Towards Open-Domain Semantic Role Labeling

Danilo Croce, Cristina Giannone, Paolo Annesi, Roberto Basili
{croce,giannone,annesi,basili}@info.uniroma2.it
Department of Computer Science, Systems and Production
University of Roma, *Tor Vergata*

Abstract

Current Semantic Role Labeling technologies are based on inductive algorithms trained over large scale repositories of annotated examples. Frame-based systems currently make use of the FrameNet database but fail to show suitable generalization capabilities in out-of-domain scenarios. In this paper, a state-of-art system for frame-based SRL is extended through the encapsulation of a distributional model of semantic similarity. The resulting argument classification model promotes a simpler feature space that limits the potential overfitting effects. The large scale empirical study here discussed confirms that state-of-art accuracy can be obtained for out-of-domain evaluations.

1 Introduction

The availability of large scale semantic lexicons, such as FrameNet (Baker et al., 1998), allowed the adoption of a wide family of learning paradigms in the automation of semantic parsing. Building upon the so called *frame* semantic model (Fillmore, 1985), the Berkeley FrameNet project has developed a semantic lexicon for the core vocabulary of English, since 1997. A frame is evoked in texts through the occurrence of its *lexical units* (LU), i.e. predicate words such verbs, nouns, or adjectives, and specifies the participants and properties of the situation it describes, the so called *frame elements* (FEs).

Semantic Role Labeling (SRL) is the task of automatic recognition of individual predicates together with their major roles (e.g. frame elements) as they are grammatically realized in input sentences. It has been a popular task since the availability of the PropBank and FrameNet annotated corpora (Palmer et al., 2005), the seminal work of (Gildea and Jurafsky, 2002) and the successful CoNLL evaluation campaigns (Carreras and Màrquez, 2005). Statistical machine learning methods, ranging from joint probabilistic models to support vector machines, have been successfully adopted to provide very accurate semantic labeling, e.g. (Carreras and Màrquez, 2005).

SRL based on FrameNet is thus not a novel task, although very few systems are known capable of completing a general frame-based annotation process over raw texts, noticeable exceptions being discussed for example in (Erk and Pado, 2006), (Johansson and Nugues, 2008b) and (Coppola et al., 2009). Some critical limitations have been outlined in literature, some of them independent from the underlying semantic paradigm.

Parsing Accuracy. Most of the employed learning algorithms are based on complex sets of syntagmatic features, as deeply investigated in (Johansson and Nugues, 2008b). The resulting recognition *is thus highly dependent on the accuracy of the underlying parser*, whereas wrong structures returned by the parser usually imply large misclassification errors.

Annotation costs. Statistical learning approaches applied to SRL are *very demanding with respect to the amount and quality of the training material.* The complex SRL architectures proposed (usually combining local and global, i.e. joint, models of argument classification, e.g. (Toutanova et al., 2008)) require a large number of annotated examples. The amount and quality of the training data required to reach a significant accuracy is a serious limitation to the exploitation of SRL in many NLP applications.

Limited Linguistic Generalization. Several studies showed that even when large training sets exist *the corresponding learning exhibits poor generalization power*. Most of the CoNLL 2005 systems show a significant performance drop when the tested corpus, i.e. Brown, differs from the training one (i.e. Wall Street Journal), e.g. (Toutanova et al., 2008). More recently, the state-of-art frame-based semantic role labeling system discussed in (Johansson and Nugues, 2008b) reports a 19% drop in accuracy for the argument classification task when a different test domain is targeted (i.e. NTI corpus). Out-of-domain tests seem to suggest the models trained on BNC do not generalize well to novel grammatical and lexical phenomena. As also suggested in (Pradhan et al., 2008), the major drawback is the poor generalization power affecting lexical features. Notice how this is also a general problem of statistical learning processes, as large fine grain feature sets are more exposed to the risks of overfitting.

The above problems are particularly critical for frame-based shallow semantic parsing where, as opposed to more syntactic-oriented semantic labeling schemes (as Propbank (Palmer et al., 2005)), *a significant mismatch exists between the semantic descriptors and the underlying syntactic annotation level.* In (Johansson and Nugues, 2008b) an upper bound of about 83.9% for the accuracy of the argument identification task is reported, it is due to the complexity in projecting frame element boundaries out from the dependency graph: more than 16% of the roles in the annotated material lack of a clear grammatical status.

The limited level of linguistic generalization outlined above is still an open research problem. Existing solutions have been proposed in literature along different lines. Learning from richer linguistic descriptions of more complex structures is proposed in (Toutanova et al., 2008). Limiting the cost required for developing large domain-specific training data sets has been also studied, e.g., (Fürstenau and Lapata, 2009). Finally, the application of semi-supervised learning is attempted to increase the lexical expressiveness of the model, e.g. (Goldberg and Elhadad, 2009).

