Lisbon: Evaluating TurboSemanticParser on Multiple Languages and Out-of-Domain Data

Mariana S. C. Almeida[*†] **André F. T. Martins**[*†]
[*]Priberam Labs, Alameda D. Afonso Henriques, 41, 2º, 1000-123 Lisboa, Portugal
[†]Instituto de Telecomunicações, Instituto Superior Técnico, 1049-001 Lisboa, Portugal
{atm,mla}@priberam.pt

Abstract

As part of the SemEval-2015 shared task on Broad-Coverage Semantic Dependency Parsing, we evaluate the performace of our last year's system (*TurboSemanticParser*) on multiple languages and out-of-domain data. Our system is characterized by a feature-rich linear model, that includes scores for first and second-order dependencies (arcs, siblings, grandparents and co-parents). For decoding this second-order model, we solve a linear relaxation of that problem using alternating directions dual decomposition (AD^3). The experiments have shown that, even though the parser's performance in Chinese and Czech attains around 80% (not too far from English performance), domain shift is a serious issue, suggesting domain adaptation as an interesting avenue for future research.

1 Introduction

The last years have witnessed a continuous progress in statistical multilingual models for syntax, thanks to shared tasks such as CoNLL 2006-7 (Buchholz and Marsi, 2006; Nivre et al., 2007) and, more recently, SPMRL 2013-14 (Seddah et al., 2013; Seddah et al., 2014). As a global trend, we observe that models that incorporate rich global features are typically more accurate, even if pruning is necessary or decoding needs to be approximate (McDonald et al., 2006; Koo and Collins, 2010; Bohnet and Nivre, 2012; Martins et al., 2009, 2013). The same rationale applies to **semantic dependency parsing**, also a structured prediction problem, but where the output variable is a **semantic graph**, rather than a syntactic tree. Indeed, the best performing systems in last year shared task on broad-coverage semantic dependency parsing follow this principle (Oepen et al., 2014). This year, a new challenge was put forth: how to handle multiple languages and out-of-domain data?

Our proposed parser (§2) is essentially the same that we submitted in the previous year to the same SemEval task (Martins and Almeida, 2014), where we scored top in the open challenge and second in the closed track. This year, we report results using new out-of-domain and multilingual data (namely, Czech and Chinese, in addition to English). For the English language, we participated in the closed and open tracks, using as additional resources the syntactic dependency annotations provided by the organizers. For Czech and Chinese, we only addressed the closed track, since no companion data were provided for these languages. We did not participate in the gold track that uses gold-standard syntactic annotations; and we did not address the prediction of predicate senses.

2 Semantic Parser

For this year's shared task, we re-run the semantic parser that we developed last year, which is fully desc1ribed in Martins and Almeida (2014), on the new datasets. Since this parser was designed to be multi-lingual, it was straightforward to apply it to the languages introduced this year (Chinese and Czech), as well as on the out-of-domain data.

We briefly describe our semantic parser (which we dub *TurboSemanticParser* and release as open-source software[1]), and refer the interested reader to

[1] http://labs.priberam.com/Resources/TurboSemanticParser

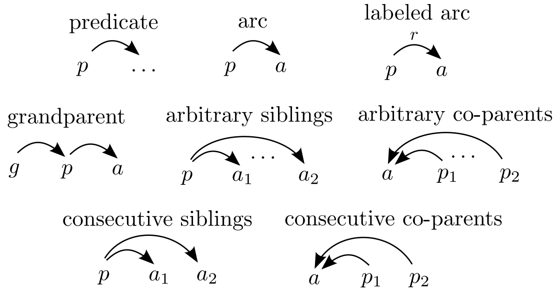

Figure 1: Parts considered by our semantic parser. The top row illustrate the *basic parts*, representing the event that a word is a predicate, or the existence of an arc between a predicate and an argument, eventually labeled with a semantic role. Our *second-order model* looks at some pairs of arcs: arcs bearing a grandparent relationship, arguments of the same predicate, predicates sharing the same argument, and consecutive versions of these two.

