Hybrid Multilingual Parsing with HPSG for SRL

Yi Zhang
Language Technology
DFKI GmbH, Germany
yzhang@coli.uni-sb.de

Rui Wang
Computational Linguistics
Saarland University, Germany
rwang@coli.uni-sb.de

Stephan Oepen
Informatics
University of Oslo, Norway
oe@ifi.uio.no

Abstract

In this paper we present our syntactic and semantic dependency parsing system submitted to both the closed and open challenges of the CoNLL 2009 Shared Task. The system extends the system of Zhang, Wang, & Uszkoreit (2008) in the multilingual direction, and achieves 76.49 average macro F1 Score on the closed joint task. Substantial improvements to the open SRL task have been observed that are attributed to the HPSG parses with handcrafted grammars. [†]

1 Introduction

The CoNLL 2009 shared task (Hajič et al., 2009) continues the exploration on learning syntactic and semantic structures based on dependency notations in previous year's shared task. The new addition to this year's shared task is the extension to multiple languages. Being one of the leading competitions in the field, the shared task received submissions from systems built on top of the state-of-the-art data-driven dependency parsing and semantic role labeling systems. Although it was originally designed as a task for machine learning approaches, CoNLL shared tasks also feature an 'open' track since 2008, which encourages the use of extra linguistic resources to further improve the performance. This makes the task a nice testbed for the cross-fertilization of various language processing techniques.

As an example of such work, Zhang et al. (2008) have shown in the past that deep linguistic parsing outputs can be integrated to help improve the performance of the English semantic role labeling task. But several questions remain unanswered. First, the integration only experimented with the semantic role labeling part of the task. It is not clear whether syntactic dependency parsing can also benefit from grammar-based parsing results. Second, the English grammar used to achieve the improvement is one of the largest and most mature hand-crafted linguistic grammars. It is not clear whether similar improvements can be achieved with less developed grammars. More specifically, the lack of coverage of hand-crafted linguistic grammars is a major concern. On the other hand, the CoNLL task is also a good opportunity for the deep processing community to (re-)evaluate their resources and software.

2 System Architecture

The overall system architecture is shown in Figure 1. It is similar to the architecture used by Zhang et al. (2008). Three major components were involved. The HPSG parsing component utilizes several hand-crafted grammars for deep linguistic parsing. The outputs of deep parsings are passed to the syntactic dependency parser and semantic role labeler. The syntactic parsing component is composed of a modified MST parser which accepts HPSG parsing results as extra features. The semantic role labeler is comprised of a pipeline of 4 sub-components (predicate identification is not necessary in this year's task). Comparing to Zhang et al. (2008), this architecture simplified the syntactic component, and puts more focus on the integration of deep parsing outputs. While Zhang et al. (2008) only used seman-

[†] We are indebted to our DELPH-IN colleagues, specifically Peter Adolphs, Francis Bond, Berthold Crysmann, and Montserrat Marimon for numerous hours of support in adapting their grammars and the PET software to parsing the CoNLL data sets. The first author thanks the German Excellence Cluster of Multimodal Computing and Interaction for the support of the work. The second author is funded by the PIRE PhD scholarship program. Participation of the third author in this work was supported by the University of Oslo, as part of its research partnership with the Center for the Study of Language and Information at Stanford University. Our deep parsing experimentation was executed on the TITAN HPC facilities at the University of Oslo.

Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL): Shared Task, pages 31–36,
Boulder, Colorado, June 2009. ©2009 Association for Computational Linguistics

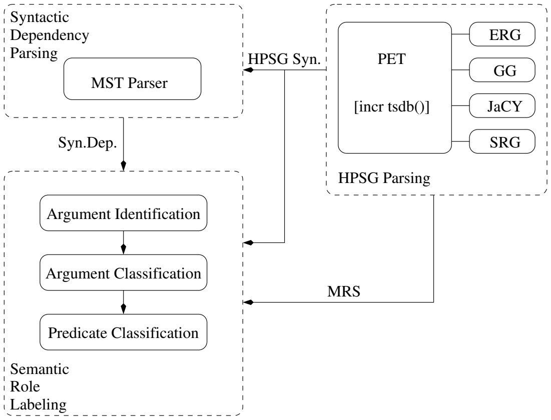

Figure 1: Joint system architecture.

tic features from HPSG parsing in the SRL task, we added extra syntactic features from deep parsing to help both tasks.

