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

Large scale annotated corpora are prerequisite to developing high-performance semantic role labeling systems. Unfortunately, such corpora are expensive to produce, limited in size, and may not be representative. Our work aims to reduce the annotation effort involved in creating resources for semantic role labeling via semi-supervised learning. Our algorithm augments a small number of manually labeled instances with unlabeled examples whose roles are inferred automatically via annotation projection. We formulate the projection task as a generalization of the linear assignment problem. We seek to find a role assignment in the unlabeled data such that the argument similarity between the labeled and unlabeled instances is maximized. Experimental results on semantic role labeling show that the automatic annotations produced by our method improve performance over using hand-labeled instances alone.

1 Introduction

Recent years have seen a growing interest in the task of automatically identifying and labeling the semantic roles conveyed by sentential constituents (Gildea and Jurafsky, 2002). This is partly due to its relevance for applications ranging from information extraction (Surdeanu et al., 2003; Moschitti et al., 2003) to question answering (Shen and Lapata, 2007), paraphrase identification (Padó and Erk, 2005), and the modeling of textual entailment relations (Tatu and Moldovan, 2005). Resources like FrameNet (Fillmore et al., 2003) and PropBank (Palmer et al., 2005) have also facilitated the development of semantic role labeling methods by providing high-quality annotations for use in training. Semantic role labelers are commonly developed using a supervised learning paradigm[1] where a classifier learns to predict role labels based on features extracted from annotated training data.

Examples of the annotations provided in FrameNet are given in (1). Here, the meaning of predicates (usually verbs, nouns, or adjectives) is conveyed by *frames*, schematic representations of situations. Semantic roles (or *frame elements*) are defined for each frame and correspond to salient entities present in the situation evoked by the predicate (or *frame evoking element*). Predicates with similar semantics instantiate the same frame and are attested with the same roles. In our example, the frame *Cause_harm* has three core semantic roles, *Agent*, *Victim*, and *Body_part* and can be instantiated with verbs such as *punch*, *crush*, *slap*, and *injure*. The frame may also be attested with non-core (peripheral) roles that are more generic and often shared across frames (see the roles *Degree*, *Reason*, and *Means*, in (1c) and (1d)).

(1)
a. [Lee]$_{Agent}$ punched [John]$_{Victim}$ [in the eye]$_{Body\_part}$.
b. [A falling rock]$_{Cause}$ crushed [my ankle]$_{Body\_part}$.
c. [She]$_{Agent}$ slapped [him]$_{Victim}$ [hard]$_{Degree}$ [for his change of mood]$_{Reason}$.
d. [Rachel]$_{Agent}$ injured [her friend]$_{Victim}$ [by closing the car door on his left hand]$_{Means}$.

The English FrameNet (version 1.3) contains 502 frames covering 5,866 lexical entries. It also comes with a set of manually annotated example sentences, taken mostly from the British National Corpus. These annotations are often used

[1] The approaches are too numerous to list; 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.

as training data for semantic role labeling systems. However, the applicability of these systems is limited to those words for which labeled data exists, and their accuracy is strongly correlated with the amount of labeled data available. Despite the substantial annotation effort involved in the creation of FrameNet (spanning approximately twelve years), the number of annotated instances varies greatly across lexical items. For instance, FrameNet contains annotations for 2,113 verbs; of these 12.3% have five or less annotated examples. The average number of annotations per verb is 29.2. Labeled data is thus scarce for individual predicates within FrameNet's target domain and would presumably be even scarcer across domains. The problem is more severe for languages other than English, where training data on the scale of FrameNet is virtually non-existent. Although FrameNets are being constructed for German, Spanish, and Japanese, these resources are substantially smaller than their English counterpart and of limited value for modeling purposes.

One simple solution, albeit expensive and time-consuming, is to manually create more annotations. A better alternative may be to begin with an initial small set of labeled examples and augment it with unlabeled data sufficiently similar to the original labeled set. Suppose we have manual annotations for sentence (1a). We shall try and find in an unlabeled corpus other sentences that are both structurally and semantically similar. For instance, we may think that *Bill will punch me in the face* and *I punched her hard in the head* resemble our initial sentence and are thus good examples to add to our database. Now, in order to use these new sentences as training data we must somehow infer their semantic roles. We can probably guess that constituents in the same syntactic position must have the same semantic role, especially if they refer to the same concept (e.g., "body parts") and thus label *in the face* and *in the head* with the role *Body_part*. Analogously, *Bill* and *I* would be labeled as *Agent* and *me* and *her* as *Victim*.

