

A General Purpose FrameNet-based Shallow Semantic Parser

Bonaventura Coppola, Alessandro Moschitti

Department of Information Engineering and Computer Science
University of Trento
Via Sommarive, 14 - 38123 Povo (Trento), Italy
coppola@disi.unitn.it, moschitti@disi.unitn.it

Abstract

In this paper we present a new FrameNet-based Shallow Semantic Parser. While Shallow Semantic Parsing has been a popular Natural Language Processing task since the 2004 and 2005 CoNLL Shared Task editions, efforts in extending such task to the FrameNet setting have been constrained by practical software engineering issues. We hereby analyze these issues, identify desirable requirements for a practical parsing framework, and show the results of our software implementation. In particular, we attempt at meeting requirements arising from both a) the need of a flexible environment supporting current ongoing research, and b) the willingness of providing an effective platform supporting preliminary application prototypes in the field. After introducing the task of FrameNet-based Shallow Semantic Parsing, we sketch the system processing workflow and summarize a set of successful experimental results, directing the reader to previous published papers for extended experiment descriptions and wider discussion of the achieved results.

1. Motivations

In this paper we introduce a general purpose multi-language and multi-domain FrameNet-based Shallow Semantic Parser. In recent years, Shallow Semantic Parsing (SSP) has been attracting a remarkable interest as it provides solutions for the design of advanced applications of natural language processing (Moschitti et al., 2003). Several SSP systems have been developed since the CoNLL shared tasks on Semantic Role Labeling (SRL) centered on the PropBank resource (Palmer et al., 2005). However, Frame Semantics (Fillmore, 1968) and the FrameNet¹ lexical/semantic resource (Baker et al., 1998) have been recently gaining increasing popularity for their capability of managing natural language semantics at a finer-grain level. Nonetheless, very few effective implementations for FrameNet-based SSP have been publicly proposed, e.g. (Erk and Pado, 2006). Moreover, most development efforts were devoted to single and specific evaluation tasks as in SemEval 2007 (Baker et al., 2007). This is not surprising given the inherent complexity of such parsers. To better understand this, it is enough to consider FrameNet parsers as an extension of PropBank-based SRL software systems. When passing from the design of the latter to the former, the very first arising issues, well known to NLP developers include unavoidably:

- The increase from ~50 Semantic Roles to more than 4000 Frame Elements requiring individual classifiers (Scalability issue).
- The introduction of Context/Frame-dependent semantic labels (modularity issue).
- The necessity of managing thousands of different machine learning models by continuous loading/unloading (speed issue).
- The presence of non-verbal predicates (ultimately, data sparseness issue)

The whole set of such practical problems, though reasonably manageable under a software engineering perspective, has kept very high the cost of developing an effective semantic parser from a research point of view.

2. Requirements and Solutions

Even considering the difficult setting described in the previous motivations, it would still be highly desirable to have a flexible experimental framework available, supporting at the same time a) the ongoing *research* activity in the field and b) the development of effective *application prototypes*. Concerning the research side, the main requirements are:

1. Easy integration of new algorithms and approaches.
2. Batch training/test execution allowing multiple configurations.
3. Sound algorithm/configuration evaluation and error analysis.

On the other hand, the requirements for applications would be:

4. Fast annotation of large scale text corpora (e.g. from Web Crawling).
5. An effective development cycle for building ad-hoc semantic parsers.
6. Quick adaptation to new languages and application domains.

Starting from the whole set of above requirements, we hereby propose a general purpose multi-language and multi-domain FrameNet-based Shallow Semantic Parser showing the following features:

Modularity has been considered the most critical development principle. It allows for the easy management of multiple learning models and of several subsystems as multiple (possibly redundant) linguistic preprocessors, e.g. different syntactic parsers.

¹Project homepage: <http://framenet.icsi.berkeley.edu>

Scalability with respect to the input text size is achieved by allowing transparent execution over multiple CPUs and multiple servers. The only requirement enabling massive parallel execution is the availability of a Network File System shared by the servers, which execute identical instances of the parser. Also, the amount of CPUs dedicated to an annotation job can be dynamically changed during the execution.

Platform Portability is achieved through a pure, clean Java implementation of the whole architecture. Nonetheless, limited portions of C code have been kept for critical subsystems as the SVM-Light machine learning package (Joachims, 1999).

Performance Optimization has been carried out by analyzing the overall system workflow and identifying the most critical performance bottlenecks. Such analysis asked for a customized rework of the SVM-Light "Tree Kernel" extension (Moschitti, 2006), which now includes a specific machine learning model caching capability.