In this paper, this last direction is pursued. A semi-supervised statistical model exploiting useful lexical information from unlabeled corpora is proposed. The model adopts a simple feature space by relying on a limited set of grammatical properties, thus reducing its learning capacity. Moreover, it generalizes lexical information about the annotated examples by applying a geometrical model, in a Latent Semantic Analysis style, inspired by a distributional paradigm (Pado and Lapata, 2007). As we will see, the accuracy reachable through a restricted feature space is still quite close to the state-of-art, but interestingly the performance drops in out-of-domain tests are avoided.

In the following, after discussing existing approaches to SRL (Section 2), a distributional approach is defined in Section 3. Section 3.2 discusses the proposed HMM-based treatment of joint inferences in argument classification. The large scale experiments described in Section 4 will allow to draw the conclusions of Section 5.

2 Related Work

State-of-art approaches to frame-based SRL are based on Support Vector Machines, trained over linear models of syntactic features, e.g. (Johansson and Nugues, 2008b), or tree-kernels, e.g. (Coppola et al., 2009). SRL proceeds through two main steps: the localization of arguments in a sentence, called *boundary detection (BD)*, and the assignment of the proper role to the detected constituents, that is the *argument classification, (AC)* step. In (Toutanova et al., 2008) a SRL model over Propbank that effectively exploits the semantic argument frame as a joint structure, is presented. It incorporates strong dependencies within a comprehensive statistical joint model with a rich set of features over multiple argument phrases. This approach effectively introduces a new step in SRL, also called *Joint Re-ranking, (RR)*, e.g. (Toutanova et al., 2008) or (Moschitti et al., 2008). First local models are applied to produce role labels over individual arguments, then the joint model is used to decide the entire argument sequence among the set of the n-best competing solutions. While these approaches increase the expressive power of the models to capture more general linguistic properties, they rely on complex feature sets, are more demanding about the amount of training information and increase the overall exposure to overfitting effects.

In (Johansson and Nugues, 2008b) the impact of different grammatical representations on the task of frame-based shallow semantic parsing is studied and the poor lexical generalization problem is outlined. An argument classification accuracy of 89.9% over the FrameNet (i.e. BNC) dataset is shown to decrease to 71.1% when a different test domain is evaluated (i.e. the Nuclear Threat Initiative corpus). The argument classification component is thus shown to be heavily domain-dependent whereas the inclusion of grammatical function features is just able to mitigate this sensitivity. In line with (Pradhan et al., 2008), it is suggested that lexical features are domain specific and their suitable generalization is not achieved.

The lack of suitable lexical information is also discussed in (Fürstenau and Lapata, 2009) through an approach aiming to support the creation of novel annotated resources. Accordingly a semi-supervised approach for reducing the costs of the manual annotation effort is proposed. Through a graph alignment algorithm triggered by annotated resources, the method acquires training instances from an unlabeled corpus also for verbs not listed as existing FrameNet predicates.

2.1 The role of Lexical Semantic Information

It is widely accepted that lexical information (as features directly derived from word forms) is crucial for training accurate systems in a number of NLP tasks. Indeed, all the best systems in the CoNLL shared task competitions (e.g. Chunking (Tjong Kim Sang and Buchholz, 2000)) make extensive use of lexical information. Also lexical features are beneficial in SRL usually either for systems on Propbank as well as for FrameNet-based annotation.

In (Goldberg and Elhadad, 2009), a different strategy to incorporate lexical features into classification models is proposed. A more expressive training algorithm (i.e. anchored SVM) coupled with an aggressive feature pruning strategy is shown to achieve high accuracy over a chunking and named entity recognition task. The suggested perspective here is that effective semantic knowledge can be collected from sources external to the annotated corpora (very large unannotated corpora or on manually constructed lexical resources) rather than learned from the raw lexical counts of the annotated corpus. Notice how this is also the strategy pursued in recent work on deep learning approaches to NLP tasks. In (Collobert and Weston, 2008) a unified architecture for Natural Language Processing that learns features relevant to the tasks at hand given very limited prior knowledge is presented. It embodies the idea that a multitask learning architecture coupled with semi-supervised learning can be effectively applied even to complex linguistic tasks such as SRL. In particular, (Collobert and Weston, 2008) proposes an embedding of lexical information using Wikipedia as source, and exploiting the resulting language model within the multitask learning process. The idea of (Collobert and Weston, 2008) to obtain an embedding of lexical information by acquiring a language model from unlabeled data is an interesting approach to the problem of performance degradation in out-of-domain tests, as already pursued by (Deschacht and Moens, 2009). The extensive use of unlabeled texts allows to achieve a significant level of lexical generalization that seems better capitalize the smaller annotated data sets.

3 A Distributional Model for Argument Classification

High quality lexical information is crucial for robust open-domain SRL, as semantic generalization highly depends on lexical information. For example, the following two sentences evoke the STATEMENT frame, through the LUs *say* and *state*, where the FEs, SPEAKER and MEDIUM, are shown.

[*President Kennedy*] SPEAKER said *to an astronaut, "Man is still the most extraordinary computer of all."* (1)

[*The report*] MEDIUM stated, *that some problems needed to be solved.* (2)

In sentence (1), for example, *President Kennedy* is the grammatical subject of the verb *say* and this justifies its role of SPEAKER. However, syntax does not entirely characterize argument semantics. In (1) and (2), the same syntactic relation is observed. It is the semantics of the grammatical heads, i.e. *report* and *Kennedy*, the main responsible for the difference between the two resulting proto-agentive roles, SPEAKER and MEDIUM.