In this paper we formalize the method sketched above in order to expand a small number of FrameNet-style semantic role annotations with large amounts of unlabeled data. We adopt a learning strategy where annotations are projected from labeled onto unlabeled instances via maximizing a similarity function measuring syntactic and semantic compatibility. We formalize the annotation projection problem as a generalization of the linear assignment problem and solve it efficiently using the simplex algorithm. We evaluate our algorithm by comparing the performance of a semantic role labeler trained on the annotations produced by our method and on a smaller dataset consisting solely of hand-labeled instances. Results in several experimental settings show that the automatic annotations, despite being noisy, bring significant performance improvements.

2 Related Work

The lack of annotated data presents an obstacle to developing many natural language applications, especially when these are not in English. It is therefore not surprising that previous efforts to reduce the need for semantic role annotation have focused primarily on non-English languages.

Annotation projection is a popular framework for transferring frame semantic annotations from one language to another by exploiting the translational and structural equivalences present in parallel corpora. The idea here is to leverage the existing English FrameNet and rely on word or constituent alignments to automatically create an annotated corpus in a new language. Padó and Lapata (2006) transfer semantic role annotations from English onto German and Johansson and Nugues (2006) from English onto Swedish. A different strategy is presented in Fung and Chen (2004), where English FrameNet entries are mapped to concepts listed in HowNet, an on-line ontology for Chinese, without consulting a parallel corpus. Then, Chinese sentences with predicates instantiating these concepts are found in a monolingual corpus and their arguments are labeled with FrameNet roles.

Other work attempts to alleviate the data requirements for semantic role labeling either by relying on unsupervised learning or by extending existing resources through the use of unlabeled data. Swier and Stevenson (2004) present an unsupervised method for labeling the arguments of verbs with their semantic roles. Given a verb instance, their method first selects a frame from VerbNet, a semantic role resource akin to FrameNet and PropBank, and labels each argument slot with sets of possible roles. The algorithm proceeds iteratively by first making initial unambiguous role assignments, and then successively updating a probabil-

ity model on which future assignments are based. Being unsupervised, their approach requires no manual effort other than creating the frame dictionary. Unfortunately, existing resources do not have exhaustive coverage and a large number of verbs may be assigned no semantic role information since they are not in the dictionary in the first place. Pennacchiotti et al. (2008) address precisely this problem by augmenting FrameNet with new lexical units if they are similar to an existing frame (their notion of similarity combines distributional and WordNet-based measures). In a similar vein, Gordon and Swanson (2007) attempt to increase the coverage of PropBank. Their approach leverages existing annotations to handle novel verbs. Rather than annotating new sentences that contain novel verbs, they find syntactically similar verbs and use their annotations as surrogate training data.

Our own work aims to reduce but not entirely eliminate the annotation effort involved in creating training data for semantic role labeling. We thus assume that a small number of manual annotations is initially available. Our algorithm augments these with unlabeled examples whose roles are inferred automatically. We apply our method in a monolingual setting, and thus do not project annotations between languages but within the same language. In contrast to Pennacchiotti et al. (2008) and Gordon and Swanson (2007), we do not aim to handle novel verbs, although this would be a natural extension of our method. Given a verb and a few labeled instances exemplifying its roles, we wish to find more instances of the same verb in an unlabeled corpus so as to improve the performance of a hypothetical semantic role labeler without having to annotate more data manually. Although the use of semi-supervised learning is widespread in many natural language tasks, ranging from parsing to word sense disambiguation, its application to FrameNet-style semantic role labeling is, to our knowledge, novel.