Flexibility of the experimental framework is obtained as a direct consequence of architectural modularity. It enables the introduction of different learning models, e.g. standard SRL features, Syntactic Tree Kernels, or any combination of them.

Language Portability is granted by the priority given to pure statistical approaches. A constituency-based syntactic parser is currently the only language-dependent (and often retrainable) module.

Domain Portability is achieved by avoiding any hard-coded knowledge and relying on a data driven approach. This brings the advantage of effortless adaptation to the training data and their annotation (i.e. new different frame set definitions, new frame elements, etc.)

Different Frame Learning Configurations allow for different data aggregation schemes, which lead to different models. They include per-frame, selective, and aggregate learning, and can be triggered in order to reduce problems due to data sparseness.

Different Execution Modalities have been implemented to enable different annotation tasks. Specific modes allow for Online (user interactive), Batch (whole corpus processing), and Client/Server (application oriented) exploitation of the parser.

A Configuration Management mechanism has been implemented, imposing that the system behavior for any specific annotation job is solely and completely defined by a set of XML configuration files, which specify any possible execution parameter.

3. FrameNet-based Semantic Annotation

Frame Semantics (Fillmore, 1968) allows for real-world knowledge to be captured by semantic frames, script-like

conceptual structures that describe particular types of situations, objects, or events along with their participating elements. For example, here is a short definition of a sample frame:

COMMERCE_SCENARIO

Core Elements: BUYER, GOODS, MONEY, SELLER

non-Core Elements: MANNER, MEANS, PURPOSE, RATE

Subframes: COMMERCIAL_TRANSACTION

where the core frame elements are participant entities which are supposed to be always present, whereas non-core are just optional, more generic participants. Frame-to-frame relations are also defined, like the *Subframe* relation which states here a hierarchical dependency of the COMMERCIAL_TRANSACTION frame. The Berkeley FrameNet Project (Baker et al., 1998) currently includes the definitions of nearly 800 frames, 4,000 frame elements, and 135,000 annotated English sentences. An example of sentence annotation for the COMMERCE_SCENARIO is reported hereafter:

Ralemborg said [he]_{SELLER} already had a [buyer]_{BUYER} [for the wine]_{GOODS}

where the underlined word *buyer* is the target word (or *lexical unit*, or predicate) which plays the role of *evoker* for this particular frame. In order to automatically parse this information from plain text exploiting a machine learning approach, we need in general (a) to represent the relation between the target word and the words compounding an argument in terms of feature vectors, and then (b) to learn classification models able to process such vectors. Such approach is presented in deeper detail in the next section.

4. The Automatic Annotation Workflow

To implement a FrameNet-based parsing system we adopt a multi-stage classification scheme over natural language text. Previous studies in this direction apply Semantic Role Labeling (SRL) approaches (Gildea and Jurafsky, 2002). We extended the same strategy developed in (Moschitti et al., 2008; Moschitti et al., 2005b), that exploits a strict-pipelined architecture and now includes the following stages:

1. *Target Word Detection*, where the semantically relevant words bringing predicative information are detected;
2. *Frame Disambiguation*, in which the correct frame for any target word is chosen;
3. *Boundary Detection (BD)*, where the sequences of words constituting the frame elements (arguments) are detected;
4. *Role Classification (RC)*, which assigns semantic labels to the frame elements detected in the previous stage.

The first two stages can be carried out in several ways (depending on the application), which include heuristics based

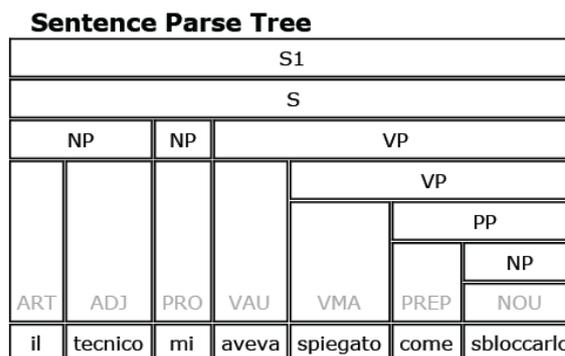
on FrameNet lexical units found in the text, or traditional supervised multi-classification approaches. *BD* is typically carried out as a binary classification problem, where the classification instances are the nodes of the syntactic parse tree of the considered sentence (or dialog turn). Indeed the arguments of a predicate, according to some linguistic theories, are univocally associated with syntactic constituents, i.e. the internal parse tree nodes. At training time, the positive examples are the nodes corresponding to arguments, whereas all the remaining nodes are negative examples. Although dependency-based syntactic analysis is often considered closer to semantics, we still exploit here constituency-based parsing as a legacy approach historically originating from the past CoNLL shared tasks on Semantic Role Labeling (Carreras and Màrquez, 2004; Carreras and Màrquez, 2005). *RC* is a multi-classification problem over the set of the possible labels for an argument (with respect to the chosen frame). Even in this case, role labels are strictly associated with internal tree nodes as selected in the previous stage. The representation of the nodes in a learning algorithm is traditionally carried out by exploiting syntactic information, since syntax is strongly linked to semantics. Many features for representing the nodes have been provided (Gildea and Jurafsky, 2002), which form the vectors to train SVMs. We further exploit the potential of SVMs by using kernel methods, so we use Tree Kernels to encode the subtree which includes a target word and one of its arguments into the learning algorithm, as shown in (Moschitti et al., 2008). It is worth emphasizing the relevance of this double approach.