Martins and Almeida (2014) for further details.

The parser was built as an extension of a recent dependency parser, *TurboParser* (Martins et al., 2010, 2013), with the goal of performing semantic parsing using any of the three formalisms considered in the shared task (DM, PAS, and PSD). We have followed prior work in semantic role labeling (Toutanova et al., 2005; Johansson and Nugues, 2008; Das et al., 2012; Flanigan et al., 2014), by adding constraints and modeling interactions among arguments within the same frame; however, we went beyond such sibling interactions to consider more complex grandparent and co-parent structures, effectively correlating different predicates. The overall set of parts used by our parser is illustrated in Figure 1; note that by using only a subset of the parts (predicate, arc, labeled arc, and sibling parts), the semantic parser decodes each predicate frame independently from other predicates; it is the co-parent and grandparent parts that have the effect of creating inter-dependence among predicates; we will analyze the effect of these dependencies in the experimental section (§3).

For each part in our model (shown in Figure 1), we computed binary features based on various combination of lexical forms, lemmas, POS tags and syntactic dependency relations of words related to the corresponding predicates and arguments. Most of these features were taken from *TurboParser* (Martins et al., 2013), and others were inspired by the semantic parser of Johansson and Nugues (2008).

To tackle all the parts, we formulate parsing as a global optimization problem and solve a relaxation through AD^3 (Martins et al., 2011), a fast dual decomposition algorithm in which several simple local subproblems are solved iteratively. Through a rich set of features, we arrive at top accuracies at parsing speeds around 1,000 tokens per second. See Martins and Almeida (2014) for details on the model, features and decoding process that were used.

3 Experimental Results

All models were trained by running 10 epochs of max-loss MIRA with $C = 0.01$ (Crammer et al., 2006). The cost function takes into account mismatches between predicted and gold dependencies, with a cost c_P on labeled arcs incorrectly predicted (false positives) and a cost c_R on gold labeled arcs that were missed (false negatives). These values were set through cross-validation in the dev set, yielding $c_P = 0.4$ and $c_R = 0.6$ in all runs, except for the English PSD dataset in the closed track, for which $c_P = 0.3$ and $c_R = 0.7$.

As in the previous work, we speed up decoding by training a probabilistic unlabeled first-order pruner and discarding the arcs whose posterior probability is below 10^{-4}. This allows a significant reduction of the search space with a very small drop in recall.

Table 1 shows our final results in the test set, for a model trained in the train and development partitions. Note that we do not report scores for complete predications, since we did not predict predicate sense. Our system achieved the best final score in 3 out of the 4 tracks for the English language, and for the in-domain closed track in the Czech language. For the remaining 3 tracks we scored relatively close to the best system (Peking), which consists of an ensemble of various methods. For all languages, the runtimes are in par with last year's submission (around 1,000 tokens per second).

As expected, the scores obtained for out-of-domain data are significantly below those obtained with in-domain data. This degradation becomes particularly striking for Czech, with F_1-scores dropping more than 15%. This suggests that domain adaptation (Blitzer et al., 2006; Daumé III, 2007) is an interesting research avenue for future work. In ad-

