Graph Alignment for Semi-Supervised Semantic Role Labeling

Hagen Fürstenau
Dept. of Computational Linguistics
Saarland University
Saarbrücken, Germany
hagenf@coli.uni-saarland.de

Mirella Lapata
School of Informatics
University of Edinburgh
Edinburgh, UK
mlap@inf.ed.ac.uk

Abstract

Unknown lexical items present a major obstacle to the development of broad-coverage semantic role labeling systems. We address this problem with a semi-supervised learning approach which acquires training instances for unseen verbs from an unlabeled corpus. Our method relies on the hypothesis that unknown lexical items will be structurally and semantically similar to known items for which annotations are available. Accordingly, we represent known and unknown sentences as graphs, formalize the search for the most similar verb as a graph alignment problem and solve the optimization using integer linear programming. Experimental results show that role labeling performance for unknown lexical items improves with training data produced automatically by our method.

1 Introduction

Semantic role labeling, the task of automatically identifying the semantic roles conveyed by sentential constituents, has recently attracted much attention in the literature. The ability to express the relations between predicates and their arguments while abstracting over surface syntactic configurations holds promise for many applications that require broad coverage semantic processing. Examples include information extraction (Surdeanu et al., 2003), question answering (Narayanan and Harabagiu, 2004), machine translation (Boas, 2005), and summarization (Melli et al., 2005).

Much progress in the area of semantic role labeling is due to the creation of resources like FrameNet (Fillmore et al., 2003), which document the surface realization of semantic roles in real world corpora. Such data is paramount for developing semantic role labelers which are usually based on supervised learning techniques and thus require training on role-annotated data. Examples of the training instances provided in FrameNet are given below:

(1) a. If [you]$_{Agent}$ [carelessly]$_{Manner}$ chance going back there, you deserve what you get.

b. Only [one winner]$_{Buyer}$ purchased [the paintings]$_{Goods}$

c. [Rachel]$_{Agent}$ injured [her friend]$_{Victim}$ [by closing the car door on his left hand]$_{Means}$.

Each verb in the example sentences evokes a *frame* which is situation-specific. For instance, *chance* evokes the *Daring* frame, *purchased* the *Commerce_buy* frame, and *injured* the *Cause_harm* frame. In addition, frames are associated with semantic roles corresponding to salient entities present in the situation evoked by the predicate. The semantic roles for the frame *Daring* are *Agent* and *Manner*, whereas for *Commerce_buy* these are *Buyer* and *Goods*. A system trained on large amounts of such *hand-annotated* sentences typically learns to identify the boundaries of the arguments of the verb predicate (argument identification) and label them with semantic roles (argument classification).

A variety of methods have been developed for semantic role labeling with reasonably good performance (F_1 measures in the low 80s on standard test collections for English; we refer the interested reader to the proceedings of the SemEval-2007 shared task (Baker et al., 2007) for an overview of the state-of-the-art). Unfortunately, the reliance on training data, which is both difficult and highly expensive to produce, presents a major obstacle to the widespread application of semantic role labeling across different languages and text genres. The English FrameNet (version 1.3) is not

a small resource — it contains 502 frames covering 5,866 lexical entries and 135,000 annotated sentences. Nevertheless, by virtue of being under development it is incomplete. Lexical items (i.e., predicates evoking existing frames) are missing as well as frames and annotated sentences (their number varies greatly across lexical items). Considering how the performance of supervised systems degrades on out-of-domain data (Baker et al., 2007), not to mention unseen events, semi-supervised or unsupervised methods seem to offer the primary near-term hope for broad coverage semantic role labeling.

In this work, we develop a semi-supervised method for enhancing FrameNet with additional annotations which could then be used for classifier training. We assume that an initial set of labeled examples is available. Then, faced with an unknown predicate, i.e., a predicate that does not evoke any frame according to the FrameNet database, we must decide (a) which frames it belongs to and (b) how to automatically annotate example sentences containing the predicate. We solve both problems jointly, using a graph alignment algorithm. Specifically, we view the task of inferring annotations for new verbs as an instance of a structural matching problem and follow a graph-based formulation for pairwise global network alignment (Klau, 2009). Labeled and unlabeled sentences are represented as dependency-graphs; we formulate the search for an optimal alignment as an integer linear program where different graph alignments are scored using a function based on semantic and structural similarity. We evaluate our algorithm in two ways. We assess how accurate it is in predicting the frame for an unknown verb and also evaluate whether the annotations we produce are useful for semantic role labeling.

In the following section we provide an overview of related work. Next, we describe our graph-alignment model in more detail (Section 3) and present the resources and evaluation methodology used in our experiments (Section 4). We conclude the paper by presenting and discussing our results.