In machine learning tasks, the manual engineering of effective features is a complex and time consuming process. For this reason, our SVM-based parsing approach exploits the combination of two different models. We first used Polynomial Kernels over handcrafted, linguistically-motivated, “standard” SRL features (Gildea and Jurafsky, 2002; Pradhan et al., 2005; Xue and Palmer, 2004). Nonetheless, since we aim at modeling a Semantic Parsing System for other languages than English (as Italian) and different possible domains, the above features may result ineffective. Thus, to achieve independence on the application domain, we exploited Tree Kernels (Collins and Duffy, 2002) over automatic structural features proposed in (Moschitti et al., 2005a; 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.

5. Current Applications and Results

The performance and effectiveness of our parser has been tested in very diverse experimental settings. We hereby refer to already published experiments and results, providing the reader with references to papers including detailed descriptions and discussions.

The evaluation on the standard Berkeley FrameNet DataBase, with a standard experiment setting exploiting the whole dataset, resulted in Precision=74.7%, Recall=54.5%, and F=63% for the task of recognizing exact text boundaries and semantic labels of the Frame Elements appearing



TARGET WORD = spiegato

FRAME = Statement [SCORE=1.0000]

Speaker	Addressee		Target	Message
il tecnico	mi	aveva	spiegato	come sbloccarlo

Figure 1: Processing example (courtesy of LUNA Project). It shows the syntactic parsing and the FrameNet-based semantic analysis for the Italian sentence “*il tecnico mi aveva spiegato come sbloccarlo*”.

in the experiment test set, i.e. the steps 3 and 4 as described in Section 4. The detailed description of this experiment setting as well as the discussion of achieved results is included in (Coppola et al., 2008a)

A first extensive, application-oriented exploitation of the parser has been conducted in the framework of the LUNA European IST Project². The LUNA Project addressed the problem of real time understanding of spontaneous speech in the context of advanced telecom services, and it applies to Italian, French and Polish. As a first step, the project has made available a benchmark collection of Italian dialogs collected at a real helpdesk service, which have been transcribed and annotated with different layers including syntax and Frame Semantics (Dinarelli et al., 2009). This corpus provided a very different experimental setting due to the Italian language, the very specific application domain (hardware/software assistance) and the very different nature of the text (spoken language transcription versus written text). Moreover, the annotation workflow of such corpus included the definition of novel frame definitions ad related Frame Elements, thus emphasizing the adaptivity of the parser. The results achieved in this application-oriented setting were Precision=77.4%, Recall=74.7%, and F=76%, exploiting exactly the same learning model which generated the results for English as reported earlier. The detailed experiment and dataset description for this Italian annotation task are included in (Coppola et al., 2008b), while the comparison between the two English and Italian settings is discussed in (Coppola et al., 2009b).

Concerning the running time, the execution performance of the parser allowed for its inclusion as a real-time analysis module for on line user interaction by the LUNA Project

²Project Homepage: <http://www.ist-luna.eu>

Consortium. An online demo of such real-time module (for Italian) is currently maintained and publicly accessible³. Figure 1 presents the typical output for a sample sentence, showing the syntactic parsing, and the FrameNet-based semantic analysis, i.e. the identified target word, the invoked frame, and the frame element instances along with their semantic role labels.

Additional exploitations of our FrameNet-based Shallow Semantic Parser are being performed in the fields of ontology learning (Coppola et al., 2009a) and development of cross-lingual development of semantic resources (Basili et al., 2009).

Acknowledgments

The authors wish to thank Daniele Pighin for the development of critical software modules, Sara Tonelli for the development of the Italian data set, and Alberto Lavelli for his longstanding effort in porting several syntactic parsers to Italian language. The experiments on Italian spoken dialogs have been supported by Giuseppe Riccardi and the LUNA IST European Project Consortium. Pierluigi Roberti created an effective graphic representation for LUNA, from which Figure 1 is taken.

