Structured Kernel-Based Learning for the Frame Labeling over Italian Texts

Danilo Croce, Emanuele Bastianelli, and Giuseppe Castellucci

Department of Enterprise Engineering
University of Roma, Tor Vergata
Via del Politecnico 1, 00133 Roma
{croce,bastianelli,castellucci}@info.uniroma2.it

Abstract. In this paper two systems participating to the Evalita *Frame Labeling over Italian Texts* challenge are presented. The first one, i.e. the SVM-SPTK system, implements the Smoothed Partial Tree Kernel that models semantic roles by implicitly combining syntactic and lexical information of annotated examples. The second one, i.e. the SVM-HMM system, realizes a flexible approach based on the Markovian formulation of the SVM learning algorithm. In the challenge, the SVM-SPTK system obtains state-of-the-art results in almost all tasks. Performances of the SVM-HMM system are interesting too, i.e. the second best scores in the Frame Prediction and Argument Classification tasks, especially considering it does not rely on a full syntactic parsing.

Keywords: Semantic Role Labeling, Structured Kernel-Based Learning, SVM

1 Introductions

Language learning systems usually generalize linguistic observations into statistical models of higher level semantic tasks, such as Semantic Role Labeling (SRL). Lexical or grammatical aspects of training data are the basic features for modeling the different inferences, then generalized into predictive patterns composing the final induced model. In SRL, the role of grammatical features has been outlined since the seminal work by [10], where symbolic expressions derived from the parse trees denote the position and the relationship between a predicate and its arguments, and they are used as features.

As discussed in [5, 8, 13], syntactic information of annotated examples can be effectively generalized in SRL through the adoption of tree kernel based learning ([4]), without the need of manual feature engineering: as tree kernels model similarity between two training examples as a function of their shared tree fragments, discriminative informations are automatically selected by the learning algorithm, e.g., Support Vector Machines (SVMs). However, when the availability of training data is limited, the information derived from structural patterns cannot be sufficient to discriminate examples. According to the Frame Semantics [3], two phrases like *"The man said ... "* and *"The mail said ... "* both evoke the JUDGMENT_COMMUNICATION frame[1] but the two logical subjects represent two different roles: *man* represents a human being, then associated to the COMMUNICATOR role, while *mail* is a media, therefore associated to the

[1] The frame is here evoked by the lexical unit *said*

MEAN role. Lexical information should be captured as it models fine grained context dependent aspects of the input data. One main limitation of tree kernels is that a hard matching among tree node labels is usually applied. If a train example contains *man* while a test case contains *child*, they are considered different without contributing to the overall similarity estimation. To overcome such issues, in [8] the definition of a semantically *Smoothed Partial Tree Kernel* (SPTK) has been provided to augment tree kernel formulation with node similarity, e.g. between the lexical nodes. The idea is to provide a similarity score among tree nodes depending on the semantic similarity among the node labels, e.g. *man* and *child*. SPTK can thus automatically provide the learning algorithm, with a huge set of generalized structural patterns by simply applying it to the structural representation of the target training instances. A meaningful similarity measure is thus crucial, as the lack of proper lexical generalization is often claimed as the main responsible for significant performance drops in out-of-domain SRL [12]. As the development of large scale lexical KBs is very expensive, corpus-driven methods are traditionally used to acquire meaning generalizations in an unsupervised fashion (e.g. [14]) through the analysis of distributions of word occurrences in texts. In line with previous works, (e.g. [7]) we extends a supervised approach through the adoption of vector based models of lexical meaning: a large-scale corpus is statistically analyzed and a geometrical space (the Word Space discussed in [16]) is defined. Here words are modeled as vectors whose dimensions reflect the words co-occurrence statistics over texts. The similarity (or distance) among vectors corresponds to a notion of semantic similarity among the corresponding words. This approach has been implemented in the **SVM-SPTK** system and his performances have been evaluated in the Evalita 2011 *Frame Labeling over Italian Texts* (*FLaIT*) challenge.