3 HPSG Parsing for the CoNLL Data

DELPH-IN (Deep Linguistic Processing with HPSG) is a repository of open-source software and linguistic resources for so-called ‘deep’ grammatical analysis.[1] The grammars are rooted in relatively detailed, hand-coded linguistic knowledge—including lexical argument structure and the linking of syntactic functions to thematic arguments—and are intended as general-purpose resources, applicable to both parsing and generation. Semantics in DELPH-IN is cast in the Minimal Recursion Semantics framework (MRS; Copestake, Flickinger, Pollard, & Sag, 2005), essentially predicate – argument structures with provision for underspecified scopal relations. For the 2009 ‘open’ task, we used the DELPH-IN grammars for English (ERG; Flickinger, 2000), German (GG; Crysmann, 2005), Japanese (JaCY; Siegel & Bender, 2002), and Spanish (SRG; Marimon, Bel, & Seghezzi, 2007). The grammars vary in their stage of development: the ERG comprises some 15 years of continuous development, whereas work on the SRG only started about five years ago, with GG and JaCY ranging somewhere inbetween.

3.1 Overall Setup

We applied the DELPH-IN grammars to the CoNLL data using the PET parser (Callmeier, 2002) running it through the [incr tsdb()] environment (Oepen & Carroll, 2000), for parallelization and distribution. Also, [incr tsdb()] provides facilities for (re-)training the MaxEnt parse selection models that PET uses for disambiguation.

The two main challenges in applying DELPH-IN resources to parsing CoNLL data were (a) mismatches in basic assumptions, specifically tokenization and the inventory of PoS tags provided as part of the input, and (b) the need to adapt the resources for new domains and genres—in particular in terms of parse disambiguation—as the English and Spanish grammars at least had not been previously applied to the corpora used in the CoNLL shared task.

The importance of the first of these two aspects is often underestimated. A detailed computational grammar, inevitably, comes with its own assumptions about tokenization—the ERG, for example, rejects the conventional assumptions underlying the PTB (and derived tools). It opts for an analysis of punctuation akin to affixation (rather than as stand-alone tokens), does not break up contracted negated auxiliaries, and splits hyphenated words like *ill-advised* into two tokens (the hyphen being part of the first component). Thus, a string like *Don’t you!* in the CoNLL data is tokenized as the four-element sequence ⟨*do*, *n’t*, *you*, *!*⟩,[2] whereas the ERG analysis has only two leaf nodes: ⟨*don’t*, *you!*⟩.

Fortunately, the DELPH-IN toolchain recently incorporated a mechanism called *chart mapping* (Adolphs et al., 2008), which allows one to map flexibly from ‘external’ input to grammar-internal assumptions, while keeping track of external token identities and their contributions to the final analysis. The February 2009 release of the ERG already had this machinery in place (with the goal of supporting extant, PTB-trained PoS taggers in pre-processing input to the deep parser), and we found that only a tiny number of additional chart mapping rules was required to ‘fix up’ CoNLL-specific deviations from the PTB tradition. With the help of the original developers, we created new chart mapping configurations for the German and Japanese grammars (with 17 and 16 such accomodation rules, respectively) in a similar spirit. All four DELPH-IN grammars in-

[1] See `http://www.delph-in.net` for background.

[2] Note that the implied analogy to a non-contracted variant is linguistically mis-leading, as **Do not you!* is ungrammatical.

clude an account of unknown words, based on underspecified 'generic' lexical entries that are activated from PoS information.

The Japenese case was interesting, in that the grammar assumes a different pre-processor (ChaSen, rather than Juman), such that not only token boundaries but also PoS tags and morphological features had to be mapped. From our limited experience to date, we found the chart mapping approach adequate in accomodating such discrepancies, and the addition of this extra layer of input processing gave substantial gains in parser coverage (see below). For the Spanish data, on the other hand, we found it impossible to make effective use of the PoS and morphological information in the CoNLL data, due to more fundamental discrepancies (e.g. the treatment of enclitics and multi-word expressions).

