A Latent Variable Model of Synchronous Syntactic-Semantic Parsing for Multiple Languages

Andrea Gesmundo
Univ Geneva
Dept Computer Sci
Andrea.Gesmundo@unige.ch

James Henderson
Univ Geneva
Dept Computer Sci
James.Henderson@unige.ch

Paola Merlo
Univ Geneva
Dept Linguistics
Paola.Merlo@unige.ch

Ivan Titov*
Univ Illinois at U-C
Dept Computer Sci
titov@uiuc.edu

Abstract

Motivated by the large number of languages (seven) and the short development time (two months) of the 2009 CoNLL shared task, we exploited latent variables to avoid the costly process of hand-crafted feature engineering, allowing the latent variables to induce features from the data. We took a pre-existing generative latent variable model of joint syntactic-semantic dependency parsing, developed for English, and applied it to six new languages with minimal adjustments. The parser's robustness across languages indicates that this parser has a very general feature set. The parser's high performance indicates that its latent variables succeeded in inducing effective features. This system was ranked third overall with a macro averaged F1 score of 82.14%, only 0.5% worse than the best system.

1 Introduction

Recent research in syntax-based statistical machine translation and the recent availability of syntactically annotated corpora for multiple languages (Nivre et al., 2007) has provided a new opportunity for evaluating the cross-linguistic validity of statistical models of syntactic structure. This opportunity has been significantly expanded with the 2009 CoNLL shared task on syntactic and semantic parsing of seven languages (Hajič et al., 2009) belonging to several different language families.

We participate in this task with a generative, history-based model proposed in the CoNLL 2008 shared task for English (Henderson et al., 2008) and further improved to tackle non-planar dependencies (Titov et al., 2009). This model maximises the joint probability of the syntactic and semantic dependencies and thereby enforces that the output structure be globally coherent, but the use of synchronous parsing allows it to maintain separate structures for the syntax and semantics. The probabilistic model is based on Incremental Sigmoid Belief Networks (ISBNs), a recently proposed latent variable model for syntactic structure prediction, which has shown very good performance for both constituency (Titov and Henderson, 2007a) and dependency parsing (Titov and Henderson, 2007b). The use of latent variables enables this architecture to be extended to learning a synchronous parse of syntax and semantics without overly restrictive assumptions about the linking between syntactic and semantic structures.

In this work, we evaluate the ability of this method to generalise across several languages. We take the model as it was developed for English, and apply it directly to all seven languages. The only fine-tuning was to evaluate whether to include one feature type which we had previously found did not help for English, but helped overall. No other feature engineering was done. The use of latent variables to induce features automatically from the data gives our method the adaptability necessary to perform well across all seven languages, and demonstrates the lack of language specificity in the models of Henderson et al. (2008) and Titov et al. (2009).

The main properties of this model, that differentiate it from other approaches, is the use of synchronous syntactic and semantic derivations and the

[0] Authors in alphabetical order.

use of online planarisation of crossing semantic dependencies. This system was ranked third overall with a macro averaged F1 score of 82.14%, only 0.5% worse than the best system.

2 The Synchronous Model

The use of synchronous parsing allows separate structures for syntax and semantics, while still modeling their joint probability. We use the approach to synchronous parsing proposed in Henderson et al. (2008), where we start with two separate derivations specifying each of the two structures, then synchronise these derivations at each word. The individual derivations are based on Nivre's shift-reduce-style parsing algorithm (Nivre et al., 2006), as discussed further below. First we illustrate the high-level structure of the model, discussed in more detail in Henderson et al. (2008).

Let T_d be a syntactic dependency tree with derivation $D_d^1, ..., D_d^{m_d}$, and T_s be a semantic dependency graph with derivation $D_s^1, ..., D_s^{m_s}$. To define derivations for the joint structure T_d, T_s, we divide the two derivations into the chunks between shifting each word onto the stack, $c_d^t = D_d^{b_d^t}, ..., D_d^{e_d^t}$ and $c_s^t = D_s^{b_s^t}, ..., D_s^{e_s^t}$, where $D_d^{b_d^t - 1} = D_s^{b_s^t - 1} = \textit{Shift}_{t-1}$ and $D_d^{e_d^t + 1} = D_s^{e_s^t + 1} = \textit{Shift}_t$. Then the actions of the synchronous derivations consist of quadruples $C^t = (c_d^t, \textit{Switch}, c_s^t, \textit{Shift}_t)$, where *Switch* means switching from syntactic to semantic mode. This gives us the following joint probability model, where n is the number of words in the input.

$$P(T_d, T_s) \quad = \prod_{t=1}^{n} P(C^t | C^1, \ldots, C^{t-1}) \qquad (1)$$

These synchronous derivations $C^1, \ldots, C^n$ only require a single input queue, since the *Shift* actions are synchronised, but they require two separate stacks, one for the syntactic derivation and one for the semantic derivation.