2 Related Work

Much previous work has focused on creating FrameNet-style annotations for languages other than English. A common strategy is to exploit parallel corpora and transfer annotations from English sentences onto their translations (Padó and Lapata, 2006; Johansson and Nugues, 2006). Other work attempts to automatically augment the English FrameNet in a monolingual setting either by extending its coverage or by creating additional training data.

There has been growing interest recently in determining the frame membership for unknown predicates. This is a challenging task, FrameNet currently lists 502 frames with example sentences which are simply too many (potentially related) classes to consider for a hypothetical system. Moreover, predicates may have to be assigned to multiple frames, on account of lexical ambiguity. Previous work has mainly used WordNet (Fellbaum, 1998) to extend FrameNet. For example, Burchardt et al. (2005) apply a word sense disambiguation system to annotate predicates with a WordNet sense and hyponyms of these predicates are then assumed to evoke the same frame. Johansson and Nugues (2007) treat this problem as an instance of supervised classification. Using a feature representation based also on WordNet, they learn a classifier for each frame which decides whether an unseen word belongs to the frame or not. Pennacchiotti et al. (2008) create "distributional profiles" for frames. Each frame is represented as a vector, the (weighted) centroid of the vectors representing the meaning of the predicates it evokes. Unknown predicates are then assigned to the most similar frame. They also propose a WordNet-based model that computes the similarity between the synsets representing an unknown predicate and those activated by the predicates of a frame.

All the approaches described above are type-based. They place more emphasis on extending the lexicon rather than the annotations that come with it. In our earlier work (Fürstenau and Lapata, 2009) we acquire new training instances, by projecting annotations from existing FrameNet sentences to new unseen ones. The proposed method is token-based, however, it only produces annotations for known verbs, i.e., verbs that FrameNet lists as evoking a given frame.

In this paper we generalize the proposals of Pennacchiotti et al. (2008) and Fürstenau and Lapata (2009) in a unified framework. We create training data for semantic role labeling of unknown predicates by projection of annotations from labeled onto unlabeled data. This projection is con-

ceptualized as a graph alignment problem where we seek to find a globally optimal alignment subject to semantic and structural constraints. Instead of predicting the same frame for each occurence of an unknown predicate, we consider a set of candidate frames and allow projection from any labeled predicate that can evoke one of these frames. This allows us to make instance-based decisions and thus account for predicate ambiguity.

3 Graph Alignment Method

Our approach acquires annotations for an unknown frame evoking verb by selecting sentences featuring this verb from a large unlabeled corpus (the *expansion* corpus). The choice is based upon a measure of similarity between the predicate-argument structure of the unknown verb and those of similar verbs in a manually labeled corpus (the *seed* corpus). We formulate the problem of finding the most similar verbs as the search for an optimal graph alignment (we represent labeled and unlabeled sentences as dependency graphs). Conveniently, this allows us to create labeled training instances for the unknown verb by projecting role labels from the most similar seed instance. The annotations can be subsequently used for training a semantic role labeler.

Given an unknown verb, the first step is to narrow down the number of frames it could potentially evoke. FrameNet provides definitions for more than 500 frames, of which we entertain only a small number. This is done using a method similar to Pennacchiotti et al. (2008). Each frame is represented in a semantic space as the centroid of the vectors of all its known frame evoking verbs. For an unknown verb we then consider as frame candidates the k closest frames according to a measure of distributional similarity (which we compute between the unknown verb's vector and the frame centroid vector). We provide details of the semantic space we used in our experiments in Section 4.

Next, we compare each sentence featuring the unknown verb in question to labeled sentences featuring known verbs which according to FrameNet evoke any of the k candidate frames. If sufficiently similar seeds exist, the unlabeled sentence is annotated by projecting role labels from the most similar one. The similarity score of this best match is recorded as a measure of the quality (or reliability) of the new instance. After carrying out this procedure for all sentences in the expansion corpus featuring an unknown verb, we collect the highest scoring new instances and add them back to our seed corpus as new training items. In the following we discuss in more detail how the similarity of predicate-argument structures is assessed.

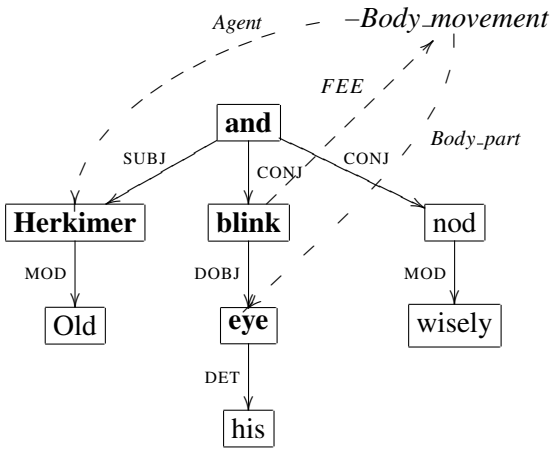

Figure 1: Annotated dependency graph for the sentence *Old Herkimer blinked his eye and nodded wisely.* The alignment domain is indicated in bold face. Labels in italics denote frame roles, whereas grammatical roles are rendered in small capitals. The verb *blink* evokes the frame *Body_Movement*.