3 Semi-Supervised Learning Method

Our method assumes that we have access to a small *seed* corpus that has been manually annotated. This represents a relatively typical situation where some annotation has taken place but not on a scale that is sufficient for high-performance supervised learning. For each sentence in the seed corpus we select a number of similar sentences

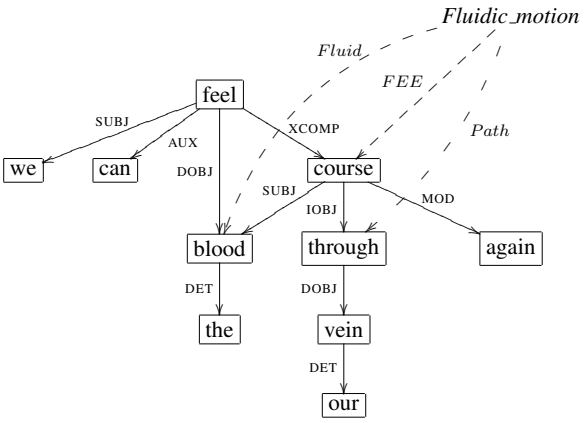

Figure 1: Labeled dependency graph with semantic role annotations for the frame evoking element (FEE) *course* in the sentence *We can feel the blood coursing through our veins again.* The frame is *Fluidic_motion*, and its roles are *Fluid* and *Path*. Directed edges (without dashes) represent dependency relations between words, edge labels denote types of grammatical relations (e.g., SUBJ, AUX).

from an unlabeled *expansion* corpus. These are automatically annotated by projecting relevant semantic role information from the labeled sentence. The similarity between two sentences is operationalized by measuring whether their arguments have a similar structure and whether they express related meanings. The seed corpus is then enlarged with the k most similar unlabeled sentences to form the expanded corpus. In what follows we describe in more detail how we measure similarity and project annotations.

3.1 Extracting Predicate-Argument Structures

Our method operates over labeled dependency graphs. We show an example in Figure 1 for the sentence *We can feel the blood coursing through our veins again.* We represent verbs (i.e., frame evoking elements) in the seed and unlabeled corpora by their predicate-argument structure. Specifically, we record the direct dependents of the predicate *course* (e.g., *blood* or *again* in Figure 1) and their grammatical roles (e.g., SUBJ, MOD). Prepositional nodes are collapsed, i.e., we record the preposition's object and a composite grammatical role (like IOBJ_THROUGH, where IOBJ stands for "prepositional object" and THROUGH for the preposition itself). In addition to direct dependents, we also

Lemma	GramRole	SemRole
blood	SUBJ	*Fluid*
vein	IOBJ_THROUGH	*Path*
again	MOD	—

Table 1: Predicate-argument structure for the verb *course* in Figure 1.

consider nodes coordinated with the predicate as arguments. Finally, for each argument node we record the semantic roles it carries, if any. All surface word forms are lemmatized. An example of the argument structure information we obtain for the predicate *course* (see Figure 1) is shown in Table 1.

We obtain information about grammatical roles from the output of RASP (Briscoe et al., 2006), a broad-coverage dependency parser. However, there is nothing inherent in our method that restricts us to this particular parser. Any other parser with broadly similar dependency output could serve our purposes.

3.2 Measuring Similarity

For each frame evoking verb in the seed corpus our method creates a labeled predicate-argument representation. It also extracts all sentences from the unlabeled corpus containing the same verb. Not all of these sentences will be suitable instances for adding to our training data. For example, the same verb may evoke a different frame with different roles and argument structure. We therefore must select sentences which resemble the seed annotations. Our hypothesis is that verbs appearing in similar syntactic and semantic contexts will behave similarly in the way they relate to their arguments.

Estimating the similarity between two predicate argument structures amounts to finding the highest-scoring alignment between them. More formally, given a labeled predicate-argument structure p^l with m arguments and an unlabeled predicate-argument structure p^u with n arguments, we consider (and score) all possible alignments between these arguments. A (partial) alignment can be viewed as an injective function $\sigma : M_\sigma \rightarrow \{1, \ldots, n\}$ where $M_\sigma \subset \{1, \ldots, m\}$. In other words, an argument i of p^l is aligned to argument $\sigma(i)$ of p^u if $i \in M_\sigma$. Note that this allows for unaligned arguments on both sides.