The probability of each synchronous derivation chunk C^t is the product of four factors, related to the syntactic level, the semantic level and the two synchronising steps. The probability of c_d^t is decomposed into one probability for each derivation action D^i, conditioned on its history using the chain rule, and likewise for c_s^t. These probabilities are estimated using the method described in section 3.

	Syn cross	Sem cross	Sem tree	No parse
Cat	0%	0%	61.4%	0%
Chi	0%	28.0%	28.6%	9.5%
Cze	22.4%	16.3%	6.1%	1.8%
Eng	7.6%	43.9%	21.4%	3.9%
Ger	28.1%	1.3%	97.4%	0.0%
Jap	0.9%	38.3%	11.2%	14.4%
Spa	0%	0%	57.1%	0%

Table 1: For each language, percentage of training sentences with crossing arcs in syntax and semantics, with semantic arcs forming a tree, and which were not parsable using the *Swap* action.

One of the main characteristics of our synchronous representation, unlike other synchronous representations of syntax and semantics (Nesson et al., 2008), is that the synchronisation is done on words, rather than on structural components. We take advantage of this freedom and adopt different methods for handling crossing arcs for syntax and for semantics.

While both syntax and semantics are represented as dependency graphs, these graphs differ substantially in their properties. Some statistics which indicate these differences are shown in table 1. For example, English syntactic dependencies form trees, while semantic dependency structures are only trees 21.4% of the time, since in general each structure does not form a connected graph and some nodes may have more than one parent. The syntactic dependency structures for only 7.6% of English sentences contain crossing arcs, while 43.9% of the semantic dependency structures contain crossing arcs. Due to variations both in language characteristics and annotation decisions across corpora, these differences between syntax and semantics vary across the seven languages, but they are consistent enough to motivate the development of new techniques specifically for handling semantic dependency structures. In particular, we use a different method for parsing crossing arcs.

For parsing crossing semantic arcs (i.e. non-planar graphs), we use the approach proposed in Titov et al. (2009), which introduces an action *Swap* that swaps the top two elements on the parser's stack. The *Swap* action allows the parser to reorder words online during the parse. This allows words to be processed in different orders during different

portions of the parse, so some arcs can be specified using one ordering, then other arcs can be specified using another ordering. Titov et al. (2009) found that only using the *Swap* action as a last resort is the best strategy for English (compared to using it preemptively to address future crossing arcs) and we use the same strategy here for all languages.

Syntactic graphs do not use a *Swap* action. We adopt the HEAD method of Nivre and Nilsson (2005) to de-projectivise syntactic dependencies outside of parsing.[1]

3 Features and New Developments

The synchronous derivations described above are modelled with a type of Bayesian Network called an Incremental Sigmoid Belief Network (ISBN) (Titov and Henderson, 2007a). As in Henderson et al. (2008), the ISBN model distinguishes two types of latent states: syntactic states, when syntactic decisions are considered, and semantic states, when semantic decision are considered. Latent states are vectors of binary latent variables, which are conditioned on variables from previous states via a pattern of connecting edges determined by the previous decisions. These latent-to-latent connections are used to engineer soft biases which reflect the relevant domains of locality in the structure being built. For these we used the set of connections proposed in Titov et al. (2009), which includes latent-to-latent connections both from syntax states to semantics states and vice versa. The latent variable vectors are also conditioned on a set of observable features of the derivation history. For these features, we start with the feature set from Titov et al. (2009), which extends the semantic features proposed in Henderson et al. (2008) to allow better handling of the non-planar structures in semantics. Most importantly, all the features previously included for the top of the stack were also included for the word just under the top of the stack. To this set we added one more type of feature, discussed below.