6. 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.
- Collin Baker, Michael Ellsworth, and Katrin Erk. 2007. Semeval-2007 task 19: Frame semantic structure extraction. In *Proceedings of the Fourth International Workshop on Semantic Evaluations (SemEval-2007)*, pages 99–104, Prague, Czech Republic, June. ACL.
- Roberto Basili, Diego De Cao, Danilo Croce, Bonaventura Coppola, and Alessandro Moschitti. 2009. Cross-language frame semantics transfer in bilingual corpora. In *CICLing*, pages 332–345.
- Xavier Carreras and Lluís Màrquez. 2004. Introduction to the conll-2004 shared task: Semantic role labeling. In Hwee Tou Ng and Ellen Riloff, editors, *HLT-NAACL 2004 Workshop: Eighth Conference on Computational Natural Language Learning (CoNLL-2004)*, pages 89–97, Boston, Massachusetts, USA, May 6 - May 7. Association for Computational Linguistics.
- Xavier Carreras and Lluís Màrquez. 2005. Introduction to the CoNLL-2005 Shared Task: Semantic Role Labeling. In *Proceedings of CoNLL-2005*, pages 152–164, Ann Arbor, Michigan, June.
- 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. 2008a. Generalized framework for syntax-based relation mining. In *IEEE-ICDM 2008*.
- Bonaventura Coppola, Alessandro Moschitti, Sara Tonelli, and Giuseppe Riccardi. 2008b. Automatic framenet-based annotation of conversational speech. In *Proceedings of the IEEE Workshop on Spoken Language Technology (SLT 2008)*, Goa, India.
- Bonaventura Coppola, Aldo Gangemi, Alfio Gliozzo, Davide Picca, and Valentina Presutti. 2009a. Frame detection over the semantic web. In *ESWC 2009 Heraklion: Proceedings of the 6th European Semantic Web Conference on The Semantic Web*, pages 126–142, Berlin, Heidelberg. Springer-Verlag.
- Bonaventura Coppola, Alessandro Moschitti, and Giuseppe Riccardi. 2009b. Shallow semantic parsing for spoken language understanding. In *Proceedings of HLT-NAACL 2009, Companion Volume: Short Papers*, pages 85–88, Boulder, Colorado, June. ACL.
- Marco Dinarelli, Silvia Quarteroni, Sara Tonelli, Alessandro Moschitti, and Giuseppe Riccardi. 2009. Annotating spoken dialogs: From speech segments to dialog acts and frame semantics. In *Proceedings of SRSLL 2009 Workshop*, pages 34–41, Athens, Greece, March. 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.
- Charles J. Fillmore. 1968. The Case for Case. In Emmon Bach and Robert T. Harms, editors, *Universals in Linguistic Theory*, pages 1–210. Holt, Rinehart, and Winston, New York.
- Daniel Gildea and Daniel Jurafsky. 2002. Automatic Labeling of Semantic Roles. *Computational Linguistics*.
- Thorsten Joachims. 1999. Making large-scale SVM learning practical. In B. Schölkopf, C. Burges, and A. Smola, editors, *Advances in Kernel Methods - Support Vector Learning*, pages 169–184.
- Alessandro Moschitti, Paul Morarescu, and Sanda M. Harabagiu. 2003. Open domain information extraction via automatic semantic labeling. In *FLAIRS Conference*, pages 397–401.
- Alessandro Moschitti, Bonaventura Coppola, Daniele Pighin, and Roberto Basili. 2005a. Engineering of syntactic features for shallow semantic parsing. In *ACL WS on Feature Engineering for ML in NLP*.
- Alessandro Moschitti, Ana-Maria Giuglea, Bonaventura Coppola, and Roberto Basili. 2005b. Hierarchical semantic role labeling. In *CoNLL 2005 shared task*.
- Alessandro Moschitti, Daniele Pighin, and Roberto Basili. 2008. Tree kernels for semantic role labeling. *Computational Linguistics*, 34(2):193–224.
- Alessandro Moschitti. 2006. Making Tree Kernels Practical for Natural Language Learning. In *Proceedings of EAACL2006*.
- Martha Palmer, Daniel Gildea, and Paul Kingsbury. 2005. The Proposition Bank: an Annotated Corpus of Semantic Roles. *Computational Linguistics*, 31(1):71–106.
- 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*.

³Demo URL: <http://cicerone.dit.unitn.it/FrameSemantics/Demo.php>