However, there is no free lunch in the adoption of grammatical features in complex NLP tasks. Methods for extracting grammatical features from parse trees are strongly biased by the parsing quality. In [15] experiments over gold parse trees are reported with an accuracy (93%) significantly higher than the ones derived by using automatically derived trees (i.e. 79%). Moreover, in [12] the adoption of the syntactic parser has been shown to restrict the correct treatment of FrameNet roles to only the 82% of them, i.e. the only ones that are grammatically recognized. A radically different approach is here pursued as a possible solution to the above problems. While parsing accuracy highly varies across corpora, the adoption of shallower features (e.g. POS n-grams) increases robustness, applicability and minimizes overfitting. In [6] the SRL task is modeled as a sequential tagging problem through the adoption of shallow grammatical features that avoid the use of a full parser. The learning framework is provided by the SVM^{hmm} formulation discussed in [1], that extends classical SVMs by learning a discriminative model isomorphic to a k-order Hidden Markov Model through the Structural SVM formulation [17]. Each word is then modeled as a set of linear features that express lexical information as well as syntactic information surrogated by POS n-grams. Another system has been thus developed for the challenge, i.e. the **SVM-HMM** based system, that aims to increase the applicability of SRL tagging without strict requirements in terms of training data. In the rest of this work, Section 2 describes both SVM-SPTK and SVM-HMM systems. Section 3 reports results achieved in the *FLaIT* challenge. Finally, in Section 4 conclusions are derived.

2 Systems Description

In this section two different systems of SRL, implementing different structured kernel-based Support Vector Machine (SVM) learning algorithms are presented.

2.1 The SVM-SPTK system

The SVM-SPTK system is based on the semantically Smoothed Partial Tree Kernel (SPTK) described in [8]. It extends the Tree Kernel formulation, which measures the structural similarity of syntactic parse trees, by accounting on the lexical information too. This is estimated according to a geometrical perspective: as discussed in [16], a large-scale corpus is statistically analyzed and a geometrical Word Space is acquired. As proposed in [8], examples are modeled according the Grammatical Relation Centered Tree (GRCT) representation from the original dependency parse structures, i.e. no manual feature engineering is needed.

The *Frame Prediction* (FP) task is modeled as a classification problem. Every lexical unit lu found in a sentence s determines an example, indicated as the pair $\langle lu, s\rangle$. Each example is modeled through the GRCT representation of s, i.e. no manual feature engineering is applied. The node corresponding to a lu is enriched with the special token `LU` to distinguish sentences containing different lus. A model for each frame, i.e. the target class, is acquired and a One-VS-All classification schema is adopted.

For the *Boundary Detection* (BD) task, each node in the dependency parse tree is a candidate node covering a word span evoking a role (i.e. a Frame Element, fe) and the classifier discriminates nodes perfectly covering a predicate argument. The frame information provided at the FP step is ignored, while models for different POS, i.e. verbs (`V`), nouns (`N`) and adjectives (`ADJ`), are acquired. This separation is needed as predicates in different POS classes may have very different syntactic behaviors. In each example the target node and the covered ones are then enriched with the `ARG` label and all nodes that do not cover a fe nor the lu are pruned out. It is useful as the complexity of parse trees grows exponentially with the sentence length, thus compromising the generalization capability of the SVM resulting model. In the *Argument Classification (AC)* task, only nodes actually covering a fe are preserved. Examples are divided by frame and a One-VS-All schema is applied, i.e. and a model for each fe is acquired.

2.2 The SVM-HMM system

The SVM-HMM implements an agile system that adopts only shallow grammatical features ignoring the full syntactic information of a sentence. The *Frame Prediction* (FP) task is modeled as a classification problem where each example is given by the $\langle lu, s\rangle$ pair and the frame f indicates the target class. Each instance is modeled as a set of manually engineered contextual features: the lexical and syntactic contexts are captured by the m words and the POS n-grams around the lu. The symbol `LU` is used to better characterize the target predicate within any n-gram. The multi-classification schema described in [11] is applied, thus defining a single classifier that implicitly compares all solution and select the most likely one.

For the *Boundary Detection* (BD) and the *Argument Classification* (AC) tasks, the approach defined in [6] is adopted. The labeling problem is modeled as a sequential tagging task thus extending a SVM by learning a discriminative model isomorphic to a k-order Hidden Markov Model. With respect to BD, each token represents the beginning (B), the inside (I) or outside (O) of an argument or it can be simply external (X) to every argument. The BD task is thus a sequence labeling process that determines the individual (correct BIO) class for each token, e.g. "*The*/B *man*/O *said*/LU ... ". Models for different lu POS are acquired as for the previous system.