We score each alignment σ using a similarity function $\text{sim}(\sigma)$ defined as:

$$\sum_{i \in M_\sigma} \Big(A \cdot \text{syn}(g_i^l, g_{\sigma(i)}^u) + \text{sem}(w_i^l, w_{\sigma(i)}^u) - B \Big)$$

where $\text{syn}(g_i^l, g_{\sigma(i)}^u)$ denotes the syntactic similarity between grammatical roles g_i^l and $g_{\sigma(i)}^u$ and $\text{sem}(w_i^l, w_{\sigma(i)}^u)$ the semantic similarity between head words w_i^l and $w_{\sigma(i)}^u$.

Our goal is to find an alignment such that the similarity function is maximized: $\sigma^* := \arg\max_\sigma \text{sim}(\sigma)$. This optimization problem is a generalized version of the linear assignment problem (Dantzig, 1963). It can be straightforwardly expressed as a linear programming problem by associating each alignment σ with a set of binary indicator variables x_{ij}:

$$x_{ij} := \begin{cases} 1 & \text{if } i \in M_\sigma \wedge \sigma(i) = j \\ 0 & \text{otherwise} \end{cases}$$

The similarity objective function then becomes:

$$\sum_{i=1}^{m} \sum_{j=1}^{n} \Big(A \cdot \text{syn}(g_i^l, g_j^u) \; + \; \text{sem}(w_i^l, w_j^u) - B \Big) \, x_{ij}$$

subject to the following constraints ensuring that σ is an injective function on some M_σ:

$$\sum_{j=1}^{n} x_{ij} \leq 1 \quad \text{for all } i = 1, \ldots, m$$

$$\sum_{i=1}^{m} x_{ij} \leq 1 \quad \text{for all } j = 1, \ldots, n$$

Figure 2 graphically illustrates the alignment projection problem. Here, we wish to project semantic role information from the seed *blood coursing through our veins again* onto the unlabeled sentence *Adrenalin was still coursing through her veins*. The predicate *course* has three arguments in the labeled sentence and four in the unlabeled sentence (represented as rectangles in the figure). There are 73 possible alignments in this example. In general, for any m and n arguments, where $m \leq n$, the number of alignments is $\sum_{k=0}^{m} \frac{m!n!}{(m-k)!(n-k)!k!}$. Each alignment is scored by taking the sum of the similarity scores of the individual alignment pairs (e.g., between *blood* and *be*, *vein* and *still*). In this example, the highest scoring alignment is between *blood* and *adrenalin*, *vein* and *vein*, and *again* and *still*, whereas *be* is

left unaligned (see the non-dotted edges in Figure 2). Note that only *vein* and *blood* carry semantic roles (i.e., *Fluid* and *Path*) which are projected onto *adrenalin* and *vein*, respectively.

Finding the best alignment crucially depends on estimating the syntactic and semantic similarity between arguments. We define the syntactic measure on the grammatical relations produced by RASP. Specifically, we set $\text{syn}(g_i^l, g_{\sigma(i)}^u)$ to 1 if the relations are identical, to $a \leq 1$ if the relations are of the same type but different subtype[2] and to 0 otherwise. To avoid systematic errors, syntactic similarity is also set to 0 if the predicates differ in voice. We measure the semantic similarity $\text{sem}(w_i^l, w_{\sigma(i)}^u)$ with a semantic space model. The meaning of each word is represented by a vector of its co-occurrences with neighboring words. The cosine of the angle of the vectors representing w^l and w^u quantifies their similarity (Section 4 describes the specific model we used in our experiments in more detail).

The parameter A counterbalances the importance of syntactic and semantic information, while the parameter B can be interpreted as the lowest similarity value for which an alignment between two arguments is possible. An optimal alignment σ^* cannot link arguments i_0 of p^l and j_0 of p^u, if $A \cdot \text{syn}(g_{i_0}^l, g_{j_0}^u) + \text{sem}(w_{i_0}^l, w_{j_0}^u) < B$ (i.e., either $i_0 \notin M_{\sigma^*}$ or $\sigma^*(i_0) \neq j_0$). This is because for an alignment σ with $\sigma(i_0) = j_0$ we can construct a better alignment σ_0, which is identical to σ on all $i \neq i_0$, but leaves i_0 unaligned (i.e., $i_0 \notin M_{\sigma_0}$). By eliminating a negative term from the scoring function, it follows that $\text{sim}(\sigma_0) > \text{sim}(\sigma)$. Therefore, an alignment σ satisfying $\sigma(i_0) = j_0$ cannot be optimal and conversely the optimal alignment σ^* can never link two arguments with each other if the sum of their weighted syntactic and semantic similarity scores is below B.