In this work we explore two different aspects. First, we propose a model that does not depend on complex syntactic information in order to minimize the risk of overfitting. Second, we improve the lexical semantic information available to the learning algorithm. The proposed "minimalistic" approach will consider only two independent features:

- the *semantic head* (h) of a role, as it can be observed in the grammatical structure. In sentence (2), for example, the MEDIUM FE is realized as the logical subject, whose head is *report*.

- the *dependency relation* (r) connecting the semantic head to the predicate words. In (2), the semantic head *report* is connected to the LU *stated* through the subject (SBJ) relation.

In the rest of the section the distributional model for the argument classification step is presented. A lexicalized model for individual semantic roles is first defined in order to achieve robust semantic classification local to each argument. Then a Hidden Markov Model is introduced in order to exploit the local probability estimators, sensitive to lexical similarity, as well as the global information on the entire argument sequence.

3.1 Distributional Local Models

As the classification of semantic roles is strictly related to the lexical meaning of argument heads, we adopt a distributional perspective, where the meaning is described by the set of textual contexts in which words appear. In distributional models, words are thus represented through vectors built over these observable contexts: similar vectors suggest *semantic relatedness* as a function of the distance between two words, capturing paradigmatic (e.g. synonymy) or syntagmatic relations (Pado, 2007). Vectors $\overrightarrow{h}$ are described by an adjacency matrix M, whose rows describe target words (h) and whose columns describe their corpus contexts. Latent Semantic Analysis (LSA) (Landauer and Dumais, 1997), is then applied to M to acquire meaningful representations $\overrightarrow{h}$. LSA exploits the linear transformation called *Singular Value Decomposition* (SVD) and produces an approximation of the original matrix M, capturing (semantic) dependencies between context vectors. M is replaced by a lower dimensional matrix M_l, capturing the same statistical information in a new l-dimensional space, where each dimension is a linear combination of some of the original features (i.e. contexts). These derived features may be thought as artificial concepts, each one representing an emerging meaning component, as the linear combination of many different words.

In the argument classification task, the similarity between two argument heads h_1 and h_2 observed in FrameNet can be computed over $\overrightarrow{h_1}$ and $\overrightarrow{h_2}$. The model for a given frame element FE^k is built around the semantic heads h observed in the role FE^k: they form a set denoted by H^{FE^k}. These LSA vectors $\overrightarrow{h}$ express the individual annotated examples as they are immerse in the LSA space acquired from the unlabeled texts. Moreover, given FE^k, a model for each individual syntactic relation r (i.e. that links h labeled as FE^k to their corresponding predicates) is a partition of the set H^{FE^k} called $H_r^{FE^k}$, i.e. the subset of H^{FE^k} produced by examples of the relation r (e.g. Subj). Given the annotated sentence (2), we have that $report \in H^{\textsc{Medium}}_{\texttt{SBJ}}$.

Role, FE^k	Clusters of semantic heads
Medium	c_1: {*article, report, statement*} c_2: {*constitution, decree, rule*}
Speaker	c_3: {*brother, father, mother, sister* } c_4: {*biographer, philosopher, ...*} c_5: {*he, she, we, you*} c_6: {*friend*}
Topic	c_7: {*privilege, unresponsiveness*} c_8: {*pattern*}

Table 1: Clusters of semantic heads in the Subj position for the frame STATEMENT with $\sigma = 0.5$

As the LSA vectors $\overrightarrow{h}$ are available for the semantic heads h, a vector representation $\overrightarrow{FE^k}$ for the role FE^k can be obtained from the annotated data. However, one single vector is a too simplistic representation given the rich nature of semantic roles FE^k. In order to better represent FE^k, multiple regions in the semantic space are used. They are obtained by a clustering process applied to the set $H_r^{FE^k}$ according to the *Quality Threshold (QT)* algorithm (Heyer et al., 1999). QT is a generalization of k-mean where a variable number of clusters can be obtained. This number depends on the minimal value of intra-cluster similarity accepted by the algorithm and controlled by a parameter, σ: lower values of σ correspond to more heterogeneous (i.e. larger grain) clusters, while values close to 1 characterize stricter policies and more fine-grained results. Given a syntactic relation r, $C_r^{FE^k}$ denotes the clusters derived by QT clustering over $H_r^{FE^k}$. Each cluster $c \in C_r^{FE^k}$ is represented by a vector $\overrightarrow{c}$, computed as the geometric centroid of its semantic heads $h \in c$. For a frame F, clusters define a geometric model of every frame elements FE^k: it consists of centroids $\overrightarrow{c}$ with $c \subseteq H_r^{FE^k}$. Each c represents FE^k through a set of similar heads, as role fillers observed in FrameNet. Table 1 represents clusters for the heads $H^{FE^k}_{\texttt{Subj}}$ of the STATEMENT frame.