	Our System							Peking
	UP	UR	UF	LP	LR	LF	Avg. LF	Avg. LF
Eng. DM, closed, id	91.13	87.88	89.48	89.84	86.64	88.21	85.15	**85.33**
Eng. PAS, closed, id	93.12	91.14	92.12	91.87	89.92	90.88		
Eng. PSD, closed, id	89.83	84.81	87.25	78.62	74.23	76.36		
Eng. DM, open, id	91.62	89.46	90.52	90.52	88.39	89.44	**86.23**	–
Eng. PAS, open, id	93.50	91.93	92.71	92.45	90.90	91.67		
Eng. PSD, open, id	91.27	86.16	88.64	79.88	75.41	77.58		
Eng. DM, closed, ood	86.78	80.74	83.65	84.81	78.90	81.75	**81.15**	80.78
Eng. PAS, closed, ood	90.17	86.89	88.50	88.52	85.30	86.88		
Eng. PSD, closed, ood	88.32	80.05	83.98	78.68	71.31	74.82		
Eng. DM, open, ood	87.56	83.52	85.49	85.79	81.84	83.77	**82.53**	–
Eng. PAS, open, ood	90.42	87.91	89.15	88.88	86.41	87.63		
Eng. PSD, open, ood	89.91	81.47	85.48	80.12	72.61	76.18		
Chi. PAS, closed, id	85.56	81.99	83.74	83.81	80.31	82.02	82.02	**83.43**
Cze. PSD, closed, id	90.15	81.55	85.63	83.52	75.54	79.33	**79.33**	78.45
Cze. PSD, closed, ood	86.58	75.97	80.93	67.93	59.61	63.50	63.50	**64.37**

Table 1: Final scores in the test data. For comparison, we show the scores of the Peking system – our best competitor.

dition, as found last year for English, the gap between labeled and unlabeled scores is much higher in the PSD formalism (for English and Czech) then it is for the DM and PAS formalism (for English and Chinese).

Finally, to assess the importance of the second order features, Table 2 reports experiments in the dev-set that progressively add several groups of features. We can see that second order features provide valuable information that improves the final scores. In particular, the higher-order features are extremely useful for Chinese and Czech, where we can observe gains of 1.5–2.0% over a sibling model that factors over predicates.

4 Conclusions

Our system, which is inspired by prior work in syntactic parsing, implements a linear model with second-order features, being able to model interactions between siblings, grandparents and co-parents. We have shown empirically that, for all the three languages, second-order features that correlate multiple predicates have a strong impact in the final scores. However, there is a large drop in accuracy when moving to out-of-domain data.

	UF	LF
Eng. DM, arc-factored	90.19	89.20
Eng. DM, arc-factored, pruned	90.13	89.16
+siblings	90.56	89.53
full system	91.21	90.12
Eng. PAS, arc-factored	92.42	91.52
Eng. PAS, arc-factored, pruned	92.44	91.54
+siblings	92.50	91.53
full system	92.98	91.98
Eng. PSD, arc-factored	87.54	79.69
Eng. PSD, arc-factored, pruned	87.47	79.73
+siblings	88.10	79.87
full system	89.82	80.08
Chi. PAS, arc-factored	81.10	79.49
Chi. PAS, arc-factored, pruned	81.06	79.43
+siblings	81.54	79.70
full system	83.48	81.62
Cze. PSD, arc-factored	84.27	79.77
Cze. PSD, arc-factored, pruned	83.96	79.39
+siblings	85.53	80.44
full system	87.90	81.82

Table 2: Unlabeled/labeled F_1 scores in the dev-set, progressively adding groups of features. English results are for the open track, while Czech and Chinese results are for the closed track.

Acknowledgements

We would like to thank the reviewers for their helpful comments. This work was partially supported by the EU/FEDER programme, QREN/POR Lisboa (Portugal), under the Intelligo project (contract 2012/24803), and by the FCT grants UID/EEA/50008/2013 and PTDC/EEI-SII/2312/2012.

References

John Blitzer, Ryan McDonald, and Fernando Pereira. 2006. Domain adaptation with structural correspondence learning. In *Proc. of Empirical Methods in Natural Language Processing (EMNLP)*, pages 120–128.

Bernd Bohnet and Joakim Nivre. 2012. A transition-based system for joint part-of-speech tagging and labeled non-projective dependency parsing. In *Proc. of the Empirical Methods in Natural Language Processing (EMNLP)*, pages 1455–1465.

Sabine Buchholz and Erwin Marsi. 2006. CoNLL-X shared task on multilingual dependency parsing. In *Proc. Int. Conf. on Natural Language Learning (CoNLL)*, pages 149–164.

