Shallow Semantic Parsing for Spoken Language Understanding

Bonaventura Coppola and **Alessandro Moschitti** and **Giuseppe Riccardi**
Department of Information Engineering and Computer Science - University of Trento, Italy
{coppola,moschitti,riccardi}@disi.unitn.it

Abstract

Most Spoken Dialog Systems are based on speech grammars and frame/slot semantics. The semantic descriptions of input utterances are usually defined ad-hoc with no ability to generalize beyond the target application domain or to learn from annotated corpora. The approach we propose in this paper exploits machine learning of frame semantics, borrowing its theoretical model from computational linguistics. While traditional automatic Semantic Role Labeling approaches on written texts may not perform as well on spoken dialogs, we show successful experiments on such porting. Hence, we design and evaluate automatic FrameNet-based parsers both for English written texts and for Italian dialog utterances. The results show that disfluencies of dialog data do not severely hurt performance. Also, a small set of FrameNet-like manual annotations is enough for realizing accurate Semantic Role Labeling on the target domains of typical Dialog Systems.

1 Introduction

Commercial services based on spoken dialog systems have consistently increased both in number and in application scenarios (Gorin et al., 1997). Despite its success, current Spoken Language Understanding (SLU) technology is mainly based on simple conceptual annotation, where just very simple semantic composition is attempted. In contrast, the availability of richer semantic models as FrameNet (Baker et al., 1998) is very appealing for the design of better dialog managers. The first step to enable the exploitation of frame semantics is to show that accurate automatic semantic labelers can be designed for processing conversational speech.

In this paper, we face the problem of performing shallow semantic analysis of speech transcriptions from real-world dialogs. In particular, we apply Support Vector Machines (SVMs) and Kernel Methods to the design of a semantic role labeler (SRL) based on FrameNet. Exploiting Tree Kernels (Collins and Duffy, 2002; Moschitti et al., 2008), we can quickly port our system to different languages and domains. In the experiments, we compare results achieved on the English FrameNet against those achieved on a smaller Italian FrameNet-like corpus of spoken dialog transcriptions. They show that the system is robust enough to disfluencies and noise, and that it can be easily ported to new domains and languages.

In the remainder of the paper, Section 2 presents our basic Semantic Role Labeling approach, Section 3 describes the experiments on the English FrameNet and on our Italian dialog corpus, and Section 4 draws the conclusions.

2 FrameNet-based Semantic Role Labeling

Semantic frames represent prototypical events or situations which individually define their own set of actors, or frame participants. For example, the COMMERCE_SCENARIO frame includes participants as SELLER, BUYER, GOODS, and MONEY. The task of FrameNet-based shallow semantic parsing can be implemented as a combination of multiple specialized semantic labelers as those in (Carreras and Màrquez, 2005), one for each frame. Therefore, the general semantic parsing work-flow includes 4 main steps: (i) *Target Word Detection*, where the semantically relevant words bringing predicative information (the frame *targets*) are detected, e.g. the verb *to purchase* for the above example; (ii) *Frame Disambiguation*, where the correct frame for every target word (which may be ambiguous) is determined, e.g. COMMERCE_SCENARIO; (iii) *Boundary Detection (BD)*, where the sequences of words realizing the frame elements (or predicate arguments) are detected; and (iv) *Role Classification (RC)* (or argument classification), which assigns semantic labels to the frame elements detected in the previous step, e.g. GOODS. Therefore, we implement the full task of FrameNet-based parsing by a combination of multiple specialized SRL-like labelers, one for each frame (Coppola et al., 2008). For the design of each single labeler, we use the state-of-the-art strategy developed in (Pradhan et al., 2005; Moschitti et al., 2008).

2.1 Standard versus Structural Features

In machine learning tasks, the manual engineering of effective features is a complex and time consuming process. For this reason, our SVM-based SRL approach exploits the combination of two different models. We use both Polynomial Kernels over handcrafted, linguistically-motivated features (Gildea and Jurafsky, 2002; Pradhan et al., 2005; Xue and Palmer, 2004), and Tree Kernels (Collins and Duffy, 2002) over automatic structural features (Moschitti et al., 2008).

Concerning the former, there is a common consensus on the set of basic features effective for SRL, which we will refer to as *standard* features. They mostly refer to unstructured information extracted from parse trees. For example, the *Phrase Type* feature indicates the syntactic type of the phrase labeled as a predicate argument; the *Parse Tree Path* feature contains the path in the parse tree between the predicate and the argument phrase; the *Predicate Word* feature is the surface form of the verbal predicate. Standard features proved to be very effective in the typical SRL setting on English. Nonetheless, since we aim at modeling an SRL system for a new language (Italian) and a new domain (dialog transcriptions), the above features may result ineffective. Thus, to achieve independence on the application domain, we exploited structural features proposed in (Moschitti et al., 2005; Moschitti et al., 2008). These are complementary to standard features and are obtained by applying Tree Kernels (Collins and Duffy, 2002; Moschitti et al., 2008) to basic tree structures expressing the syntactic relation between arguments and predicates.