In argument classification we assume that the evoking predicate word for the frame F in an input sentence s is known. A sentence s can be seen as a sequence of role-relation pairs:

$s = \{(r_1, h_1), ..., (r_n, h_n)\}$ where the heads h_i are in the syntactic relation r_i with the underlying lexical unit of F.

For every head h in s, the vector $\overrightarrow{h}$ can be then used to estimate its similarity with the different candidate roles FE^k. Given the syntactic relation r, the clusters $c \in C_r^{FE^k}$ whose centroid vector $\vec{c}$ is closer to $\vec{h}$ are selected. $D_{r,h}$ is the set of the representations semantically related to h:

$$D_{r,h} = \bigcup_k \{c_{kj} \in C_r^{FE^k} | sim(h, c_{kj}) \geq \tau\} \quad (3)$$

where the similarity between the *j-th* cluster for the FE^k, i.e. $c_{kj} \in C_r^{FE^k}$, and h is the usual *cosine similarity*: $sim_{cos}(h, c_{kj}) = \frac{\vec{h} \cdot \vec{c}_{kj}}{\|\vec{h}\| \, \|\vec{c}_{kj}\|}$

Then, through a *k-nearest neighbours* (*k*-NN) strategy within $D_{r,h}$, the m clusters c_{kj} most similar to h are retained in the set $D_{r,h}^{(m)}$. A probabilistic preference for the role FE^k is estimated for h through a cluster-based voting scheme,

$$prob(FE^k | r, h) = \frac{|C_r^{FE^k} \cap D_{r,h}^{(m)}|}{|D_{r,h}^{(m)}|} \quad (4)$$

or, alternatively, an instance-based one over $D_{r,h}^{(m)}$:

$$prob(FE^k | r, h) = \frac{\sum_{c \in C_r^{FE^k} \cap D_{r,h}^{(m)}} |c|}{\sum_{c \in D_{r,h}^{(m)}} |c|} \quad (5)$$

In Fig. 1 the preference estimation for the incoming head $h = professor$ connected to a LU by the `Subj` relation is shown. Clusters for the heads in Table 1 are also reported. First, in the set of clusters whose similarity with $professor$ is higher than a threshold τ the $m = 5$ most similar clusters are selected. Accordingly, the preferences given by Eq. 4 are $prob(\text{SPEAKER}|\text{SBJ}, h) = 3/5$, $prob(\text{MEDIUM}|\text{SBJ}, h) = 2/5$ and $prob(\text{TOPIC}|\text{SBJ}, h) = 0$. The strategy modeled by Eq. 5 amplifies the role of larger clusters, e.g. $prob(\text{SPEAKER}|\text{SBJ}, h) = 9/14$ and $prob(\text{MEDIUM}|\text{SBJ}, h) = 5/14$. We call *Distributional*, the model that applies Eq. 5 to the source (r, h) arguments, by rejecting cases *only when* no information about the head h is available from the unlabeled corpus or no example of relation r for the role FE^k is available from the annotated corpus. Eq. 4 and 5 in fact do not cover all possible cases. Often the incoming head h or the relation r may be unavailable:

1. If the head h has never been met in the unlabeled corpus or the high grammatical ambiguity of the sentence does not allow to locate it reliably, Eq. 4 (or 5) should be backed off to a purely syntactic model, that is $prob(FE^k|r)$
2. If the relation r can not be properly located in s, h is also unknown: the prior probability of individual arguments, i.e. $prob(FE^k)$, is here employed.

Both $prob(FE^k|r)$ and $prob(FE^k)$ can be estimated from the training set and smoothing can be also applied[1]. A more robust argument preference function for all arguments $(r_i, h_i) \in s$ of the frame F is thus given by:

$$prob(FE^k | r_i, h_i) = \lambda_1 prob(FE^k | r_i, h_i) + \lambda_2 prob(FE^k | r_i) + \lambda_3 prob(FE^k) \quad (6)$$

where weights λ_1, λ_2, λ_3 can be heuristically assigned or estimated from the training set[2]. The resulting model is hereafter called *Backoff model*: although simply based on a single feature (i.e. the syntactic relation r), it accounts for information at different reliability degrees.

3.2 A Joint Model for Argument Classification

Eq. 6 defines roles preferences local to individual arguments (r_i, h_i). However, an argument frame is a joint structure, with strong dependencies between arguments. We thus propose to model the reranking phase (RR) as a HMM sequence labeling task. It defines a stochastic inference over multiple (locally justified) alternative sequences through a Hidden Markov Model (HMM). It infers the best sequence $FE^{(k_1,...,k_n)}$ over all the possible hidden state sequences (i.e. made by the target FE^{k_i}) given the observable emissions, i.e. the arguments (r_i, h_i). Viterbi inference is applied to build the best (role) interpretation for the input sentence.

Once Eq. 6 is available, the best frame element sequence $FE^{(\theta(1),...,\theta(n))}$ for the entire sentence s can be selected by defining the function $\theta(\cdot)$ that maps arguments $(r_i, h_i) \in s$ to frame elements FE^k:

$$\theta(i) = k \;\; s.t. \;\; FE^k \in F \quad (7)$$

[1] Lindstone smoothing was applied with $\delta = 1$.

