval (Salton et al., 1975) and harvesting thesauri (Lin, 1998). The intuition is that the meaning of a word can be described by the set of textual contexts in which it appears (*Distributional Hypothesis* (Harris, 1964)), and that words with similar vectors are semantically related. In our setting, the goal is to find a semantic space model able to capture the notion of *frame* – i.e. the property of *"being characteristic of a frame"*. In such a model, an unknown LU is induced by first computing the similarity between its vector and the vectors of the existing frames, and then assigning the LU to the frame with the highest similarity.

3.1 Assigning unknown LUs to frames

In our model, a LU l is represented by a vector $\vec{l}$ whose dimensions represent the set of contexts C of the semantic space. The value of each dimension is given by the co-occurrence value of the LU with a contextual feature $c \in C$, computed over a large corpus using an association measure. We experiment with two different association measures: normalized frequency and pointwise mutual information. We approximate these measures by using Maximum Likelihood Estimation, as follows:

$$
\begin{aligned}
F(l,c) &=_{MLE} \frac{|l,c|}{|*,*|} \\
MI(l,c) &=_{MLE} \frac{|l,c||*,*|}{|*,c||l,*|}
\end{aligned} \quad (1)
$$

where $|l,c|$ denotes the co-occurrence counts of the pair (l,c) in the corpus, $|*,c| = \sum_{l \in L} |l,c|$, $|l,*| = \sum_{c \in C} |l,c|$ and finally $|*,*| = \sum_{l \in L, c \in C} |l,c|$.

A frame f is modeled by a vector $\vec{f}$, representing the distributional profile of the frame in the semantic space. We here assume that a frame can be fully described by the set of its lexical units F. We implement this intuition by computing $\vec{f}$ as the weighted centroid of the set F, as follows:

$$\vec{f} = \sum_{l \in F} w_{lf} * \vec{l} \quad (2)$$

where w_{lf} is a weighting factor, accounting for the relevance of a given lexical unit with respect to the frame, estimated as:

$$w_{lf} = \frac{|l|}{\sum_{l \in F} |l|} \quad (3)$$

where $|l|$ denotes the counts of l in the corpus. From a more cognitive perspective, the vector $\vec{f}$ represents the prototypical lexical unit of the frame.

Given the set of all frames $\mathcal{N}$ and an unknown lexical unit ul, we assign ul to the frame $fmax_{ul}$ which is distributionally most similar – i.e. we intuitively map an unknown lexical unit to the frame whose prototypical lexical unit $\vec{f}$ has the highest similarity with $\vec{ul}$:

$$fmax_{ul} = argmax_{f \in \mathcal{N}} sim_D(\vec{ul}, \vec{f}) \quad (4)$$

In our model, we used the traditional cosine similarity:

$$sim_{cos}(ul, f) = \frac{\vec{ul} \cdot \vec{f}}{|\vec{ul}| * |\vec{f}|} \quad (5)$$

3.2 Choosing the space

Different types of contexts C define spaces with different semantic properties. We are here looking for a space able to capture the properties which characterise a frame. The most relevant of these properties is that LUs in the same frame tend to be either co-occurring or substitutional words (e.g. *assassin/kill* or *assassinate/kill*) – i.e. they are either in paradigmatic and syntagmatic relation. In an ideal space, a high similarity value sim_D would be then given both to *assassinate/kill* and to *assassin/kill*. We explore three spaces which seem to capture the above property well:

Word-based space: Contexts are words appearing in a n-window of the lexical unit. Such spaces model a generic notion of *semantic relatedness*. Two LUs close in the space are likely to be related by some type of generic semantic relation, either paradigmatic (e.g. synonymy, hyperonymy, antonymy) or syntagmatic (e.g. meronymy, conceptual and phrasal association).[1]

Syntax-based space: Contexts are syntactic relations (e.g. *X-VSubj-man* where X is the LU), as described in (Padó, 2007). These spaces are good at modeling *semantic similarity*. Two LUs close in the space are likely to be in a paradigmatic relation, i.e. to be close in a is-a hierarchy (Budanitsky and Hirst, 2006; Lin, 1998; Padó, 2007). Indeed, as contexts are syntactic relations, targets with the same part of speech are much closer than targets of different types.