We made some modifications to reflect differences in the task definition between the 2008 and 2009 shared tasks, and experimented with one type of features which had been previously implemented. For the former modifications, the system was adapted to allow the use of the PFEAT and FILLPRED fields in the data, which both resulted in improved accuracy for all the languages. The PFEAT data field (automatically predicted morphological features) was introduced in the system in two ways, as an atomic feature bundle that is predicted when predicting the word, and split into its elementary components when conditioning on a previous word, as was done in Titov and Henderson (2007b). Because the testing data included a specification of which words were annotated as predicates (the FILLPRED data field), we constrained the parser's output so as to be consistent with this specification. For rare predicates, if the predicate was not in the parser's lexicon (extracted from the training set), then a sense was taken from the list of senses reported in the Lexicon and Frame Set resources available for the closed challenge. If this information was not available, then a default sense was constructed based on the automatically predicted lemma (PLEMMA) of the predicate.

We also made use of a previously implemented type of feature that allows the prediction of a semantic link between two words to be conditioned on the syntactic dependency already predicted between the same two words. While this feature had previously not helped for English, it did result in an overall improvement across the languages.

Also, in comparison with previous experiments, the search beam used in the decoding phase was increased from 50 to 80, producing a small improvement in the overall development score.

All development effort took about two person-months, mostly by someone who had no previous experience with the system. Most of this time was spent on the above differences in the task definition between the 2008 and 2009 shared tasks.

4 Results and Discussion

We participated in the joint task of the closed challenge, as described in Hajič et al. (2009). The datasets used in this challenge are described in Taulé et al. (2008) (Catalan and Spanish), Palmer and Xue (2009) (Chinese), Hajič et al. (2006) (Czech), Surdeanu et al. (2008) (English), Burchardt et al. (2006) (German), and Kawahara et al. (2002) (Japanese).

[1] The statistics in Table 1 suggest that, for some languages, swapping might be beneficial for syntax as well.

	Rank	Average	Catalan	Chinese	Czech	English	German	Japanese	Spanish
macro F1	3	82.14	82.66	76.15	83.21	86.03	79.59	84.91	82.43
syntactic acc	1	@85.77	@87.86	76.11	@80.38	88.79	87.29	92.34	@87.64
semantic F1	3	78.42	77.44	76.05	86.02	83.24	71.78	77.23	77.19

Table 2: The three main scores for our system. Rank is within task.

	Rank	Ave	Cze-ood	Eng-ood	Ger-ood
macro F1	3	75.93	@80.70	75.76	71.32
syn Acc	2	78.01	@76.41	80.84	76.77
sem F1	3	73.63	84.99	70.65	65.25

Table 3: Results on out-of-domain for our system. Rank is within task.

The official results on the testing set are shown in tables 2, 3, and 4. The symbol "@" indicates the best result across systems. In table 5, we show our rankings across the different datasets, amongst systems submitted for the same task.

The overall score used to rank systems is the unweighted average of the syntactic labeled accuracy and the semantic labeled F1 measure, across all languages ("macro F1" in table 2). We were ranked third, out of 14 systems. There was only a 0.5% difference between our score and that of the best system, while there was a 1.29% difference between our score and the fourth ranked system. Only considering syntactic accuracy, we had the highest average score of all systems, with the highest individual score for Catalan, Czech, and Spanish. Only considering semantic F1, we were again ranked third. Our results for out-of-domain data (table 3) achieved a similar level of success, although here we were ranked second for average syntactic accuracy. Our precision on semantic arcs was generally much better than our recall (shown in table 4). However, other systems had a similar imbalance, resulting in no change in our third place ranking for semantic precision and for semantic recall. Only when the semantic precision is averaged with syntactic accuracy do we squeeze into second place ("macro Prec").

To get a more detailed picture of the strengths and weaknesses of our system, we computed its rank within each dataset, shown in table 5. Overall, our system is robust across languages, with little fluctuation in ranking for the overall score, including for out-of-domain data. The one noticeable exception to this consistency is the syntactic score for English out-of-domain data. The other ranks for English out-of-domain and English in-domain scores are also on the poor side. These results support our claim that our parser has not undergone much hand-tuning, since it was originally developed for English. It is not currently clear whether this relative difference reflects a English-specific weakness in our system, or that many of the other systems have been fine-tuned for English.

	data	time (min)	macro F1
Czech	25%	5007	73.84
	50%	3699	77.57
	75%	4201	79.10
	100%	6870	80.55
English	25%	1300	79.02
	50%	1899	81.61
	75%	3196	82.41
	100%	3191	83.27

Table 6: Training times and development set accuracies using different percentages of the training data, for Czech and English.