3.1 Alignment Scoring

Let s be a semantically labeled dependency graph in which node n_{FEE} represents the frame evoking verb. Here, we use the term "labeled" to indicate that the graph contains semantic role labels in addition to grammatical role labels (e.g., subject or object). Let g be an unlabeled graph and n_{target} a verbal node in it. The "unlabeled" graph contains grammatical roles but no semantic roles. We wish to find an alignment between the predicate-argument structures of n_{FEE} and n_{target}, respectively. Such an alignment takes the form of a function σ from a set M of nodes of s (the *alignment domain*) to a set N of nodes of g (the *alignment range*). These two sets represent the relevant predicate-argument structures within the two graphs; nodes that are not members of these sets are excluded from any further computations.

If there were no mismatches between (frame) semantic arguments and syntactic arguments, we would expect all roles in s to be instantiated by syntactic dependents in n_{FEE}. This is usually the case but not always. We cannot therefore sim-

ply define M as the set of direct dependents of the predicate, but also have to consider *complex paths* between n_{FEE} and role bearing nodes. An example is given in Figure 1, where the role *Agent* is filled by a node which is not dominated by the frame evoking verb *blink*; instead, it is connected to *blink* by the complex path (CONJ^{-1}, SUBJ). For a given seed s we build a list of all such complex paths and also include all nodes of s connected to n_{FEE} by one of these paths. We thus define the alignment domain M as:

1. the predicate node n_{FEE}
2. all direct dependents of n_{FEE}, except auxiliaries
3. all nodes on complex paths originating in n_{FEE}
4. single direct dependents of any preposition or conjunction node which is in (2) or end-point of a complex path covered in (3)

The last rule ensures that the semantic heads of prepositional phrases and conjunctions are included in the alignment domain.

The alignment range N is defined in a similar way. However, we cannot extract complex paths from the unlabeled graph g, as it does not contain semantic role information. Therefore, we use the same list of complex paths extracted from s. Note that this introduces an unavoidable asymmetry into our similarity computation.

An alignment is a function $\sigma : M \rightarrow N \cup \{\varepsilon\}$ which is injective for all values except ε, i.e., $\sigma(n_1) = \sigma(n_2) \neq \varepsilon \Rightarrow n_1 = n_2$. We score the similarity of two subgraphs expressed by an alignment function σ by the following term:

$$\sum_{\substack{n \in M \\ \sigma(n) \neq \varepsilon}} \text{sem}(n, \sigma(n)) + \alpha \sum_{\substack{(n_1,n_2) \in E(M) \\ (\sigma(n_1),\sigma(n_2)) \in E(N)}} \text{syn}\left(r_{n_2}^{n_1}, r_{\sigma(n_2)}^{\sigma(n_1)}\right) \quad (2)$$

Here, sem represents a semantic similarity measure between graph nodes and syn a syntactic similarity measure between the grammatical role labels of graph edges. $E(M)$ and $E(N)$ are the sets of all graph edges between nodes of M and nodes of N, respectively, and $r_{n_2}^{n_1}$ denotes the grammatical relation between nodes n_1 and n_2.

Equation (2) expresses the similarity between two predicate-argument structures in terms of the sum of semantic similarity scores of aligned graph nodes and the sum of syntactic similarity scores of aligned graph edges. The relative weight of these two sums is determined by the parameter α. Figure 2 shows an example of an alignment between two dependency graphs. Here, the aligned node pairs *thud* and *thump*, *back* and *rest*, *against* and *against*, as well as *wall* and *front* contribute semantic similarity scores, while the three edge pairs SUBJ and SUBJ, IOBJ and IOBJ, as well as DOBJ and DOBJ contribute syntactic similarity scores.

We normalize the resulting score so that it always falls within the interval $[0, 1]$. To take into account unaligned nodes in both the alignment domain and the alignment range, we divide Equation (2) by:

$$\sqrt{|M| \cdot |N|} + \alpha \sqrt{|E(M)| \cdot |E(N)|} \quad (3)$$

A trivial alignment of a seed with itself where all semantic and syntactic scores are 1 will thus receive a score of:

$$\frac{|M| \cdot 1 + \alpha \cdot |E(M)| \cdot 1}{\sqrt{|M|^2} + \alpha \sqrt{E(M)^2}} = 1 \quad (4)$$

which is the largest possible similarity score. The lowest possible score is obviously 0, assuming that the semantic and syntactic scores cannot be negative.