The AC task is realized in a similar fashion, i.e. once the BIO notation for each argument is available, each token inside a boundary is classified with respect to the corresponding role. Each frame is characterized by a single classifier as the SVM^{hmm} formulation implicitly realizes a multi-classification as well as a re-ranking schema. The role label most frequently assigned to the inner members of a boundary is retained as the unique role. For both BD and AC each instance, i.e. each words, is modeled as a set of manually engineered features as in [6] and a linear kernel is applied to compare feature vectors in all the three tasks. The cutting-plane learning algorithm allows to train our linear classifiers very efficiently as described in [11].

3 Results

In this section, results achieved in the Evalita 2011 *FLaIT* challenge are reported. Both systems are trained using 1255 annotated sentences provided as the training set. Parameter tuning has been carried out according a 5-fold cross validation schema. Syntactic trees of the 318 test sentences have been manually checked, as the TANL parser [2] diverged in several sentences, providing inconsistent syntactic labeling. However the training sentences were not checked to measure the system robustness when trained over real but noisy data. The lexical generalization is provided by a word space acquired from the Italian Wikipedia corpus[2]. Here lemmatized and POS tagged words that occur in the corpus more than 200 times have been selected, thus reducing data-sparseness. Each target word tw corresponds to a row in the adjacency matrix M, i.e. a point in the resulting space. Each column of M represents a word in the corpus and each item determines the point-wise mutual information (*pmi*) score that estimates the number of times this word co-occurs with tw in a window of size ± 3. The most frequent 20,000 items are thus selected. A dimensionality space reduction based on Singular Value Decomposition is then applied as described in [9] to reduce the space dimensionality to N=250. The similarity between words is thus expressed as the cosine similarity between the corresponding vectors in such reduced space.

Frame Prediction (FP): In the FP task, the SVM-SPTK system correctly determined the evoked frame for the 80.82% of test sentences, thus achieving best results with respect to this task. The SVM-HMM achieved a close accuracy score, i.e. 78.62%. It seems that the syntactic information of the sentence was not discriminative for this particular task and the shallow grammatical patterns represent a valuable information.

[2] The corpus is developed by the WaCky community and it is available in the Wacky project web page at http://medialab.di.unipi.it/Project/QA/wikiCoNLL.bz2

Table 1. Evalita 2011 - Boundary Detection Results

	System	Argument-Based			Token-Based		
		P	R	F1	P	R	F1
First Run	SVM-SPTK	**66.67%**	**72.50%**	**69.46%**	**81.99%**	**84.34%**	**83.15%**
	SVM-HMM	50.70%	51.43%	51.06%	68.02%	77.18%	72.31%
Second Run	SVM-SPTK	66.67%	**72.50%**	**69.46%**	81.99%	84.34%	83.15%
	SVM-HMM	49.91%	50.36%	50.13%	68.14%	76.69%	72.16%

Boundary Detection (BD): In Table 1 results obtained in the BD task are reported. In the First challenge run, gold standard frame information is not provided and it must be automatically induced. On the contrary this information is provided by the organizers in the Second challenge run. In both cases, SVM-SPTK system achieves state-of-the-art results for the perfect detection of semantic roles, i.e. the SPTK based classifier can effectively exploit the combination of syntactic information and lexical generalization to acquire a robust model of semantic roles. In the token based BD measure, a different system achieved better (even if very close) results, i.e. our approach tends to neglect some words in the role spans. This phenomenon need to be investigated when gold-standard results will be provided by organizers. According to the perfect role detection measure, the SVM-HMM system shows an important performance drop of nearly 19% in terms of F1. Even if this drop is reduced according to the token based measure (i.e. nearly 11% of F1), the adoption of shallow grammatical information seems not to be the best solution in this such training condition, i.e. only 1255 training sentences. Here different arguments are not retrieved at all. The sequences of part-of-speech patterns represent a sparse source of information that penalizes the resulting system recall.
Argument Classification (AC): In Table 2 results for the AC task are reported. Notice that in the Third challenge run, also gold-standard argument boundaries are provided. Again, the SVM-SPTK system achieves state-of-the-art results in all challenge runs, confirming how the combination of syntactic and lexical information provides a robust model of semantic roles. When gold standard boundaries are provided, i.e. the Third run, the SVM-HMM system achieves the second best results in the challenge. As discussed in [7], this task strictly depends on lexical information and these results confirm that a shallower grammatical information can properly generalize the syntactic behavior of different roles. Notice that SVM-HMM produces the most likely labeling for the entire sentence, so that the implicit re-ranking further contributes to the system robustness. Finally, higher results in the token based measures show that both systems better classify semantic roles with larger spans, i.e. with more syntactical and lexical material.