Mixed space: In a combination of the two above spaces, contexts are words connected to the LU by a dependency path of at most length n. Unlike word-based spaces, contexts are selected in a more principled way: only syntactically related words are contexts, while other (possibly noisy) material is filtered out. Unlike syntax-based spaces, the context c does not explicitly state the type of syntactic relation with the LU: this usually allows to capture both paradigmatic and syntagmatic relations.

4 WordNet-based model

In a departure from previous work, our WordNet-based model does not rely on standard WordNet similarity measures (Pedersen et al., 2004), as these measures can only be applied to *pairs* of words, while we here need to capture the meaning of whole frames, which typically consist of larger sets of LUs. Our intuition is that senses able to evoke a frame can be detected via WordNet, by jointly considering the WordNet synsets activated by *all* LUs of the frame.

We implement this intuition in a weakly-supervised model, where each frame f is represented as a set of specific sub-graphs of the WordNet

[1] See (Padó, 2007; Sahlgren, 2006) for an in depth analysis.

hyponymy hierarchy. As different parts of speech have different WordNet hierarchies, we build a sub-graph for each of them: S_f^n for nouns, S_f^v for verbs and S_f^a for adjectives.[2] These sub-graphs represent the lexical semantic properties characterizing the frame. An unknown LU ul of a given part of speech is assigned to the frame whose corresponding sub-graph is semantically most similar to one of the senses of ul:

$$fmax_{ul} = argmax_{f \in \mathcal{N}} sim_{WN}(ul, f) \qquad (6)$$

where sim_{WN} is a WordNet-based similarity measure. In the following subsections we will describe how we build sub-graphs and model the similarity measure for the different part of speech.

Figure 1 reports an excerpt of the noun sub-graph for the frame PEOPLE_BY_AGE, covering the suitable senses of its nominal LUs $\{adult, baby, boy, kid, youngster, youth\}$. The relevant senses (e.g. sense 1 of $youth$ out of the 6 potential ones) are generally selected, as they share the most specific generalizations in WordNet with the other words.

Nouns. To compute similarity for *nouns* we adopt *conceptual density* (*cd*) (Agirre and Rigau, 1996), a semantic similarity model previously applied to word sense disambiguation tasks.

Given a frame f and its set of nominal lexical units F_n, the nominal subgraph S_f^n is built as follows. All senses of all words in F_n are activated in WordNet. All hypernyms H_f^n of these senses are then retrieved. Every synset $\sigma \in H_f^n$ is given a *cd* score, representing the *density* of the WordNet sub-hierarchy rooted at σ in representing the set of nouns F_n. The intuition behind this model is that the larger the number of LUs in F_n that are generalized by σ is, the better it captures the lexical semantics intended by the frame f. Broader generalizations are penalized as they give rise to bigger hierarchies, not well correlated with the full set of targets F_n.

To build the final sub-graph S_f^n, we apply the greedy algorithm proposed by Basili and colleagues (2004). It first computes the set of WordNet synsets that generalize at least two LUs in F_n, and then selects the subset of most dense ones $S_f^n \subset H_f^n$ that cover F_n. If a LU has no common hypernym with other members of F_n, it is not represented in S_f^n, and its similarity is set to 0 . S_f^n disambiguates words in F_n as only the lexical senses with at least one hypernym in S_f^n are considered.

Figure 1 shows the nominal sub-graph automatically derived using conceptual density for the frame PEOPLE_BY_AGE. The word *boy* is successfully disambiguated, as its only hypernym in the sub-graph refers to its third sense (*a male human offspring*) which correctly maps to the given frame. Notice that this model departs from the first sense heuristics largely successful in word sense disambiguation: most frames in fact are characterized by non predominant senses. The only questionable disambiguation is for the word *adult*: the wrong sense (*adult mammal*) is selected. However, even in these cases, the *cd* values are very low (about 10^{-4}), so that they do not impact much on the quality of the resulting inference.