Considerable latitude is available in selecting the semantic and syntactic similarity measures. With regard to semantic similarity, WordNet is a prime contender and indeed has been previously used to acquire new predicates in FrameNet (Pennacchiotti et al., 2008; Burchardt et al., 2005; Johansson and Nugues, 2007). Syntactic similarity may be operationalized in many ways, for example by taking account a hierarchy of grammatical relations (Keenan and Comrie, 1977). Our experiments employed relatively simple instantiations of these measures. We did not make use of WordNet, as we were interested in exploring the setting where WordNet is not available or has limited coverage. Therefore, we approximate the semantic similarity between two nodes via distributional similarity. We present the details of the semantic space model we used in Section 4.

If n and n' are both nouns, verbs or adjectives, we set:

$$\text{sem}(n, n') := \cos(\vec{v}_n, \vec{v}_{n'}) \quad (5)$$

where $\vec{v}_n$ and $\vec{v}_{n'}$ are the vectors representing the lemmas of n and n' respectively. If n and n'

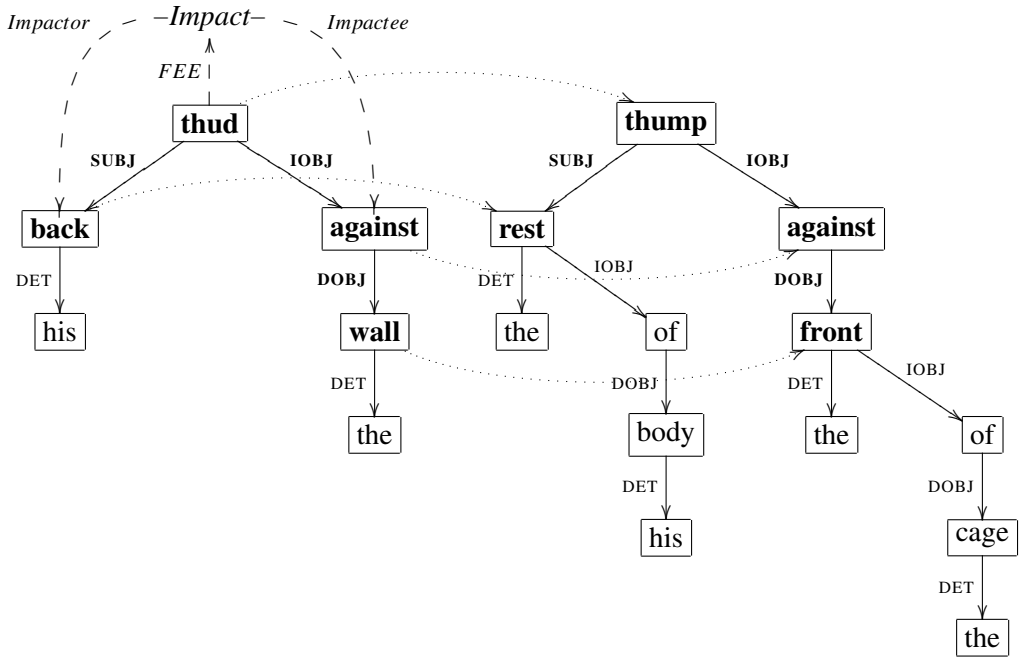

Figure 2: The dotted arrows show aligned nodes in the graphs for the two sentences *His back thudded against the wall.* and *The rest of his body thumped against the front of the cage.* (Graph edges are also aligned to each other.) The alignment domain and alignment range are indicated in bold face. The verb *thud* evokes the frame *Impact*.

are identical prepositions or conjunctions we set $\text{sem}(n,n') := 1$. In all other cases $\text{sem}(n,n') := 0$. As far as syntactic similarity is concerned, we chose the simplest metric possible and set:

$$\text{syn}\left(r,r'\right) := \begin{cases} 1 & \text{if } r = r' \\ 0 & \text{otherwise} \end{cases} \quad (6)$$

3.2 Alignment Search

The problem of finding the best alignment according to the scoring function presented in Equation (2) can be formulated as an integer linear program. Let the binary variables x_{ik} indicate whether node n_i of graph s is aligned to node n_k of graph g. Since it is not only nodes but also graph edges that must be aligned we further introduce binary variables y_{ijkl}, where $y_{ijkl} = 1$ indicates that the edge between nodes n_i and n_j of graph s is aligned to the edge between nodes n_k and n_l of graph g. This follows a general formulation of the graph alignment problem based on maximum structural matching (Klau, 2009). In order for the x_{ik} and y_{ijkl} variables to represent a valid alignment, the following constraints must hold:

1. Each node of s is aligned to at most one node of g: $\sum_k x_{ik} \leq 1$
2. Each node of g is aligned to at most one node of s: $\sum_i x_{ik} \leq 1$
3. Two edges may only be aligned if their adjacent nodes are aligned: $y_{ijkl} \leq x_{ik}$ and $y_{ijkl} \leq x_{jl}$