[2] In each test discussed hereafter, λ_1, λ_2, λ_3 were assigned to .9,.09 and .01, in order to impose a strict priority to the model contributions.

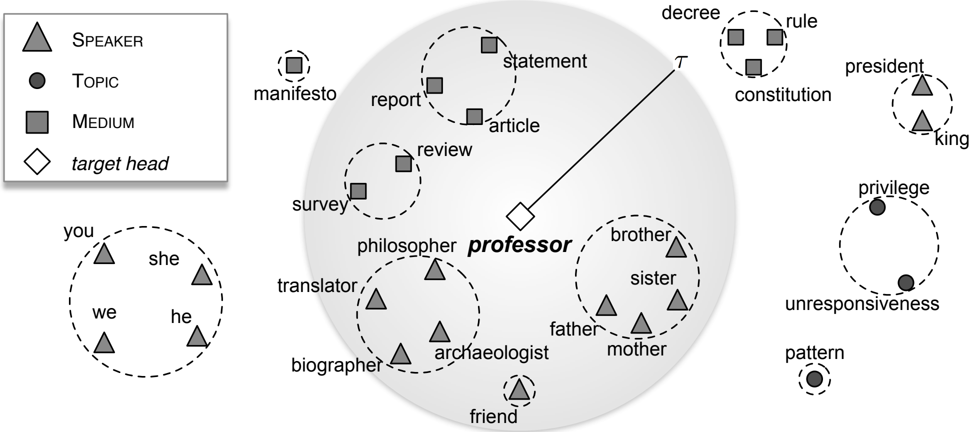

Figure 1: A k-NN approach to the role classification for $h_i = professor$

Notice that different transfer functions $\theta(\cdot)$ are usually possible. By computing their probability we can solve the SRL task by selecting the most likely interpretation, $\widehat{\theta}(\cdot)$, via $\mathrm{argmax}_\theta P\big(\theta(\cdot) \mid s\big)$, as follows:

$$\widehat{\theta}(\cdot) = \underset{\theta}{\mathrm{argmax}}\, P\big(s|\theta(\cdot)\big)P\big(\theta(\cdot)\big) \tag{8}$$

In Eq. 8, the emission probability $P\big(s|\theta(\cdot)\big)$ and the transition probability $P\big(\theta(\cdot)\big)$ are explicit. Notice that the emission probability corresponds to an argument interpretation (e.g. Eq. 5) and it is assigned independently from the rest of the sentence. On the other hand, transition probabilities model role sequences and support the expectations about argument frames of a sentence.

The emission probability is approximated as:

$$P\big(s \mid \theta(1)\ldots\theta(n)\big) \approx \prod_{i=1}^{n} P(r_i, h_i \mid FE^{\theta(i)}) \tag{9}$$

as it is made independent from previous states in a Viterbi path. Again the emission probability can be rewritten as:

$$P(r_i, h_i|FE^{\theta(i)}) = \frac{P(FE^{\theta(i)}|r_i, h_i)\ P(r_i, h_i)}{P(FE^{\theta(i)})} \tag{10}$$

Since $P(r_i, h_i)$ does not depend on the role labeling, maximizing Eq. 10 corresponds to maximize:

$$\frac{P(FE^{\theta(i)}|r_i, h_i)}{P(FE^{\theta(i)})} \tag{11}$$

whereas $P(FE^{\theta(i)}|r_i, h_i)$ is thus estimated through Eq. 6.

The transition probability, estimated through

$$P\big(\theta(1)\ldots\theta(n)\big) \approx \prod_{i=1}^{n} P\big(FE^{\theta(i)}|FE^{\theta(i-1)}, FE^{\theta(i-2)}\big) \tag{12}$$

accounts FEs sequence via a *3*-gram model[3] .

4 Empirical Analysis

The aim of the evaluation is to measure the reachable accuracy of the simple model proposed and to compare its impact over in-domain and out-of-domain semantic role labeling tasks. In particular, we will evaluate the argument classification (AC) task in Section 4.2.

Experimental Set-Up. The in-domain test has been run over the FrameNet annotated corpus, derived from the British National Corpus (BNC). The splitting between train and test set is 90%-10% according to the same data set of (Johansson and Nugues, 2008b). In all experiments, the FrameNet 1.3 version and the dependency-based system using the LTH parser (Johansson and Nugues, 2008a) have been employed. Out-of-domain tests are run over the two training corpora as made available by the Semeval 2007 Task 19[4] (Baker et al., 2007): the Nuclear Threat Initiative (NTI) and the American National Corpus

[3] Two empty states are added at the beginning of any sequence. Moreover, Laplace smoothing was also applied to each estimator.