On the higher end of our dataset rankings, we do relatively well on Catalan, Czech, and Spanish. Catalan and Spanish are unique amongst these datasets in that they have no crossing arcs in their semantic structure. Czech seems to have semantic structures which are relatively well handled by our derivations with *Swap*. As indicated above in table 1, only 2% of sentences are unparsable, despite 16% requiring the *Swap* action. However, this argument does not explain why our parser did relatively poorly on German semantic dependencies. Regardless, these observations would suggest that our system is still having trouble with crossing dependencies, despite the introduction of the *Swap* operation, and that our learning method could achieve better performance with an improved treatment of crossing semantic dependencies.

Table 6 shows how accuracies and training times vary with the size of the training dataset, for Czech and English. Training times vary in part because

	Rank	Ave	Cat	Chi	Cze	Eng	Ger	Jap	Spa	Cze-ood	Eng-ood	Ger-ood
semantic Prec	3	81.60	79.08	80.93	87.45	84.92	75.60	83.75	79.44	85.90	72.89	75.19
semantic Rec	3	75.56	75.87	71.73	@84.64	81.63	68.33	71.65	75.05	@84.09	68.55	57.63
macro Prec	2	83.68	83.47	78.52	83.91	86.86	81.44	88.05	83.54	81.16	76.86	@75.98
macro Rec	3	80.66	@81.86	73.92	@82.51	85.21	77.81	81.99	81.35	@80.25	74.70	67.20

Table 4: Semantic precision and recall and macro precision and recall for our system. Rank is within task.

Rank by	Ave	Cat	Chi	Cze	Eng	Ger	Jap	Spa	Ave-ood	Cze-ood	Eng-ood	Ger-ood
macro F1	3	2	3	2	4	4	3	2	3	1	4	3
syntactic Acc	1	1	4	1	3	2	2	1	2	1	7	2
semantic F1	3	2	4	2	4	5	4	2	3	2	4	3

Table 5: Our system's rank within task according to the three main measures, for each dataset.

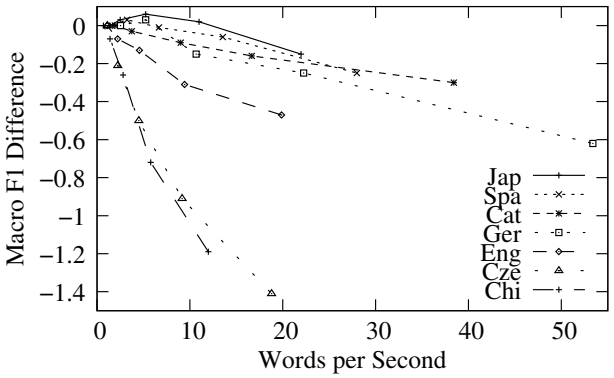

Figure 1: Difference in development set macro F1 as the search beam is decreased from the submitted beam (80) to 40, 20, 10, and 5, plotted against parser speed.

random variations can result in different numbers of training cycles before convergence. Accuracies appear to be roughly log-linear with data size.

Figure 1 shows how the accuracy of the parser degrades as we speed it up by decreasing the search beam used in decoding, for each language. For some languages, a slightly smaller search beam is actually more accurate,[2] but for smaller beams the trade-off of accuracy versus words-per-second is roughly linear. Parsing time per word is also linear in beam width, with a zero intercept.

5 Conclusion

In the joint task of the closed challenge of the CoNLL 2009 shared task (Hajič et al., 2009), we investigated how well a model of syntactic-semantic dependency parsing developed for English would generalise to the other six languages. This model provides a single generative probability of the joint syntactic and semantic dependency structures, but allows separate representations for these two structures by parsing the two structures synchronously. Finding the statistical correlations both between and within these structures is facilitated through the use of latent variables, which induce features automatically from the data, thereby greatly reducing the need for hand-coded feature engineering.

This latent variable model proved very robust across languages, achieving a ranking of between second and fourth on each language, including for out-of-domain data. The extent to which the parser does not rely on hand-crafting is underlined by the fact that its worst ranking is for English, the language for which it was developed (particularly for out-of-domain data). The parser was ranked third overall out of 14 systems, with a macro averaged F1 score of 82.14%, only 0.5% worse than the best system.

Both joint learning and conditioning decisions about semantic dependencies on latent representations of syntactic parsing states were crucial to the success of our model, as was previously demonstrated in Henderson et al. (2008). There, removing this conditioning led to a 3.5% drop in the SRL score. This result seems to contradict the general trend in the CoNLL-2008 shared task, where joint learning had only limited success. The latter fact may be explained by recent theoretical results demonstrating that pipelines can be preferable to joint learning (Roth et al., 2009) when no shared hidden representation is learnt. Our system (Henderson et al., 2008) was the only one which attempted to

[2]This fact suggests that we could have gotten improved results by tailoring the search beam to individual languages.

learn a common hidden representation for this multitask learning problem and also was the only one which achieved significant gain from joint parameter estimation. We believe that learning shared hidden representations for related NLP problems is a very promising direction for further research.