3.3 Projecting Annotations

Once we obtain the best alignment σ^* between p^l and p^u, we can simply transfer the role of each role-bearing argument i of p^l to the aligned argument $\sigma^*(i)$ of p^u, resulting in a labeling of p^u.

To increase the accuracy of our method we discard projections if they fail to transfer all roles of the labeled to the unlabeled dependency graph.

[2]This concerns fine-grained distinctions made by the parser, e.g., the underlying grammatical roles in passive constructions.

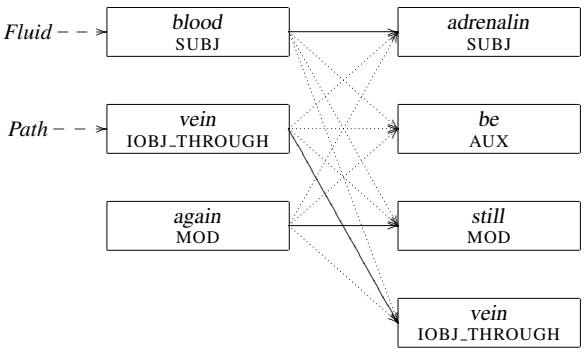

Figure 2: Alignments between the argument structures representing the clauses *blood coursing through our veins again* and *Adrenalin was still coursing through her veins*; non-dotted lines illustrate the highest scoring alignment.

This can either be the case if p^l does not cover all roles annotated on the graph (i.e., there are role-bearing nodes which we do not recognize as arguments of the frame evoking verb) or if there are unaligned role-bearing arguments (i.e., $i \notin M_{\sigma^*}$ for a role-bearing argument i of p^l).

The remaining projections form our expansion corpus. For each seed instance we select the k most similar neighbors to add to our training data. The parameter k controls the trade-off between annotation confidence and expansion size.

4 Experimental Setup

In this section we discuss our experimental setup for assessing the usefulness of the method presented above. We give details on our training procedure and parameter estimation, describe the semantic labeler we used in our experiments and explain how its output was evaluated.

Corpora Our seed corpus was taken from FrameNet. The latter contains approximately 2,000 verb entries out of which we randomly selected a sample of 100. We next extracted all annotated sentences for each of these verbs. These sentences formed our gold standard corpus, 20% of which was reserved as test data. We used the remaining 80% as seeds for training purposes. We generated seed corpora of various sizes by randomly reducing the number of annotation instances per verb to a maximum of n. An additional (non-overlapping) random sample of 100 verbs was used as development set for tuning the parameters for our method. We gathered unlabeled sentences from the BNC.

The seed and unlabeled corpora were parsed with RASP (Briscoe et al., 2006). The FrameNet annotations in the seed corpus were converted into dependency graphs (see Figure 1) using the method described in Fürstenau (2008). Briefly, the method works by matching nodes in the dependency graph with role bearing substrings in FrameNet. It first finds the node in the graph which most closely matches the frame evoking element in FrameNet. Next, individual graph nodes are compared against labeled substrings in FrameNet to transfer all roles onto their closest matching graph nodes.

Parameter Estimation The similarity function described in Section 3.2 has three free parameters. These are the weight A which determines the relative importance of syntactic and semantic information, the parameter B which determines when two arguments cannot be aligned and the syntactic score a for almost identical grammatical roles. We optimized these parameters on the development set using Powell's direction set method (Brent, 1973) with F_1 as our loss function. The optimal values for A, B and a were 1.76, 0.41 and 0.67, respectively.

Our similarity function is further parametrized in using a semantic space model to compute the similarity between two words. Considerable latitude is allowed in specifying the parameters of vector-based models. These involve the definition of the linguistic context over which cooccurrences are collected, the number of components used (e.g., the k most frequent words in a corpus), and their values (e.g., as raw cooccurrence frequencies or ratios of probabilities).