[4] The NTI and ANC annotated collections are downloadable at:
nlp.cs.swarthmore.edu/semeval/tasks/task19/data/train.tar.gz

	Corpus	Predicates	Arguments
training	FN-BNC	134,697	271,560
test			
in-domain	FN-BNC	14,952	30,173
out-of-domain	NTI	8,208	14,422
	ANC	760	1,389

Table 2: Training and Testing data sets

	Frames with a number of annotated examples					
Eq. - σ	>0	>100	>500	>1K	>3K	>5K
(5) - .85	**86.3**	86.5	87.2	**88.3**	85.9	82.0
(4) - .5	85.1	85.5	85.8	87.2	83.5	79.4
(4) - .1	84.5	84.8	85.1	86.5	83.0	78.7

Table 3: Accuracy on Arg classification tasks wrt different clustering policies

(ANC)[5]. Table 2 shows the predicates and arguments in each data set. All null-instantiated arguments were removed from the training and test sets.

Vectors $\vec{h}$ representing semantic heads have been computed according to the "dependency-based" vector space discussed in (Pado and Lapata, 2007). The entire BNC corpus has been parsed and the dependency graphs derived from individual sentences provided the basic observable contexts: every co-occurrence is thus syntactically justified by a dependency arc. The most frequent 30,000 basic features, i.e. (syntactic relation,lemma) pairs, have been used to build the matrix M, vector components corresponding to point-wise mutual information scores. Finally, the final space is obtained by applying the SVD reduction over M, with a dimensionality cut of $l = 250$.

In the evaluation of the AC task, accuracy is computed over the nodes of the dependency graph, in line with (Johansson and Nugues, 2008b) or (Coppola et al., 2009). Accordingly, also recall, precision and F-measure are reported on a *per node* basis, against the binary BD task or for the full $BD + AC$ chain.

4.1 The Role of Lexical Clustering

The first study aims at detecting the impact of different clustering policies on the resulting AC accuracy. Clustering, as discussed in Section 3.1, allows to generalize lexical information: similar heads within the latent semantic space are built from the annotated examples and they allow to predict the behavior of new unseen words as found in the test sentences. The system performances have been here measured under different clustering conditions, i.e. grains at which the clustering of annotated examples is applied. This grain is determined by the parameter σ of the applied Quality Threshold algorithm (Heyer et al., 1999). Notice that small values of σ imply large clusters, while if $\sigma \approx 1$ then many singleton clusters are promoted (i.e. one cluster for each example). By varying the threshold σ we thus account for prototype-based as well exemplar-based strategies, as discussed in (Erk, 2009).

We measured the performance on the argument classification tasks of different models obtained by combing different choices of σ with Eq. (4) or (5). Results are reported in Table 3. The leftmost column reports the different clustering settings, while in the remaining columns we see performances over test sentences related to different frames: we selected frames for which an increasing number of annotated examples are available: from all frames (for more than 0 examples) to the only frame (i.e. SELF_MOTION) that has more than 5,000 examples in our training data set.

The reported accuracies suggest that Eq. (5), promoting an example driven strategy, better captures the role preference, as it always outperforms alternative settings (i.e. more prototype oriented methods). It limits overgeneralization and promotes fine grained clusters. An interesting result is that a per-node accuracy of 86.3 (i.e. only 3 points under the state of-the art on the same data set, (Johansson and Nugues, 2008b)) is achieved. All the remaining tests have been run with the clustering configuration characterized by Eq. (5) and $\sigma = 0.85$.

4.2 Argument Classification Accuracy

In these experiments we evaluate the quality of the argument classification step against the lexical knowledge acquired from unlabeled texts and the reranking step. The accuracy reachable on the gold standard argument boundaries has been compared across several experimental settings. Two baseline systems have been obtained. The *Local Prior* model outputs the sequence that maximizes the prior probability locally to individual arguments. The *Global Prior* model is obtained by applying re-ranking (Section 3.2) to the best $n = 10$ candidates provided by the *Local Prior* model. Fi-

[5] Sentences whose arguments were not represented in the FrameNet training material were removed from all tests.

Model	FN-BNC	NTI	ANC
Local Prior	43.9	50.9	50.4
Global Prior	67.7 (+54.2%)	75.9 (+49.0%)	68.8 (+36.4%)
Distributional	81.1 (+19.8%)	82.3 (+8.4%)	69.7 (+1.3%)
Backoff	84.6 (+4.3%)	87.2 (+6.0%)	76.2 (+9.3%)
Backoff+HMMRR	86.3 (+2.0%)	90.5 (+3.8%)	79.9 (+5.0%)
(Johansson&Nugues, 2008)	89.9	71.1	-

Table 4: Accuracy of the Argument Classification task over the different corpora. In parenthesis the relative increment with respect to the immediately simpler model, previous row

nally, the application of the backoff strategies (as in Eq. 6) and the HMM-based reranking characterize the final two configurations. Table 4 reports the accuracy results obtained over the three corpora (defined as in Table 2): the accuracy scores are averaged over different values of m in Eq. 5, ranging from 3 to 30. In the in-domain scenario, i.e. the FN-BNC dataset reported in column 2, it is worth noticing that the proposed model, with backoff and global reranking, is quite effective with respect to the state-of-the-art.