The scoring function then becomes:

$$\sum_{i,k} \text{sem}(n_i,n_k)x_{ik} + \alpha \cdot \sum_{i,j,k,l} \text{syn}\left(r^{n_i}_{n_j}, r^{n_k}_{n_l}\right) y_{ijkl} \quad (7)$$

We solve this optimization problem with a version of the branch-and-bound algorithm (Land and Doig, 1960). In general, this graph alignment problem is NP-hard (Klau, 2009) and usually solved approximately following a procedure similar to beam search. However, the special structure of constraints 1 to 3, originating from the required injectivity of the alignment function, allows us to solve the optimization exactly. Our implementation of the branch-and-bound algorithm does not generally run in polynomial time, however, we found that in practice we could efficiently compute optimal alignments in almost all cases (less than 0.1% of alignment pairs in our data could not be solved in reasonable time). This relatively benign behavior depends crucially on the fact that we do not have to consider alignments between

full graphs, and the number of nodes in the aligned subgraphs is limited.

4 Experimental Design

In this section we present our experimental set-up for assessing the performance of our method. We give details on the data sets we used, describe the baselines we adopted for comparison with our approach, and explain how our system output was evaluated.

Data Our experiments used annotated sentences from FrameNet as a seed corpus. These were augmented with automatically labeled sentences from the BNC which we used as our expansion corpus. FrameNet sentences were parsed with RASP (Briscoe et al., 2006). In addition to phrase structure trees, RASP delivers a dependency-based representation of the sentence which we used in our experiments. FrameNet role annotations were mapped onto those dependency graph nodes that corresponded most closely to the annotated substring (see Fürstenau (2008) for a detailed description of the mapping algorithm). BNC sentences were also parsed with RASP (Andersen et al., 2008).

We randomly split the FrameNet corpus[1] into 80% training set, 10% test set, and 10% development set. Next, all frame evoking verbs in the training set were ordered by their number of occurrence and split into two groups, *seen* and *unseen*. Every other verb from the ordered list was considered unseen. This quasi-random split covers a broad range of predicates with a varying number of annotations. Accordingly, the FrameNet sentences in the training and test sets were divided into the sets *train_seen*, *train_unseen*, *test_seen*, and *test_unseen*. As we explain below, this was necessary for evaluation purposes.

The *train_seen* dataset consisted of 24,220 sentences, with 1,238 distinct frame evoking verbs, whereas *train_unseen* contained 24,315 sentences with the same number of frame evoking verbs. Analogously, *test_seen* had 2,990 sentences and 817 unique frame evoking verbs; the number of sentences in *test_unseen* was 3,064 (with 847 unique frame evoking verbs).

Model Parameters The alignment model presented in Section 3 crucially relies on the similarity function that scores potential alignments (see Equation (2)). This function has a free parameter, the weight α for determining the relative contribution of semantic and syntactic similarity. We tuned α using leave-one-out cross-validation on the development set. For each annotated sentence in this set we found its most similar other sentence and determined the best alignment between the two dependency graphs representing them. Since the true annotations for each sentence were available, it was possible to evaluate the accuracy of our method for any α value. We did this by comparing the true annotation of a sentence to the annotation its nearest neighbor would have induced by projection. Following this procedure, we obtained best results with $\alpha = 0.2$.

The semantic similarity measure relies on a semantic space model which we built on a lemmatized version of the BNC. Our implementation followed closely the model presented in Fürstenau and Lapata (2009) as it was used in a similar task and obtained good results. Specifically, we used a context window of five words on either side of the target word, and 2,000 vector dimensions. These were the common context words in the BNC. Their values were set to the ratio of the probability of the context word given the target word to the probability of the context word overall. Semantic similarity was measured using the cosine of the angle between the vectors representing any two words. The same semantic space was used to create the distributional profile of a frame (which is the centroid of the vectors of its verbs). For each unknown verb, we consider the k most similar frame candidates (again similarity is measured via cosine). Our experiments explored different values of k ranging from 1 to 10.

Evaluation Our evaluation assessed the performance of a semantic frame and role labeler with and without the annotations produced by our method. The labeler followed closely the implementation described in Johansson and Nugues (2008). We extracted features from dependency parses corresponding to those routinely used in the semantic role labeling literature (see Baker et al. (2007) for an overview). SVM classifiers were trained[2] with the LIBLINEAR library (Fan et al., 2008) and learned to predict the frame name, role spans, and role labels. We followed

[1] Here, we consider only FrameNet example sentences featuring verbal predicates.

[2] The regularization parameter C was set to 0.1.