4 Conclusion

In this work two different statistical learning methods for the FrameNet based SRL are investigated and implemented by two SRL systems that participated to the Evalita *FLaIT* challenge. The SVM-SPTK system is based on the Smoothed Partial Tree Kernel, a convolution kernel that models semantic roles by implicitly combining syntactic and lexical information of annotated examples. This system achieves the state-of-the-art in almost all challenge tasks. The SVM-HMM system represents a very flexible

Table 2. Evalita 2011 - Argument Classification Results

	System	Argument-Based			Token-Based		
		P	R	F1	P	R	F1
First Run	SVM-SPTK	**48.44%**	**52.68%**	**50.47%**	**62.58%**	**64.38%**	**63.47%**
	SVM-HMM	33.10%	33.57%	33.33%	46.77%	53.06%	49.72%
Second Run	SVM-SPTK	51.23%	**55.71%**	**53.38%**	**69.01%**	**70.99%**	**69.99%**
	SVM-HMM	37.52%	37.86%	37.69%	54.63%	61.48%	57.86%
Third Run	SVM-SPTK	70.36%	**70.36%**	**70.36%**	78.35%	**78.35%**	**78.35%**
	SVM-HMM	66.67%	65.36%	66.01%	77.71%	77.46%	77.59%

approach for SRL based on the Markovian formulation of the Structural SVM learning algorithm. Results achieved by this system are lower with respect to the SVM-SPTK, but in line with the other systems in most runs. It is a straightforward result, if considering that SVM-HMM does not rely on a full syntactic parsing of sentences.

Acknowledgements. We want to acknowledge Roberto Basili and Alessandro Moschitti: this work would not exist without their ideas, inspiration and invaluable support.

References

1. Altun, Y., Tsochantaridis, I., Hofmann, T.: Hidden Markov support vector machines. In: Proceedings of the International Conference on Machine Learning (2003)
2. Attardi, G., Rossi, S.D., Simi, M.: The tanl pipeline. In: Proc. of LREC Workshop on WSPP. Valletta, Malta (2010)
3. Baker, C.F., Fillmore, C.J., Lowe, J.B.: The Berkeley FrameNet project. In: Proc. of COLING-ACL. Montreal, Canada (1998)
4. Collins, M., Duffy, N.: Convolution kernels for natural language. In: Proceedings of Neural Information Processing Systems (NIPS). pp. 625–632 (2001)
5. Coppola, B., Moschitti, A., Riccardi, G.: Shallow semantic parsing for spoken language understanding. In: Proceedings of NAACL '09. pp. 85–88. Morristown, NJ, USA (2009)
6. Croce, D., Basili, R.: Structured learning for semantic role labeling. In: AI*IA (2011)
7. Croce, D., Giannone, C., Annesi, P., Basili, R.: Towards open-domain semantic role labeling. In: ACL. pp. 237–246 (2010)
8. Croce, D., Moschitti, A., Basili, R.: Structured lexical similarity via convolution kernels on dependency trees. In: Proceedings of EMNLP. Edinburgh, Scotland, UK. (2011)
9. Deerwester, S.C., Dumais, S.T., Landauer, T.K., Furnas, G.W., Harshman, R.A.: Indexing by latent semantic analysis. JASIS 41(6), 391–407 (1990)
10. Gildea, D., Jurafsky, D.: Automatic Labeling of Semantic Roles. Computational Linguistics 28(3), 245–288 (2002)
11. Joachims, T., Finley, T., Yu, C.N.: Cutting-plane training of structural SVMs. Machine Learning 77(1), 27–59 (2009)
12. Johansson, R., Nugues, P.: The effect of syntactic representation on semantic role labeling. In: Proceedings of COLING. Manchester, UK (August 18-22 2008)
13. Moschitti, A., Pighin, D., Basili, R.: Tree kernels for semantic role labeling. Computational Linguistics 34 (2008)
14. Pado, S., Lapata, M.: Dependency-based construction of semantic space models. Computational Linguistics 33(2) (2007)

15. Pradhan, S., Hacioglu, K., Krugler, V., Ward, W., Martin, J.H., Jurafsky, D.: Support vector learning for semantic argument classification. Machine Learning Journal (2005)
16. Sahlgren, M.: The Word-Space Model. Ph.D. thesis, Stockholm University (2006)
17. Tsochantaridis, I., Joachims, T., Hofmann, T., Altun, Y.: Large margin methods for structured and interdependent output variables. J. Machine Learning Reserach. 6 (December 2005)