Koby Crammer, Ofer Dekel, Joseph Keshet, Shai Shalev-Shwartz, and Yoram Singer. 2006. Online passive-aggressive algorithms. *Journal of Machine Learning Research*, 7:551–585.

Dipanjan Das, André F. T. Martins, and Noah A. Smith. 2012. An exact dual decomposition algorithm for shallow semantic parsing with constraints. In *Proc. of First Joint Conference on Lexical and Computational Semantics (*SEM 2012)*, pages 209–217.

Hall Daumé III. 2007. Frustratingly easy domain adaptation. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 256–263.

Jeffrey Flanigan, Sam Thomson, Jaime Carbonell, Chris Dyer, and Noah A. Smith. 2014. A discriminative graph-based parser for the abstract meaning representation. In *Proc. of the Annual Meeting of the Association for Computational Linguistics*, pages 1426–1436.

Richard Johansson and Pierre Nugues. 2008. Dependency-based syntactic–semantic analysis with PropBank and NomBank. *Int. Conf. on Natural Language Learning (CoNLL)*, pages 183–187.

Terry Koo and Michael Collins. 2010. Efficient third-order dependency parsers. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1–11.

André F. T. Martins and Mariana S. C. Almeida. 2014. Priberam: A turbo semantic parser with second order features. In *Proc. of the 8th Int. Workshop on Semantic Evaluation (SemEval 2014)*, pages 471–476.

André F. T. Martins, Noah A. Smith, and Eric P. Xing. 2009. Concise integer linear programming formulations for dependency parsing. In *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 342–350.

André F. T. Martins, Noah A. Smith, Eric P. Xing, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2010. Turbo parsers: Dependency parsing by approximate variational inference. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, pages 34–44.

André F. T. Martins, Noah A. Smith, Pedro M. Q. Aguiar, and Mário A. T. Figueiredo. 2011. Dual decomposition with many overlapping components. In *Proc. of Empirical Methods for Natural Language Processing (EMNLP)*, pages 238–249.

André F. T. Martins, Miguel B. Almeida, and Noah A. Smith. 2013. Turning on the turbo: Fast third-order non-projective turbo parsers. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 617–622.

Ryan McDonald, Kevin Lerman, and Fernando Pereira. 2006. Multilingual dependency analysis with a two-stage discriminative parser. In *Proc. of Int. Conf. on Natural Language Learning (CoNLL)*, pages 216–220.

Joakim Nivre, Johan Hall, Sandra Kübler, Ryan McDonald, Jens Nilsson, Sebastian Riedel, and Deniz Yuret. 2007. The CoNLL 2007 shared task on dependency parsing. In *Proc. of the CoNLL Shared Task Session of Empirical Methods for Natural Language Processing*, volume 7, pages 915–932.

Stephan Oepen, Marco Kuhlmann, Yusuke Miyao, Daniel Zeman, Dan Flickinger, Jan Hajič, Angelina Ivanova, and Yi Zhang. 2014. Semeval 2014 task 8: broad-coverage semantic dependency parsing. In *Proc. of the 8th Int. Workshop on Semantic Evaluation (SemEval 2014)*, pages 63–72.

Djamé Seddah, Reut Tsarfaty, Sandra Kübler, Marie Candito, Jinho D. Choi, Richárd Farkas, Jennifer Foster, Iakes Goenaga, Koldo Gojenola, Yoav Goldberg, et al. 2013. Overview of the SPMRL 2013 shared task: cross-framework evaluation of parsing morphologically rich languages. In *Proc. of the 4th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2013)*, pages 146–182.

Djamé Seddah, Sandra Kübler, and Reut Tsarfaty. 2014. Introducing the SPMRL 2014 shared task on parsing morphologically-rich languages. In *Proc. of the 5th Workshop on Statistical Parsing of Morphologically Rich Languages (SPMRL 2014)*, pages 23–29.

Kristina Toutanova, Aria Haghighi, and Christopher D. Manning. 2005. Joint learning improves semantic role labeling. In *Proc. of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 589–596.