Although results on the FN-BNC do not outperform the state-of-the-art for the FrameNet corpus, we still need to study the generalization capability of our SRL model in out-of-domain conditions. In a further experiment, we applied the same system, as trained over the FN-BNC data, to the other corpora, i.e. NTI and ANC, used entirely as test sets. Results, reported in column 3 and 4 of Table 4 and shown in Figure 2, confirm that no major drop in performance is observed. Notice how the positive impact of the backoff models and the HMM reranking policy is similarly reflected by all the collections. Moreover, the results on the NTI corpus are even better than those obtained on the BNC, with a resulting 90.5% accuracy on the AC task.

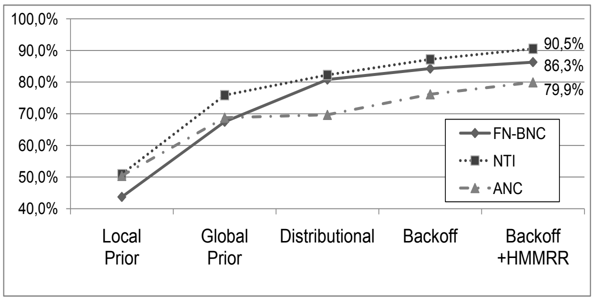

Figure 2: Accuracy of the AC task over different corpora

4.3 Discussion

The above empirical findings are relevant if compared with the outcome of a similar test on the NTI collection, discussed in (Johansson and Nugues, 2008b)[6]. There, under the same training conditions, a performance drop of about -19% is reported (from 89.9 to 71.1%) over gold standard argument boundaries. The model proposed in this paper exhibits no such drop in any collection (NTI and ANC). This seems to confirm the hypothesis that the model is able to properly generalize the required lexical information across different domains.

It is interesting to outline that the individual stages of the proposed model play different roles in the different domains, as Table 4 suggests. Although the positive contributions of the individual processing stages are uniformly confirmed, some differences can be outlined:

- The beneficial impact of the lexical information (i.e. the distributional model) applies differently across the different domains. The ANC domain seems not to significantly benefit when the distributional model (Eq. 5) is applied. Notice how Eq. 5 depends both from the evidence gathered in the corpus about lexical heads h as well as about the relation r. In ANC the percentage of times that the Eq. 5 is backed off against test instances (as h or r are not available from the training data) is twice as high as in the BNC-FN or in the NTI domain (i.e. 15.5 vs. 7.2 or 8.7, respectively). The different syntactic style of ANC seems thus the main responsible of the poor impact of distributional information, as it is often unapplicable to ANC test cases.

- The complexity of the three test sets is different, as the three plots show. The NTI col-

[6] Notice that in this paper only the training portion of the NTI data set is employed as reported in Table 2 and results are not directly comparable to (Johansson and Nugues, 2008b).

lections seems characterized by a lower level of complexity (see for example the accuracy of the *Local prior* model, that is about 51% as for the ANC). It then gets benefits from all the analysis stages, in particular the final HMM reranking. The BNC-FN test collection seems the most complex one, and the impact of the lexical information brought by the distributional model is here maximal. This is mainly due to the coherence between the distributions of lexical and grammatical phenomena in the test and training data.

- The role of HMM reranking is an effective way to compensate errors in the local argument classifications for all the three domains. However, it is particularly effective for the outside domain cases, while, in the BNC corpus, it produces just a small improvement instead (i.e. +2%, as shown in Table 4). It is worth noticing that the average length of the sentences in the BNC test collection is about 23 words per sentence, while it is higher for the NTI and ANC data sets (i.e. 34 and 31, respectively). It seems that the HMM model well captures some information on the global semantic structure of a sentence: this is helpful in cases where errors in the grammatical recognition (of individual arguments or at sentence level) are more frequent and afflict the local distributional model. The more complex is the syntax of a corpus (e.g. in the NTI and ANC data sets), the higher seems the impact of the reranking phase.

The significant performance of the AC model here presented suggest to test it when integrated within a full SRL architecture. Table 5 reports the results of the processing cascade over three collections. Results on the Boundary Detection BD task are obtained by training an SVM model on the same feature set presented in (Johansson and Nugues, 2008b) and are slightly below the state-of-the art BD accuracy reported in (Coppola et al., 2009). However, the accuracy of the complete $BD + AC + RR$ chain (i.e. 68%) improves the corresponding results of (Coppola et al., 2009). Given the relatively simple feature set adopted here, this result is very significant as for its resulting efficiency. The overall BD recognition process is, on a standard architecture, performed at about 6.74 sentences per second, that is basically the same as the time needed for applying the entire $BD + AC + RR$ chain, i.e. 6.21 sentence per second.