3 Experiments

Our purpose is to show that an accurate automatic FrameNet parser can be designed with reasonable effort for Italian conversational speech. For this purpose, we designed and evaluated both a semantic parser for the English FrameNet (Section 3.1) and one for a corpus of Italian spoken dialogs (Section 3.2). The accuracy of the latter and its comparison against the former can provide evidence to sustain out thesis or not.

3.1 Evaluation on the English FrameNet

In this experiment we trained and tested boundary detectors (BD) and role classifiers (RC) as described in Section 2. More in detail, (a) we trained 5 BDs according to the syntactic categories of the possible target predicates, namely nouns, verbs, adjectives, adverbs and prepositions; (b) we trained 782 one-versus-all multi-role classifiers RC, one for each available frame and predicate syntactic category, for a total of 5,345 binary classifiers; and (c) we applied the above models for recognizing predicate arguments and their associated semantic labels in sentences, where the frame label and the target predicate were considered as given.

3.1.1 Data Set

We exploited the FrameNet 1.3 data base. After preprocessing and parsing the sentences with Charniak's parser, we obtained 135,293 semantically-annotated and syntactically-parsed sentences.

The above dataset was partitioned into three subsets: 2% of data (2,782 sentences) for training the BDs, 90% (121,798 sentences) for training RC, and 1% (1,345 sentences) as test set. The remaining data were discarded. Accordingly, the number of positive and negative training examples for BD were: 2,764 positive and 37,497 negative examples for verbal, 1,189 and 35,576 for nominal, 615 and 14,544 for adjectival, 0 and 40 for adverbial, and 7 and 177 for prepositional predicates (for a total of 4,575 and 87,834). For RC, the total numbers were 207,662 and 1,960,423, which divided by the number of role types show the average number of 39 positive versus 367 negative examples per role label.

3.1.2 Results

We tested several kernels over standard features (Gildea and Jurafsky, 2002; Pradhan et al., 2005) and structured features (Moschitti et al., 2008): the Polynomial Kernel (*PK*, with a degree of 3), the Tree Kernel (TK) and its combination with the bag of word kernel on the tree leaves (TKL). Also, the combinations PK+TK and PK+TKL were tested.

The 4 rows of Table 1 report the performance of different classification tasks. They show in turn: (1) the "pure" performance of the BD classifiers, i.e. considering correct the classification decisions also

	PK			TK			PK+TK			TKL			PK+TKL		
Eval setting	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1	P	R	F_1
BD	.887	.675	.767	.949	.652	.773	.915	.698	**.792**	.938	.659	.774	.908	.701	.791
BD Proj.	.850	.647	.735	.919	.631	.748	.875	.668	**.758**	.906	.636	.747	.868	.670	.757
BD+RC	.654	.498	.565	.697	.479	.568	.680	.519	**.588**	.689	.484	.569	.675	.521	.588
BD+RC Proj.	.625	.476	.540	.672	.462	.548	.648	.495	**.561**	.663	.466	.547	.644	.497	.561

Table 1: Results on FrameNet dataset: Polynomial Kernel, two different Tree Kernels, and their combinations (see Section 3.1.2) with 2% training for BD and 90% for RC.

when a correctly classified tree *node* does not exactly correspond to its argument's *word* boundaries. Such mismatch frequently happens when the parse tree (which is automatically generated) contains incorrect node attachments; (2) the real performance of the BD classification when actually "projected" on the tree leaves, i.e. when matching not only the constituent node as in 1, but also exactly matching the selected *words* (leaves) with those in the FrameNet gold standard. This also implies the exact automatic syntactic analysis for the subtree; (3) the same as in (1), with the argument role classification (RC) also performed (frame element labels must also match); (4) the same as in (2), with RC also performed. For each classification task, the Precision, Recall and F_1 measure achieved by means of different kernel combinations are shown in the columns of the table. Only for the best configuration in Table 1 (PK+TK, results in bold) the amount of training data for the BD model was increased from 2% to 90%, resulting in a popular splitting for this task(Erk and Pado, 2006). Results are shown in Table 2: the PK+TK kernel achieves 1.0 Precision, 0.732 Recall, and 0.847 F_1. These figures can be compared to 0.855 Precision, 0.669 Recall and 0.751 F_1 of the system described in (Erk and Pado, 2006) and trained over the same amount of data. In conclusion, our best learning scheme is currently capable of tagging FrameNet data with exact boundaries and role labels at 63% F_1. Our next steps will be (1) further improving the RC models using FrameNet-specific information (such as Frame and role inheritance), and (2) introducing an effective Frame classifier to automatically choose Frame labels.

3.2 Evaluation on Italian Spoken Dialogs

In this section, we present the results of BD and RC of our FrameNet parser on the smaller Italian spoken dialog corpus. We assume here as well that the target word (i.e. the predicate for which arguments have to be extracted) along with the correct frame are given.