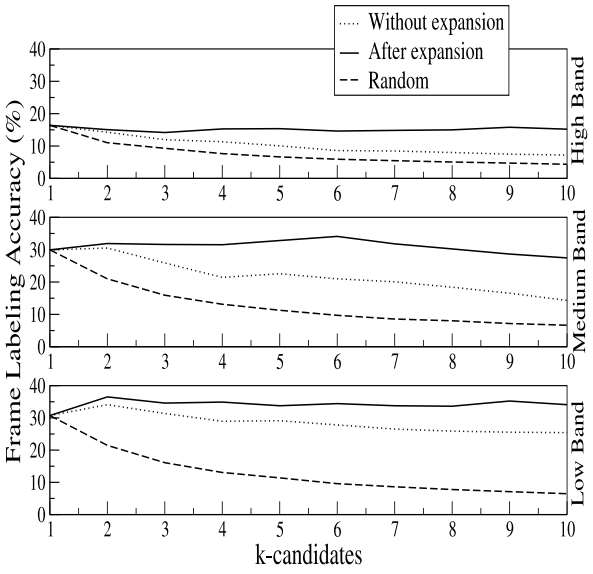

Figure 3: Frame labeling accuracy on high, medium and low frequency verbs, before and after applying our expansion method; the labeler decides among $k = 1, \ldots, 10$ candidate frames.

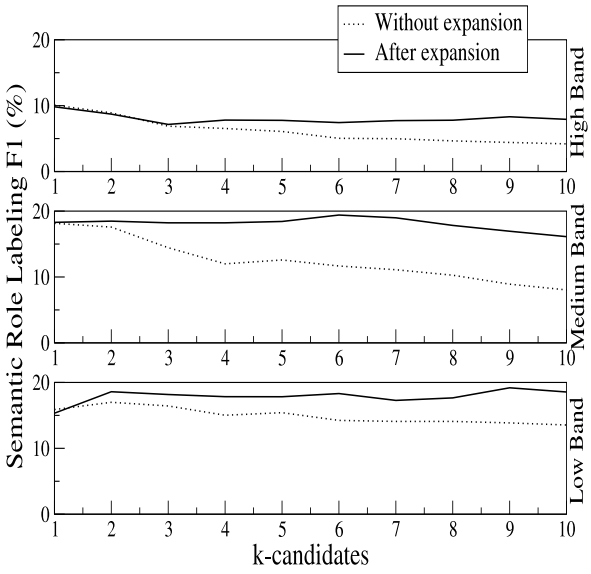

Figure 4: Role labeling F_1 for high, medium, and low frequency verbs (roles of mislabeled frames are counted as wrong); the labeler decides among $k = 1, \ldots, 10$ candidate frames.

the one-versus-one strategy for multi-class classification (Friedman, 1996).

Specifically, the labeler was trained on the *train_seen* data set without any access to training instances representative of the "unknown" verbs in *test_unseen*. We then trained the labeler on a larger set containing *train_seen* and new training examples obtained with our method. To do this, we used *train_seen* as the seed corpus and the BNC as the expansion corpus. For each "unknown" verb in *train_unseen* we obtained BNC sentences with annotations projected from their most similar seeds. The quality of these sentences as training instances varies depending on their similarity to the seed. In our experiments we added to the training set the 20 highest scoring BNC sentences per verb (adding less or more instances led to worse performance).

The average number of frames which can be evoked by a verb token in the set *test_unseen* was 1.96. About half of them (1,522 instances) can evoke only one frame, 22% can evoke two frames, and 14 instances can evoke up to 11 different frames. Finally, there are 120 instances (4%) in *test_unseen* for which the correct frame is not annotated on any sentence in *train_seen*.

5 Results

We first examine how well our method performs at frame labeling. We partitioned the frame evoking verbs in our data set into three bands (High, Medium, and Low) based on an equal division of the range of their occurrence frequency in the BNC. As frequency is strongly correlated with polysemy, the division allows us to assess how well our method is performing at different degrees of ambiguity. Figure 3 summarizes our results for High, Medium, and Low frequency verbs. The number of verbs in each band are 282, 282, and 283, respectively. We compare the frame accuracy of a labeler trained solely on the annotations available in FrameNet (Without expansion) against a labeler that also uses annotations created with our method (After expansion). Both classifiers were employed in a setting where they had to decide among k candidate frames. These were the k most similar frames to the unknown verb in question. We also show the accuracy of a simple baseline labeler, which randomly chooses one of the k candidate frames.

The graphs in Figure 3 show that for verbs in the Medium and Low frequency bands, both classifiers (with and without expansion) outperform the baseline of randomly choosing among k candidate frames. Interestingly, rather than defaulting to the most similar frame ($k = 1$), we observe that ac-

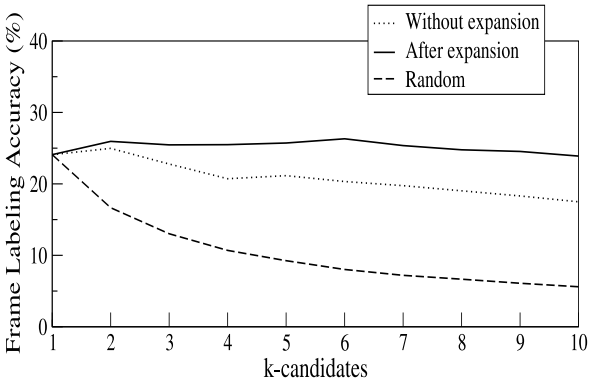

Figure 5: Hybrid frame labeling accuracy ($k = 1$ for High frequency verbs).