We created a vector-based model from a lemmatized version of the BNC. Following previous work (Bullinaria and Levy, 2007), we optimized the parameters of our model on a word-based semantic similarity task. The task involves examining the degree of linear relationship between the human judgments for two individual words and vector-based similarity values. We experimented with a variety of dimensions (ranging from 50 to 500,000), vector component definitions (e.g., pointwise mutual information or log likelihood ratio) and similarity measures (e.g., cosine or confusion probability). We used WordSim353, a benchmark dataset (Finkelstein et al., 2002), consisting of relatedness judgments (on a scale of 0 to 10) for 353 word pairs.

We obtained best results with a model using a context window of five words on either side of the target word, the cosine measure, and 2,000 vector dimensions. The latter were the most common context words (excluding a stop list of function words). 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. This configuration gave high correlations with the WordSim353 similarity judgments using the cosine measure.

Solving the Linear Program A variety of algorithms have been developed for solving the linear assignment problem efficiently. In our study, we used the simplex algorithm (Dantzig, 1963). We generate and solve an LP of every unlabeled sentence we wish to annotate.

Semantic role labeler We evaluated our method on a semantic role labeling task. Specifically, we compared the performance of a generic semantic role labeler trained on the seed corpus and a larger corpus expanded with annotations produced by our method. Our semantic role labeler followed closely the implementation of 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 to identify the arguments and label them with appropriate roles. For the latter we performed multi-class classification following the one-versus-one method[3] (Friedman, 1996). For the experiments reported in this paper we used the LIBLINEAR library (Fan et al., 2008). The misclassification penalty C was set to 0.1.

To evaluate against the test set, we linearized the resulting dependency graphs in order to obtain labeled role bracketings like those in example (1) and measured labeled precision, labeled recall and labeled F_1. (Since our focus is on role labeling and not frame prediction, we let our role labeler make use of gold standard frame annotations, i.e., labeling of frame evoking elements with frame names.)

5 Results

The evaluation of our method was motivated by three questions: (1) How do different training set sizes affect semantic role labeling performance?

[3]Given n classes the one-versus-one method builds $n(n-1)/2$ classifiers.

TrainSet	Size	Prec (%)	Rec (%)	F_1 (%)
0-NN	849	35.5	42.0	38.5
1-NN	1205	36.4	43.3	39.5
2-NN	1549	38.1	44.1	40.9*
3-NN	1883	37.9	43.7	40.6*
4-NN	2204	38.0	43.9	40.7*
5-NN	2514	37.4	43.9	40.4*
self train	1609	34.0	41.0	37.1

Table 2: Semantic role labeling performance using different amounts of training data; the seeds are expanded with their k nearest neighbors; *: F_1 is significantly different from 0-NN ($p < 0.05$).

Training size varies depending on the number of unlabeled sentences added to the seed corpus. The quality of these sentences also varies depending on their similarity to the seed sentences. So, we would like to assess whether there is a trade-off between annotation quality and training size. (2) How does the size of the seed corpus influence role labeling performance? Here, we are interested to find out what is the least amount of manual annotation possible for our method to have some positive impact. (3) And finally, what are the annotation savings our method brings?

Table 2 shows the performance of our semantic role labeler when trained on corpora of different sizes. The seed corpus was reduced to at most 10 instances per verb. Each row in the table corresponds to adding the k nearest neighbors of these instances to the training data. When trained solely on the seed corpus the semantic role labeler yields a (labeled) F_1 of 38.5%, (labeled) recall is 42.0% and (labeled) precision is 35.5% (see row 0-NN in the table). All subsequent expansions yield improved precision and recall. In all cases except $k = 1$ the improvement is statistically significant ($p < 0.05$). We performed significance testing on F_1 using stratified shuffling (Noreen, 1989), an instance of assumption-free approximative randomization testing. As can be seen, the optimal trade-off between the size of the training corpus and annotation quality is reached with two nearest neighbors. This corresponds roughly to doubling the number of training instances. (Due to the restrictions mentioned in Section 3.3 a 2-NN expansion does not triple the number of instances.)