Enhanced PK+TK			
Eval Setting	P	R	F_1
BD (nodes)	1.0	.732	.847
BD (words)	.963	.702	.813
BD+RC (nodes)	.784	.571	.661
BD+RC (words)	.747	.545	.630

Table 2: Results on the FrameNet dataset. Best configuration from Table 1, raised to 90% of training data for BD and RC.

Eval Setting	P	R	F_1	P	R	F_1
					PK	
BD	-	-	-	.900	.869	.884
BD+RC	-	-	-	.769	.742	.756
		TK			**PK+TK**	
BD	.887	.856	.871	.905	.873	**.889**
BD+RC	.765	.738	.751	.774	.747	**.760**

Table 3: Experiment Results on the Italian dialog corpus for different learning schemes and kernel combinations.

3.2.1 Data Set

The Italian dialog corpus includes 50 real human-human dialogs recorded and manually transcribed at the call center of the help-desk facility of an Italian Consortium for Information Systems. The dialogs are fluent and spontaneous conversations between a caller and an operator, concerning hardware and software problems. The dialog turns contain 1,677 annotated frame instances spanning 154 FrameNet frames and 20 new ad hoc frames specific for the domain. New frames mostly concern data processing such as Navigation, Display_Data, Lose_Data, Create_data. Being intended as a reference resource, this dataset includes partially human-validated syntactic analysis, i.e. lower branches corrected to fit arguments. We divided such dataset into 90% training (1,521 frame instances) and 10% testing (156 frame instances). Each frame instance brings its own set of frame participant (or predicate argument) instances.

For BD, the very same approach as in Section 3.1 was followed. For RC, we also followed the same approach but, in order to cope with data sparseness, we also attempted a different RC strategy by merging data related to different syntactic predicates within the same frame. So, within each frame, we merged data related to verbal predicates, nominal predicates, and so on. Due to the short space available, we will just report results for this latter approach, which performed sensitively better.

3.2.2 Results

The results are reported in Table 3. Each table block shows Precision, Recall and F_1 for either PK, TK, or PK+TK. The rows marked as BD show the results for the task of marking the exact constituent boundaries of every frame element (argument) found. The rows marked as BD+RC show the results for the two-stage pipeline of *both* marking the exact constituent boundaries and *also* assigning the correct semantic label. A few observations hold.

First, the highest F_1 has been achieved using the PK+TK combination. On this concern, we underline that kernel combinations *always* gave the best performance in any experiment we run.

Second, we emphasize that the F_1 of PK is surprisingly high, since it exploits the set of standard SRL feature (Gildea and Jurafsky, 2002; Pradhan et al., 2005), originally developed for English and left unmodified for Italian. Nonetheless, their performance is comparable to the Tree Kernels and, as we said, their combination improves the result. Concerning the structured features exploited by Tree Kernels, we note that they work as well without any tuning when ported to Italian dialogs.

Finally, the achieved F_1 is extremely good. In fact, our corresponding result on the FrameNet corpus (Table 2) is P=0.784, R=0.571, F_1=0.661, where the corpus contains much more data, its sentences come from a standard written text (no disfluencies are present) and it is in English language, which is morphologically simpler than Italian. On the other hand, the Italian corpus includes optimal syntactic annotation which exactly fits the frame semantics, and the number of frames is lower than in the FrameNet experiment.

4 Conclusions

The good performance achieved for Italian dialogs shows that FrameNet-based parsing is viable for labeling conversational speech in any language using a few training data. Moreover, the approach works well for very specific domains, like help-desk/customer conversations. Nonetheless, additional tests based on fully automatic transcription and syntactic parsing are needed. However, our current results show that future research on complex spoken dialog systems is enabled to exploit automatically generated frame semantics, which is our very direction.

Acknowledgments

The authors wish to thank Daniele Pighin for the SRL subsystem and Sara Tonelli for the Italian corpus. This work has been partially funded by the European Commission - LUNA project (contract n.33549), and by the Marie Curie Excellence Grant for the ADAMACH project (contract n.022593).

References

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley FrameNet project. In *Proceedings of COLING-ACL '98*, pages 86–90.

Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2005*.

Michael Collins and Nigel Duffy. 2002. New Ranking Algorithms for Parsing and Tagging: Kernels over Discrete structures, and the voted perceptron. In *ACL02*.

Bonaventura Coppola, Alessandro Moschitti, and Daniele Pighin. 2008. Generalized framework for syntax-based relation mining. In *IEEE-ICDM 2008*.

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

Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*.

A. L. Gorin, G. Riccardi, and J. H. Wright. 1997. How may i help you? *Speech Communication*.

Alessandro Moschitti, Bonaventura Coppola, Daniele Pighin, and Roberto Basili. 2005. Engineering of syntactic features for shallow semantic parsing. In *ACL WS on Feature Engineering for ML in NLP*.

Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224.

Sameer Pradhan, Kadri Hacioglu, Valerie Krugler, Wayne Ward, James H. Martin, and Daniel Jurafsky. 2005. Support Vector Learning for Semantic Argument Classification. *Machine Learning*.

Nianwen Xue and Martha Palmer. 2004. Calibrating features for semantic role labeling. In *Proceedings of EMNLP 2004*.