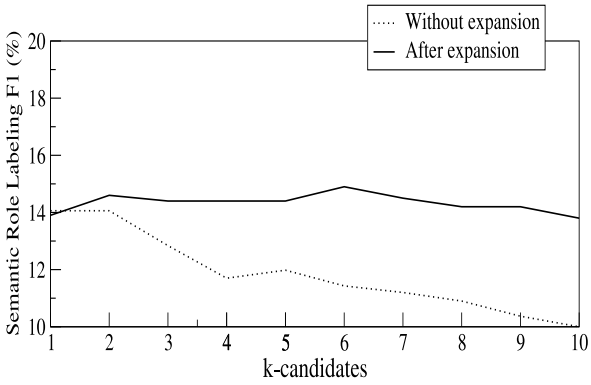

Figure 6: Hybrid role labeling F_1 ($k = 1$ for High frequency verbs).

curacy improves when frame selection is viewed as a classification task. The classifier trained on the expanded training set consistently outperforms the one trained on the original training set. While this is also true for the verbs in the High frequency band, labeling accuracy peaks at $k = 1$ and does not improve when more candidate frames are considered. This is presumably due to the skewed sense distributions of high frequency verbs, and defaulting to the most likely sense achieves relatively good performance.

Next, we evaluated our method on role labeling, again by comparing the performance of our role labeler on the expanded and original training set. Since role and frame labeling are interdependent, we count all predicted roles of an incorrectly predicted frame as wrong. This unavoidably results in low role labeling scores, but allows us to directly compare performance across different settings (e.g., different number of candidate frames, with or without expansion). Figure 4 reports labeled F_1 for verbs in the High, Medium and Low frequency bands. The results are similar to those obtained for frame labeling; the role labeler trained on the the expanded training set consistently outperforms the labeler trained on the unexpanded one. (There is no obvious baseline for role labeling, which is a complex task involving the prediction of frame labels, identification of the role bearing elements, and assignment of role labels.) Again, for High frequency verbs simply defaulting to $k = 1$ performs best.

Taken together, our results on frame and role labeling indicate that our method is not very effective for High frequency verbs (which in practice should be still annotated manually). We therefore also experimented with a hybrid approach that lets the classifier choose among k candidates for Medium and Low frequency verbs and defaults to the most similar candidate for High frequency verbs. Results for this approach are shown in Figures 5 and 6. All differences between the expanded and the unexpanded classifier when choosing between the same $k > 1$ candidates are significant according to McNemar's test ($p < .05$). The best frame labeling accuracy (26.3%) is achieved by the expanded classifier when deciding among $k = 6$ candidate frames. This is significantly better ($p < .01$) than the best performance of the unexpanded classifier (25.0%), which is achieved at $k = 2$. Role labeling results follow a similar pattern. The best expanded classifier (F_1=14.9% at $k = 6$) outperforms the best unexpanded one (F_1=14.1% at $k = 2$). The difference in performance as significant at $p < 0.05$, using stratified shuffling (Noreen, 1989).

6 Conclusions

This paper presents a novel semi-supervised approach for reducing the annotation effort involved in creating resources for semantic role labeling. Our method acquires training instances for unknown verbs (i.e., verbs that are not evoked by existing FrameNet frames) from an unlabeled corpus. A key assumption underlying our work is that verbs with similar meanings will have similar argument structures. Our task then amounts to finding the seen instances that resemble the unseen instances most, and projecting their annotations. We represent this task as a graph alignment problem, and formalize the search for an optimal alignment as an integer linear program under an

objective function that takes semantic and structural similarity into account.

Experimental results show that our method improves frame and role labeling accuracy, especially for Medium and Low frequency verbs. The overall frame labeling accuracy may seem low. There are at least two reasons for this. Firstly, the unknown verb might have a frame for which no manual annotation exists. And secondly, many errors are due to near-misses, i.e., we assign the unknown verb a wrong frame which is nevertheless very similar to the right one. In this case, accuracy will not give us any credit.

An obvious direction for future work concerns improving our scoring function. Pennacchiotti et al. (2008) show that WordNet-based similarity measures outperform their simpler distributional alternatives. An interesting question is whether the incorporation of WordNet-based similarity would lead to similar improvements in our case. Also note that currently our method assigns unknown lexical items to existing frames. A better alternative would be to decide first whether the unknown item can be classified at all (because it evokes a known frame) or whether it represents a genuinely novel frame for which manual annotation must be provided.