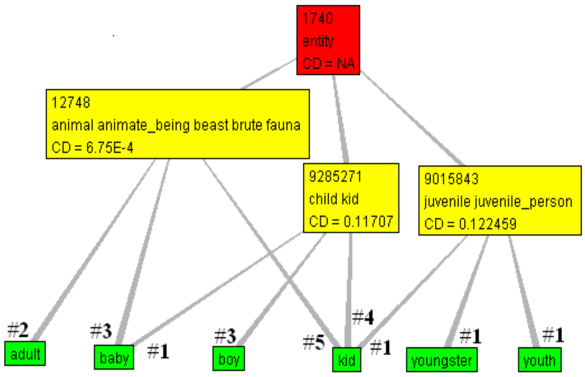

Figure 1: The noun sub-graph for the frame PEOPLE_BY_AGE as evoked by a subset of the words. Sense numbers #n refers to WordNet 2.0.

Using this model, LU induction is performed as follows. Given an unknown lexical unit ul, for each frame $f \in \mathcal{N}$ we first build the sub-graph S_f^n from the set $F_n \cup \{ul\}$. We then compute $sim_{WN}(f, ul)$ as the maximal *cd* of any synset $\sigma \in S_f^n$ that generalizes one of the lexical senses of ul. In the example *baby* would receive a score of 0.117 according to its first sense in WordNet 2.0 ("*baby,babe,infant*"). In a final step, we assign the LU to the most similar frame, according to Eq. 6

Verbs and Adjectives. As the conceptual density algorithm can be used only for nouns, we apply different similarity measures for verbs and adjectives.

[2] Our WordNet model does not cover the limited number of LUs which are not nouns, verbs or adjectives.

For *verbs* we exploit the co-hyponymy relation: the sub-graph S_f^v is given by all hyponyms of all verbs F_v in the frame f. Similarity $sim_{WN}(f, ul)$ is computed as follows:

$$sim_{WN}(ul, f) = \begin{cases} 1 & \textbf{iff } \exists K \subset F \text{ such that} \\ & |K| > \tau \textbf{ AND} \\ & \forall l \in K, l \text{ is a co-hyponym of } ul \\ \epsilon & \text{otherwise} \end{cases} \tag{7}$$

As for *adjectives*, WordNet does not provide a hyponymy hierarchy. We then compute similarity simply on the basis of the synonymy relation, as follows:

$$sim_{WN}(ul, f) = \begin{cases} 1 & \textbf{iff } \exists l \in F \text{ such that} \\ & l \text{ is a synonym of } ul \\ \epsilon & \text{otherwise} \end{cases} \tag{8}$$

5 Combined model

The methods presented so far use two independent information sources to induce LUs: distributional similarity sim_D and WordNet similarity sim_{WN}. We also build a joint model, leveraging both approaches: we expect the combination of different information to raise the overall performance. We here choose to combine the two approaches using a simple back-off model, that uses the WordNet-based model as a default and backs-off to the distributional one when no frame is proposed by the former. The intuition is that WordNet should guarantee the highest precision in the assignment, while distributional similarity should recover cases of low coverage.

6 Experiments

In this section we present a comparative evaluation of our models on the task of inducing LUs, in a leave-one-out setting over a reference gold standard.

6.1 Experimental Setup

Our gold standard is the FrameNet 1.3 database, containing 795 frames and a set L of 7,522 unique LUs (in all there are 10,196 LUs possibly assigned to more than one frame). Given a lexical unit $l \in L$, we simulate the induction task by executing a leave-one-out procedure, similarly to Burchardt and colleagues (2005). First, we remove l from all its original frames. Then, we ask our models to reassign it to the most similar frame(s) f, according to the similarity measure[3]. We repeat this procedure for all lexical units. Though our experiment is not completely realistic (we test over LUs already in FrameNet), it has the advantage of a reliable gold standard produced by expert annotators. A second, more realistic, small-scale experiment is described in Section 6.2.

We compute *accuracy* as the fraction of LUs in L that are correctly re-assigned to the original frame. Accuracy is computed at different levels k: a LU l is correctly assigned if its gold standard frame appears among the best-*k* frames f ranked by the model using the $sim(l, f)$ measure. As LUs can have more than one correct frame, we deem as correct an assignment for which at least one of the correct frames is among the best-*k*.