Corpus	Eval. Setting	Recall	Precision	F1
BNC	BD	72.6	85.1	78.4
	BD+AC+RR	62.6	74.5	**68.0**
NTI	BD	63.9	80.0	71.0
	BD+AC+RR	56.7	72.1	63.5
ANC	BD	64.0	81.5	71.7
	BD+AC+RR	47.4	62.5	53.9

Table 5: Accuracy of the full cascade of the SRL system over three domain

5 Conclusions

In this paper, a distributional approach for acquiring a semi-supervised model of argument classification (AC) preferences has been proposed. It aims at improving the generalization capability of the inductive SRL approach by reducing the complexity of the employed grammatical features and through a distributional representation of lexical features. The obtained results are close to the state-of-art in FrameNet semantic parsing. State of the art accuracy is obtained instead in out-of-domain experiments. The model seems to capitalize from simple methods of lexical modeling (i.e. the estimation of lexico-grammatical preferences through distributional analysis over unlabeled data), estimation (through syntactic or lexical back-off where necessary) and reranking. The result is an accurate and highly portable SRL cascade. Experiments on the integrated SRL architecture (i.e. $BD + AC + RR$ chain) show that state-of-art accuracy (i.e. 68%) can be obtained on raw texts. This result is also very significant as for the achieved efficiency. The system is able to apply the entire $BD + AC + RR$ chain at a speed of 6.21 sentences per second. This significant efficiency confirms the applicability of the SRL approach proposed here in large scale NLP applications. Future work will study the application of the flexible SRL method proposed to other languages, for which less resources are available and worst training conditions are the norm. Moreover, dimensionality reduction methods alternative to LSA, as currently studied on semi-supervised spectral learning (Johnson and Zhang, 2008), will be experimented.

References

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

Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of SemEval-2007*, pages 99–104, Prague, Czech Republic, June. Association for Computational Linguistics.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proc. of CoNLL-2005*, pages 152–164, Ann Arbor, Michigan, June.

Ronan Collobert and Jason Weston. 2008. A unified architecture for natural language processing: deep neural networks with multitask learning. In *In Proceedings of ICML '08*, pages 160–167, New York, NY, USA. ACM.

Bonaventura Coppola, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Shallow semantic parsing for spoken language understanding. In *Proceedings of NAACL '09*, pages 85–88, Morristown, NJ, USA.

Koen Deschacht and Marie-Francine Moens. 2009. Semi-supervised semantic role labeling using the latent words language model. In *EMNLP '09: Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing*, pages 21–29, Morristown, NJ, USA. Association for Computational Linguistics.

Katrin Erk and Sebastian Pado. 2006. Shalmaneser - a flexible toolbox for semantic role assignment. In *Proceedings of LREC 2006*, Genoa, Italy.

Katrin Erk. 2009. Representing words as regions in vector space. In *In Proceedings of CoNLL '09*, pages 57–65, Morristown, NJ, USA. Association for Computational Linguistics.

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

Hagen Fürstenau and Mirella Lapata. 2009. Graph alignment for semi-supervised semantic role labeling. In *In Proceedings of EMNLP '09*, pages 11–20, Morristown, NJ, USA.

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*, 28(3):245–288.

Yoav Goldberg and Michael Elhadad. 2009. On the role of lexical features in sequence labeling. In *In Proceedings of EMNLP '09*, pages 1142–1151, Singapore, August. Association for Computational Linguistics.

L.J. Heyer, S. Kruglyak, and S. Yooseph. 1999. Exploring expression data: Identification and analysis of coexpressed genes. *Genome Research*, (9):1106–1115.

Richard Johansson and Pierre Nugues. 2008a. Dependency-based syntactic-semantic analysis with propbank and nombank. In *Proceedings of CoNLL-2008*, Manchester, UK, August 16-17.

Richard Johansson and Pierre Nugues. 2008b. The effect of syntactic representation on semantic role labeling. In *Proceedings of COLING*, Manchester, UK, August 18-22.

Rie Johnson and Tong Zhang. 2008. Graph-based semi-supervised learning and spectral kernel design. *IEEE Transactions on Information Theory*, 54(1):275–288.

Tom Landauer and Sue Dumais. 1997. A solution to plato's problem: The latent semantic analysis theory of acquisition, induction and representation of knowledge. *Psychological Review*, 104.

A. Moschitti, D. Pighin, and R. Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34.

Sebastian Pado and Mirella Lapata. 2007. Dependency-based construction of semantic space models. *Computational Linguistics*, 33(2).

Sebastian Pado. 2007. *Cross-Lingual Annotation Projection Models for Role-Semantic Information*. Ph.D. thesis, Saarland University.

Martha Palmer, Dan Gildea, and Paul Kingsbury. 2005. The proposition bank: A corpus annotated with semantic roles. *Computational Linguistics*, 31(1), March.

Sameer S. Pradhan, Wayne Ward, and James H. Martin. 2008. Towards robust semantic role labeling. *Comput. Linguist.*, 34(2):289–310.

Erik F. Tjong Kim Sang and Sabine Buchholz. 2000. Introduction to the conll-2000 shared task: chunking. In *Proceedings of the 2nd workshop on Learning language in logic and the 4th conference on Computational natural language learning*, pages 127–132, Morristown, NJ, USA. Association for Computational Linguistics.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. 2008. A global joint model for semantic role labeling. *Comput. Linguist.*, 34(2):161–191.