We also compared our results against a self-training procedure (see last row in Table 2). Here, we randomly selected unlabeled sentences corresponding in number to a 2-NN expansion, labeled them with our role labeler, added them to the training set, and retrained. Self-training resulted in performance inferior to the baseline of adding no unlabeled data at all (see the first row in Table 2). Performance decreased even more with the addition of more self-labeled instances. These results indicate that the similarity function is crucial to the success of our method.

An example of the annotations our method produces is given below. Sentence (2a) is the seed. Sentences (2b)–(2e) are its most similar neighbors. The sentences are presented in decreasing order of similarity.

(2)
a. [He]$_{Theme}$ stared and came [slowly]$_{Manner}$ [towards me]$_{Goal}$.
b. [He]$_{Theme}$ had heard the shooting and come [rapidly]$_{Manner}$ [back towards the house]$_{Goal}$.
c. Without answering, [she]$_{Theme}$ left the room and came [slowly]$_{Manner}$ [down the stairs]$_{Goal}$.
d. [Then]$_{Manner}$ [he]$_{Theme}$ won't come [to Salisbury]$_{Goal}$.
e. Does [he]$_{Theme}$ always come round [in the morning]$_{Goal}$ [then]$_{Manner}$?

As we can see, sentences (2b) and (2c) accurately identify the semantic roles of the verb *come* evoking the frame *Arriving*. In (2b) *He* is labeled as *Theme*, *rapidly* as *Manner*, and *towards the house* as *Goal*. Analogously, in (2c) *she* is the *Theme*, *slowly* is *Manner* and *down the stairs* is *Goal*. The quality of the annotations decreases with less similar instances. In (2d) *then* is marked erroneously as *Manner*, whereas in (2e) only the *Theme* role is identified correctly.

To answer our second question, we varied the size of the training corpus by varying the number of seeds per verb. For these experiments we fixed k = 2. Table 3 shows the performance of the semantic role labeler when the seed corpus has one annotation per verb, five annotations per verb, and so on. (The results for 10 annotations are repeated from Table 2). With 1, 5 or 10 instances per verb our method significantly improves labeling performance. We observe improvements in F_1 of 1.5%, 2.1%, and 2.4% respectively when adding the 2 most similar neighbors to these training corpora. Our method also improves F_1 when a 20 seeds

TrainSet	Size	Prec (%)	Rec (%)	F_1 (%)
$\leq$ 1 seed	95	24.9	31.3	27.7
+ 2-NN	170	26.4	32.6	29.2*
$\leq$ 5 seeds	450	29.7	38.4	33.5
+ 2-NN	844	31.8	40.4	35.6*
$\leq$ 10 seeds	849	35.5	42.0	38.5
+ 2-NN	1549	38.1	44.1	40.9*
$\leq$ 20 seeds	1414	38.7	46.1	42.1
+ 2-NN	2600	40.5	46.7	43.4
all seeds	2323	38.3	47.0	42.2
+ 2-NN	4387	39.5	46.7	42.8

Table 3: Semantic role labeling performance using different numbers of seed instances per verb in the training corpus; the seeds are expanded with their $k = 2$ nearest neighbors; *: F_1 is significantly different from seed corpus ($p < 0.05$).

corpus or all available seeds are used, however the difference is not statistically significant.

The results in Table 3 also allow us to draw some conclusions regarding the relative quality of manual and automatic annotation. Expanding a seed corpus with 10 instances per verb improves F_1 from 38.5% to 40.9%. We can compare this to the labeler's performance when trained solely on the 20 seeds corpus (without any expansion). The latter has approximately the same size as the expanded 10 seeds corpus. Interestingly, F_1 on this exclusively hand-annotated corpus is only 1.2% better than on the expanded corpus. So, using our expansion method on a 10 seeds corpus performs almost as well as using twice as many manual annotations. Even in the case of the 5 seeds corpus, where there is limited information for our method to expand from, we achieve an improvement from 33.5% to 35.6%, compared to 38.5% for manual annotation of about the same number of instances. In sum, while additional manual annotation is naturally more effective for improving the quality of the training data, we can achieve substantial proportions of these improvements by automatic expansion alone. This is a promising result suggesting that it is possible to reduce annotation costs without drastically sacrificing quality.