We also measure *coverage*, intended as the percentage of LUs that have been assigned to at least one frame by the model. Notice that when no sense preference can be found above the threshold ϵ, the WordNet-based model cannot predict any frame, thus decreasing coverage.

We present results for the following models and parametrizations (further parametrizations have revealed comparable performance).

Dist-word : the word-based space described in Section 3. Contextual features correspond to the set of the 4,000 most frequent words in the BNC.[4] The association measure between LUs and contexts is the pointwise mutual information. Valid contexts for LUs are fixed to a 20-window.

Dist-syntax : the syntax-based space described in Section 3. Context features are the 10,000 most frequent syntactic relations in the BNC[5]. As association measure we apply log-likelihood ratio (Dunning, 1993) to normalized frequency. Syntactic relations are extracted using the Minipar parser.

Dist-mixed : the mixed space described in Sec-

[3] In the distributional model, we recompute the centroids for each frame f in which the LU appeared, applying Eq. 2 to the set $F - \{l\}$.

[4] We didn't use the FrameNet corpus directly, as it is too small to obtain reliable statistics.

[5] Specifically, we use the minimum context selection function and the plain path value function described in Pado (2007).

tion 3. As for the *Dist-word* model, contextual features are 4,000 and pointwise mutual information is the association measure. The maximal dependency path length for selecting each context word is 3. Syntactic relations are extracted using Minipar.

WNet-full : the WordNet based model described in Section 4.

WNet-bsense : this model is computed as *WNet-full* but using only the most frequent sense for each LU as defined in WordNet.

Combined : the combined method presented in Section 5. Specifically, it uses *WNet-full* as a default and *Dist-word* as back-off.

Baseline-rnd : a baseline model, randomly assigning LUs to frames.

Baseline-mostfreq : a model predicting as best-*k* frames the most likely ones in FrameNet – i.e. those containing the highest number of LUs.

6.2 Experimental Results

Table 1 reports accuracy and coverage results for the different models, considering only 6792 LUs with frequency higher than 5 in the BNC, and frames with more than 2 lexical units (to allow better generalizations in all models). Results show that all our models largely outperform both baselines, achieving a good level of accuracy and high coverage. In particular, accuracy for the best-10 frames is high enounugh to support tasks such as the semi-automatic creation of new FrameNets. This claim is supported by a further task-driven experiment, in which we asked 3 annotators to assign 60 unknown LUs (from the Detour system log) to frames, with and without the support of the *Dist-word* model's predictions as suggestions[6]. We verified that our model guarantee an annotation speed-up of 25% – i.e. in average an annotator saves 25% of annotation time by using the system's suggestions.

Distributional vs. WordNet-based models. WordNet-based models are significantly better than distributional ones, for several reasons. First, distributional models acquire information only from the contexts in the corpus. As we do not use a FrameNet annotated corpus, there is no guarantee that the usage of a LU in the texts reflects exactly the semantic properties of the LU in FrameNet. In the extreme cases of polysemous LUs, it may happen that the textual contexts refer to senses which are not accounted for in FrameNet. In our study, we explicitly ignore the issue of polisemy, which is a notoriously hard task to solve in semantics spaces (see (Schütze, 1998)), as the occurrences of different word senses need to be clustered separately. We will approach the problem in future work. The WordNet-based model suffers from the problem of polisemy to a much lesser extent, as all senses are explicitly represented and separated in WordNet, including those related to the FrameNet gold standard.

A second issue regards data sparseness. The vectorial representation of LUs with few occurrences in the corpus is likely to be semantically incomplete, as not enough statistical evidence is available. Particularly skewed distributions can be found when some frames are very rarely represented in the corpus. A more in-depth descussion on these two issues is given later in this section.