Acknowledgments The authors acknowledge the support of DFG (IRTG 715) and EPSRC (grant GR/T04540/01). We are grateful to Richard Johansson for his help with the re-implementation of his semantic role labeler. Special thanks to Manfred Pinkal for valuable feedback on this work.

References

Øistein E. Andersen, Julien Nioche, Ted Briscoe, and John Carroll. 2008. The BNC Parsed with RASP4UIMA. In *Proceedings of the 6th International Language Resources and Evaluation Conference*, pages 865–869, Marrakech, Morocco.

Collin F. Baker, Michael Ellsworth, and Katrin Erk. 2007. SemEval-2007 Task 19: Frame Semantic Structure Extraction. In *Proceedings of the 4th International Workshop on Semantic Evaluations*, pages 99–104, Prague, Czech Republic.

Hans C. Boas. 2005. Semantic frames as interlingual representations for multilingual lexical databases. *International Journal of Lexicography*, 18(4):445–478.

Ted Briscoe, John Carroll, and Rebecca Watson. 2006. The Second Release of the RASP System. In *Proceedings of the COLING/ACL 2006 Interactive Presentation Sessions*, pages 77–80, Sydney, Australia.

Aljoscha Burchardt, Katrin Erk, and Anette Frank. 2005. A WordNet Detour to FrameNet. In *Proceedings of the GLDV 200 Workshop GermaNet II*, Bonn, Germany.

Rong-En Fan, Kai-Wei Chang, Cho-Jui Hsieh, Xiang-Rui Wang, and Chih-Jen Lin. 2008. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 9:1871–1874.

Christiane Fellbaum, editor. 1998. *WordNet: An Electronic Database*. MIT Press, Cambridge, MA.

Charles J. Fillmore, Christopher R. Johnson, and Miriam R. L. Petruck. 2003. Background to FrameNet. *International Journal of Lexicography*, 16:235–250.

Jerome H. Friedman. 1996. Another approach to polychotomous classification. Technical report, Department of Statistics, Stanford University.

Hagen Fürstenau and Mirella Lapata. 2009. Semi-supervised semantic role labeling. In *Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics*, pages 220–228, Athens, Greece.

Hagen Fürstenau. 2008. Enriching frame semantic resources with dependency graphs. In *Proceedings of the 6th Language Resources and Evaluation Conference*, pages 1478–1484, Marrakech, Morocco.

Richard Johansson and Pierre Nugues. 2006. A FrameNet-based semantic role labeler for Swedish. In *Proceedings of the COLING/ACL 2006 Main Conference Poster Sessions*, pages 436–443, Sydney, Australia.

Richard Johansson and Pierre Nugues. 2007. Using WordNet to extend FrameNet coverage. In Richard Johansson and Pierre Nugues, editors, *FRAME 2007: Building Frame Semantics Resources for Scandinavian and Baltic Languages*, pages 27–30, Tartu, Estonia.

Richard Johansson and Pierre Nugues. 2008. The effect of syntactic representation on semantic role labeling. In *Proceedings of the 22nd International Conference on Computational Linguistics*, pages 393–400, Manchester, UK.

E. Keenan and B. Comrie. 1977. Noun phrase accessibility and universal grammar. *Linguistic Inquiry*, 8:62–100.

Gunnar W. Klau. 2009. A new graph-based method for pairwise global network alignment. *BMC Bioinformatics*, 10 (Suppl 1).

A.H. Land and A.G. Doig. 1960. An automatic method for solving discrete programming problems. *Econometrica*, 28:497–520.

Gabor Melli, Yang Wang, Yurdong Liu, Mehdi M. Kashani, Zhongmin Shi, Baohua Gu, Anoop Sarkar, and Fred Popowich. 2005. Description of SQUASH, the SFU question answering summary handler for the duc-2005 summarization task. In *Proceedings of the HLT/EMNLP Document Understanding Workshop*, Vancouver, Canada.

Srini Narayanan and Sanda Harabagiu. 2004. Question answering based on semantic structures. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 693–701, Geneva, Switzerland.

E. Noreen. 1989. *Computer-intensive Methods for Testing Hypotheses: An Introduction*. John Wiley and Sons Inc.

Sebastian Padó and Mirella Lapata. 2006. Optimal constituent alignment with edge covers for semantic projection. In *Proceedings of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the Association for Computational Linguistics*, pages 1161–1168, Sydney, Australia.

Marco Pennacchiotti, Diego De Cao, Roberto Basili, Danilo Croce, and Michael Roth. 2008. Automatic induction of FrameNet lexical units. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 457–465, Honolulu, Hawaii.

Mihai Surdeanu, Sanda Harabagiu, John Williams, and Paul Aarseth. 2003. Using predicate-argument structures for information extraction. In *Proceedings of the 41st Annual Meeting of the Association for Computational Linguistics*, pages 8–15, Sapporo, Japan.