6 Conclusions

This paper presents a novel method for reducing the annotation effort involved in creating resources for semantic role labeling. Our strategy is to expand a manually annotated corpus by projecting semantic role information from labeled onto unlabeled instances. We formulate the projection problem as an instance of the linear assignment problem. We seek to find role assignments that maximize the similarity between labeled and unlabeled instances. Similarity is measured in terms of structural and semantic compatibility between argument structures.

Our method improves semantic role labeling performance in several experimental conditions. It is especially effective when a small number of annotations is available for each verb. This is typically the case when creating frame semantic corpora for new languages or new domains. Our experiments show that expanding such corpora with our method can yield almost the same relative improvement as using exclusively manual annotation.

In the future we plan to extend our method in order to handle novel verbs that are not attested in the seed corpus. Another direction concerns the systematic modeling of diathesis alternations (Levin, 1993). These are currently only captured implicitly by our method (when the semantic similarity overrides syntactic dissimilarity). Ideally, we would like to be able to systematically identify changes in the realization of the argument structure of a given predicate. Although our study focused solely on FrameNet annotations, we believe it can be adapted to related annotation schemes, such as PropBank. An interesting question is whether the improvements obtained by our method carry over to other role labeling frameworks.

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 reimplementation of his semantic role labeler.

References

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.

R. P. Brent. 1973. *Algorithms for Minimization without Derivatives*. Prentice-Hall, Englewood Cliffs, NJ.

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.

J. A. Bullinaria and J. P. Levy. 2007. Extracting semantic representations from word co-occurrence statistics: A computational study. *Behavior Research Methods*, 39:510–526.

George B. Dantzig. 1963. *Linear Programming and Extensions*. Princeton University Press, Princeton, NJ, USA.

R.-E. Fan, K.-W. Chang, C.-J. Hsieh, X.-R. Wang, and C.-J. Lin. 2008. LIBLINEAR: A library for large linear classification. *Journal of Machine Learning Research*, 9:1871–1874.

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

Lev Finkelstein, Evgeniy Gabrilovich, Yossi Matias, Ehud Rivlin, Zach Solan, Gadi Wolfman, and Eytan Ruppin. 2002. Placing search in context: The concept revisited. *ACM Transactions on Information Systems*, 20(1):116–131.

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

Pascale Fung and Benfeng Chen. 2004. BiFrameNet: Bilingual frame semantics resources construction by cross-lingual induction. In *Proceedings of the 20th International Conference on Computational Linguistics*, pages 931–935, Geneva, Switzerland.

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

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

Andrew Gordon and Reid Swanson. 2007. Generalizing semantic role annotations across syntactically similar verbs. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 192–199, Prague, Czech Republic.

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. 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.

Beth Levin. 1993. *English Verb Classes and Alternations: A Preliminary Investigation*. University of Chicago Press.

Alessandro Moschitti, Paul Morarescu, and Sanda Harabagiu. 2003. Open-domain information extraction via automatic semantic labeling. In *Proceedings of FLAIRS 2003*, pages 397–401, St. Augustine, FL.

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

Sebastian Padó and Katrin Erk. 2005. To cause or not to cause: Cross-lingual semantic matching for paraphrase modelling. In *Proceedings of the EUROLAN Workshop on Cross-Linguistic Knowledge Induction*, pages 23–30, Cluj-Napoca, Romania.

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.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The Proposition Bank: An annotated corpus of semantic roles. *Computational Linguistics*, 31(1):71–106.

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.

Dan Shen and Mirella Lapata. 2007. Using semantic roles to improve question answering. In *Proceedings of the joint Conference on Empirical Methods in Natural Language Processing and Conference on Computational Natural Language Learning*, pages 12–21, Prague, Czech Republic.

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.

Robert S. Swier and Suzanne Stevenson. 2004. Unsupervised semantic role labelling. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 95–102. Bacelona, Spain.

Marta Tatu and Dan Moldovan. 2005. A semantic approach to recognizing textual entailment. In *Proceedings of the joint Human Language Technology Conference and Conference on Empirical Methods in Natural Language Processing*, pages 371–378, Vancouver, BC.