Regarding the WordNet-based models, *WNet-full* in most cases outperforms *WNet-bsense*. The first sense heuristic does not seem to be as effective as in other tasks, such as Word Sense Disambiguation. Although sense preferences (or predominance) across two general purpose resources, such as WordNet and FrameNet, should be a useful hint, the conceptual density algorithm seems to produce better distributions (i.e. higher accuracy), especially when several solutions are considered. Indeed, for many LUs the first WordNet sense is not the one represented in the FrameNet database.

As for distributional models, results show that the *Dist-word* model performs best. In general, syntactic relations (*Dist-syntax* model) do not help to capture frame semantic properties better than a simple window-based approach. This seems to indicate that LUs in a same frame are related both by paradigmatic and syntagmatic relations, in accordance to the definition given in Section 3.2 – i.e. they are mostly semantically *related*, but not *similar*.

Coverage. Distributional models show a coverage 15% higher than WordNet-based ones. Indeed, as far as corpus evidence is available (i.e. the unknown LU appears in the corpus), distributional methods are always able to predict a frame. WordNet-based mod-

[6] For this purpose, the dataset is evenly split in two parts.

MODEL	B-1	B-2	B-3	B-4	B-5	B-6	B-7	B-8	B-9	B-10	COVERAGE
Dist-word	0.27	0.36	0.42	0.46	0.49	0.51	0.53	0.55	0.56	0.57	95%
Dist-syntax	0.22	0.29	0.34	0.38	0.41	0.44	0.46	0.48	0.50	0.51	95%
Dist-mixed	0.25	0.35	0.40	0.44	0.47	0.49	0.51	0.53	0.54	0.56	95%
WNet-full	0.47	0.59	0.65	0.69	0.72	0.73	0.75	0.76	0.77	0.78	80%
WNet-bsense	0.52	0.61	0.64	0.66	0.67	0.68	0.69	0.69	0.70	0.70	72%
Combined	0.43	0.54	0.60	0.64	0.66	0.68	0.70	0.71	0.72	0.73	95%
Baseline-rnd	*0.02*	*0.03*	*0.05*	*0.06*	*0.08*	*0.10*	*0.11*	*0.12*	*0.14*	*0.15*	
Baseline-mostfreq	*0.02*	*0.05*	*0.07*	*0.08*	*0.10*	*0.11*	*0.13*	*0.14*	*0.15*	*0.17*	

Table 1: Accuracy and coverage of different models on best-*k* ranking with frequency threshold 5 and frame threshold 2

els cannot make predictions in two specific cases. First, when the LU is not present in WordNet. Second, when the function sim_{WN} does not has sufficient relational information to find a similar frame. This second factor is particularly evident for adjectives, as Eq. 8 assigns a frame only when a synonym of the unknown LU is found. It is then not surprising that 68% of the missed assignment are indeed adjectives.

Results for the *Combined* model suggest that the integration of distributional and WordNet-based methods can offer a viable solution to the coverage problem, as it achieves an accuracy comparable to the pure WordNet approaches, while keeping the coverage high.

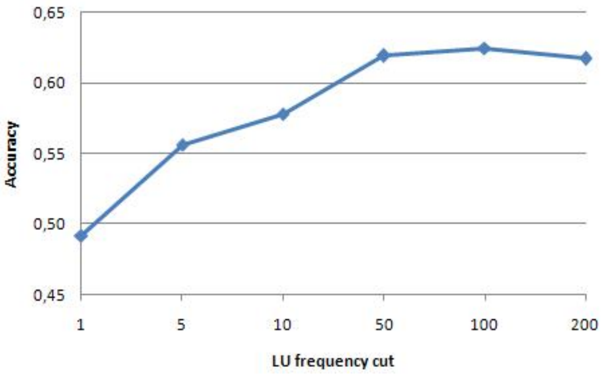

Figure 2: *Dist-word* model accuracy at different LU frequency cuts.