Acknowledgements

We thank Gabriele Musillo and Dan Roth for help and advice. This work was partly funded by Swiss NSF grants 100015-122643 and PBGE22-119276, European Community FP7 grant 216594 (CLASSiC, www.classic-project.org), US NSF grant SoD-HCER-0613885 and DARPA (Bootstrap Learning Program).

References

Aljoscha Burchardt, Katrin Erk, Anette Frank, Andrea Kowalski, Sebastian Padó, and Manfred Pinkal. 2006. The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 5th International Conference on Language Resources and Evaluation (LREC-2006)*, Genoa, Italy.

Jan Hajič, Jarmila Panevová, Eva Hajičová, Petr Sgall, Petr Pajas, Jan Štěpánek, Jiří Havelka, Marie Mikulová, and Zdeněk Žabokrtský. 2006. Prague Dependency Treebank 2.0.

Jan Hajič, Massimiliano Ciaramita, Richard Johansson, Daisuke Kawahara, Maria Antònia Martí, Lluís Màrquez, Adam Meyers, Joakim Nivre, Sebastian Padó, Jan Štěpánek, Pavel Straňák, Mihai Surdeanu, Nianwen Xue, and Yi Zhang. 2009. The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning (CoNLL-2009), June 4-5*, Boulder, Colorado, USA.

James Henderson, Paola Merlo, Gabriele Musillo, and Ivan Titov. 2008. A latent variable model of synchronous parsing for syntactic and semantic dependencies. In *Proceedings of CONLL 2008*, pages 178–182, Manchester, UK.

Daisuke Kawahara, Sadao Kurohashi, and Kôiti Hasida. 2002. Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation (LREC-2002)*, pages 2008–2013, Las Palmas, Canary Islands.

Rebecca Nesson, Giorgio Satta, and Stuart M. Shieber. 2008. Optimal k-arization of synchronous tree-adjoining grammar. In *Proceedings of ACL-08: HLT*, pages 604–612, Columbus, Ohio, June.

Joakim Nivre and Jens Nilsson. 2005. Pseudo-projective dependency parsing. In *Proc. 43rd Meeting of Association for Computational Linguistics*, pages 99–106, Ann Arbor, MI.

Joakim Nivre, Johan Hall, Jens Nilsson, Gulsen Eryigit, and Svetoslav Marinov. 2006. Pseudo-projective dependency parsing with support vector machines. In *Proc. of the Tenth Conference on Computational Natural Language Learning*, pages 221–225, New York, USA.

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 *Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007*, pages 915–932, Prague, Czech Republic, June.

Martha Palmer and Nianwen Xue. 2009. Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, 15(1):143–172.

Dan Roth, Kevin Small, and Ivan Titov. 2009. Sequential learning of classifiers for structured prediction problems. In *AISTATS*, Clearwater, Florida, USA.

Mihai Surdeanu, Richard Johansson, Adam Meyers, Lluís Màrquez, and Joakim Nivre. 2008. The CoNLL-2008 shared task on joint parsing of syntactic and semantic dependencies. In *Proceedings of the 12th Conference on Computational Natural Language Learning (CoNLL-2008)*.

Mariona Taulé, Maria Antònia Martí, and Marta Recasens. 2008. AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation (LREC-2008)*, Marrakesh, Morroco.

Ivan Titov and James Henderson. 2007a. Constituent parsing with Incremental Sigmoid Belief Networks. In *Proceedings of the 45th Annual Meeting of the Association of Computational Linguistics*, pages 632–639, Prague, Czech Republic.

Ivan Titov and James Henderson. 2007b. Fast and robust multilingual dependency parsing with a generative latent variable model. In *Proc. Joint Conf. on Empirical Methods in Natural Language Processing and Computational Natural Language Learning (EMNLP-CoNLL 2007)*, Prague, Czech Republic. (CoNLL Shared Task).

Ivan Titov, James Henderson, Paola Merlo, and Gabriele Musillo. 2009. Online graph planarisation for synchronous parsing of semantic and syntactic dependencies. In *Proc. Twenty-First International Joint Conference on Artificial Intelligence (IJCAI-09)*, Pasadena, California.