Data Sparseness. A major issue when using distributional approaches is that words with low frequency tend to have a very sparse non-meaningful representation in the vector space. This highly impacts on the accuracy of the models. To measure the impact of data sparseness, we computed the accuracy at different frequency cuts – i.e. we exclude LUs below a given frequency threshold from centroid computation and evaluation. Figure 2 reports the results for best-*10* assignment at different cuts, for the *Dist-word* model. As expected, accuracy improves by excluding infrequent LUs. Only at a frequency cut of 200 performance becomes stable, as statistical evidence is enough for a reliable prediction. Yet, in a real setting the improvement in accuracy implies a lower coverage, as the system would not classify LUs below the threshold. For example, by discarding LUs occurring less than 200 times in the corpus, we obtain a +0.12 improvement in accuracy, but the coverage decreases to 57%. However, uncovered LUs are also the most rare ones and their relevance in an application may be negligible.

Lexical Semantics, Ambiguity and Plausible Assignments. The overall accuracies achieved by our methods are "pessimistic", in the sense that they should be intended as lower-bounds. Indeed, a qualitative analysis of erroneous predictions reveals that in many cases the frame assignments produced by the models are semantically plausible, even if they are considered incorrect in the leave-one-out test. Consider for example the LU *guerrilla*, assigned in FrameNet to the frame PEOPLE BY VOCATION. Our mixed model proposes as two most similar frames MILITARY and TERRORISM, which could still be considered plausible assignment. The same holds for the LU *caravan*, for which the most similar frame is VEHICLE, while in FrameNet the LU is assigned only to the frame BUILDINGS. These cases are due to the low FrameNet coverage, i.e LUs are not fully annotated and they appear only in a subset of their potential frames. The real accuracy of our models is therefore expected to be higher.

To explore the issue, we carried out a qualitative analysis of 5 words (i.e. *abandon.v*, *accuse.v*, *body.n*, *charge.v* and *partner.n*). For each of them, we randomly picked 60 sentences from the BNC corpus, and asked two human annotators to assign to the correct frame the occurrence of the word in the given sentence. For 2 out of 5 words, no frame could be found for most of the sentences, suggesting that the most frequent frames for these words were missing from FrameNet[7]. We can then conclude that 100% accuracy cannot be considered as the upperbound of our experiment, as word usage in texts is not well reflected in the FrameNet modelling.

Further experiments. We also tested our models on a realistic gold-standard set of 24 unknown LUs extracted from the SemEval-2007 corpus (Baker et al., 2007). These are words not present in FrameNet 1.3 which have been assigned by human annotators to an existing frame[8]. *WNet-full* achieves an accuracy of 0.25 for best-1 and 0.69 for best-10, with a coverage of 67%. A qualitative analysis showed that the lower performance wrt to our main experiment is due to higher ambiguity of the LUs (e.g. we assign *tea* to SOCIAL_EVENT instead of FOOD).

Comparison to other approaches. We compare our models to the system presented by Johansson and Nugues (2007) and Burchardt and colleagues (2005). Johansson and Nugues (2007) evaluate their machine learning system using 7,000 unique LUs to train the Support Vector Machine, and the remaining LUs as test. They measure accuracy at different coverage levels. At 80% coverage accuracy is about 0.42, 10 points below our best WordNet-based system. At 90% coverage, the system shows an accuracy below 0.10 and is significantly outperformed by both our distributional and combined methods. These results confirm that WordNet-based approaches, while being highly accurate wrt distributional ones, present strong weaknesses as far as coverage is concerned. Furthermore, Johansson and Nugues (2007) show that their machine learning approach outperforms a simple approach based on WordNet similarity: thus, our results indirectly prove that our WordNet-based method is more effective than the application of the similarity measure presented in (Pedersen et al., 2004).

We also compare our results to those reported by Burchardt and colleagues (2005) for Detour. Though the experimental setting is slightly different (LU assignment is done at the text-level), they use the same gold standard and leave-one-out technique, reporting a best-1 accuracy of 0.38 and a coverage of 87%. Our WordNet-based models significantly outperform Detour on best-1 accuracy, at the cost of lower coverage. Yet,our *combined* model is significantly better both on accuracy (+5%) and coverage (+8%). Also, in most cases Detour cannot predict more than one frame (best-1), while our accuracies can be improved by relaxing to any best-k level.

7 Conclusions

In this paper we presented an original approach for FrameNet LU induction. Results show that models combining distributional and WordNet information offer the most viable solution to model the notion of frame, as they allow to achieve a reasonable trade-off between accuracy and coverage. We also showed that in contrast to previous work, simple semantic spaces are more helpful than complex syntactic ones. Results are accurate enough to support the creation and the development of new FrameNets.

As future work, we will evaluate new types of spaces (e.g. dimensionality reduction methods) to improve the generalization capabilities of the space models. We will also address the data sparseness issue, by testing smoothing techniques to better model low frequency LUs. Finally, we will implement the presented models in a complex architecture for semi-supervised FrameNets development, both for specializing the existing English FrameNet in specific domains, and for creating new FrameNets in other languages.

Acknowledgements

This work has partly been funded by the German Research Foundation DFG (grant PI 154/9-3). Thanks to Richard Johansson and Aljoscha Burchardt for providing the data of their systems.

[7] Note that the need of new frames to account for semantic phenomena in free texts has been also demonstrated by the SemEval-2007 competition.

[8] The set does not contain 4 LUs which have no frame in FrameNet.

References

E. Agirre and G. Rigau. 1996. Word Sense Disambiguation using Conceptual Density. In *Proceedings of COLING-96*, Copenhagen, Denmark.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL*, Montreal, Canada.

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 19: Frame Semantic Structure Extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic, June.

Roy Bar-Haim, Idan Szpektor, and Oren Glickman. 2005. Definition and Analysis of Intermediate Entailment Levels. In *ACL-05 Workshop on Empirical Modeling of Semantic Equivalence and Entailment*, Ann Arbor, Michigan.

R. Basili, M. Cammisa, and F.M. Zanzotto. 2004. A semantic similarity measure for unsupervised semantic disambiguation. In *Proceedings of LREC-04*, Lisbon, Portugal.

Alexander Budanitsky and Graeme Hirst. 2006. Evaluating WordNet-based measures of semantic distance. *Computational Linguistics*, 32(1):13–47.

Aljoscha Burchardt and Anette Frank. 2006. Approximating Textual Entailment with LFG and FrameNet Frames. In *Proceedings of PASCAL RTE2 Workshop*.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet Detour to FrameNet. In *Sprachtechnologie, mobile Kommunikation und linguistische Resourcen*, volume 8 of *Computer Studies in Language and Speech*. Peter Lang, Frankfurt/Main.

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of LREC*, Genova, Italy.

Ted Dunning. 1993. Accurate methods for the statistics of surprise and coincidence. *Computational Linguistics*, 18(1):61–74.

Charles J. Fillmore. 1985. Frames and the semantics of understanding. *Quaderni di Semantica*, 4(2):222–254.

K. Garoufi. 2007. Towards a better understanding of applied textual entailment: Annotation and evaluation of the rte-2 dataset. *M.Sc. thesis*, saarland university.

Zellig Harris. 1964. Distributional structure. In Jerrold J. Katz and Jerry A. Fodor, editors, *The Philosophy of Linguistics*, New York. Oxford University Press.

Richard Johansson and Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. In *Proceedings of the Workshop on Building Frame-semantic Resources for Scandinavian and Baltic Languages, at NODALIDA*, Tartu, Estonia, May 24.

Dekang Lin. 1998. Automatic retrieval and clustering of similar word. In *Proceedings of COLING-ACL*, Montreal, Canada.

Sebastian Padó. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Saarland University.

M. Palmer, D. Gildea, and P. Kingsbury. 2005. The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1).

Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi. 2004. WordNet::Similarity - Measuring the Relatedness of Concept. In *Proc. of 5th NAACL*, Boston, MA.

Magnus Sahlgren. 2006. *The Word-Space Model*. Department of Linguistics, Stockholm University.

G. Salton, A. Wong, and C. Yang. 1975. A vector space model for automatic indexing. *Communications of the ACM*, 18:613620.

Hinrich Schütze. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1):97–124.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of EMNLP-CoNLL*, pages 12–21, Prague.

C. Subirats and M. Petruck. 2003. Surprise! Spanish FrameNet! In *Proceedings of the Workshop on Frame Semantics at the XVII. International Congress of Linguists*, Prague.