3.2 Retraining Disambiguation Models

The ERG includes a domain-specific parse selection model (for tourism instructions); GG only a stub model trained on a handful of test sentences. For use on the CoNLL data, thus, we had to train new parse selections models, better adapted to the shared task corpora. Disambiguation in PET is realized by conditional MaxEnt models (Toutanova, Manning, Flickinger, & Oepen, 2005), usually trained on full HPSG treebanks. Lacking this kind of training material, we utilized the CoNLL dependency information instead, by defining an unlabeled *dependency accuracy* (DA) metric for HPSG analyses, essentially quantifying the degree of overlap in head – dependent relations against the CoNLL annotations.

Calculating DA for HPSG trees is similar to the procedure commonly used for extracting bi-lexical dependencies from phrase structure trees, in a sense even simpler as HPSG analyses fully determine headeness. Taking into account the technical complication of token-level mismatches, our DA metric loosely corresponds to the unlabeled attachment score. To train CoNLL-specific parse selection models, we parsed the development sections in 500-best mode (using the existing models) and then mechanically 'annotated' the HPSG analyses with maximum DA as preferred, all others as dis-preferred. In other words, this procedure constructs a 'binarized' empirical distribution where estimation of log-linear model parameters amounts to adjusting conditional probabilities towards higher DA values.[3]

Grammar	Coverage	Time
ERG	80.4%	10.06 s
GG	28.6%	3.41 s
JaCY	42.7%	2.13 s
SRG	7.5%	0.80 s

Table 1: Performance of the DELPH-IN grammars.

Using the [incr tsdb()] MaxEnt experimentation facilities, we trained new parse selection models for English and German, using the first 16,000 sentences of the English training data and the full German training corpus; seeing that only inputs that (a) parse successfully and (b) have multiple readings, with distinct DA values are relevant to this step, the final models reflect close to 13,000 sentences for English, and a little more than 4,000 items for German. Much like in the SRL component, these experiments are carried out with the TADM software, using ten-fold cross-validation and exact match ranking accuracy (against the binarized training distribution) to optimize estimation hyper-parameters

3.3 Deep Parsing Features

HPSG parsing coverage and average cpu time per input for the four languages with DELPH-IN grammars are summarized in Table 1. The PoS-based unknown word mechanism was active for all grammars but no other robustness measures (which tend to lower the quality of results) were used, i.e. only complete spanning HPSG analyses were accepted. Parse times are for 1-best parsing, using selective unpacking (Zhang, Oepen, & Carroll, 2007).

HPSG parsing outputs are available in several different forms. We investigated two types of structures: syntactic derivations and MRS meaningrepresentations. Representative features were extracted from both structures and selectively used in the statistical syntactic dependency parsing and semantic role labeling modules for the 'open' challenge.

[3] We also experimented with using DA scores directly as empirical probabilities in the training distribution (or some function of DA, to make it fall off more sharply), but none of these methods seemed to further improve parse selection performance.

Deep Semantic Features Similar to Zhang et al. (2008), we extract a set of features from the semantic outputs (MRS) of the HPSG parses. These features represent the basic predicate-argument structure, and provides a simplified semantic view on the target sentence.

Deep Syntactic Dependency Features A HPSG derivation is a tree structure. The internal nodes are labeled with identifiers of grammar rules, and leaves with lexical entries. The derivation tree provides complete information about the actual HPSG analysis, and can be used together with the grammar to reproduce complete feature structure and/or MRS. Given that the shared task adopts dependency representation, we further map the derivation trees into token-token dependencies, labeled by corresponding HPSG rules, by defining a set of head-finding rules for each grammar. This dependency structure is different from the dependencies in CoNLL dataset, and provides an alternative HPSG view on the sentences. We refer to this structure as the dependency backbone (DB) of the HPSG anaylsis. A set of features were extracted from the deep syntactic dependency structures. This includes: i) the POS of the DB parent from the predicate and/or argument; ii) DB label of the argument to its parent (only for AI/AC); iii) labeled path from predicate to argument in DB (only for AI/AC); iv) POSes of the predicate's DB dependents

4 Syntactic Dependency Parsing

For the syntactic dependency parsing, we use the MST Parser (McDonald et al., 2005), which is a graph-based approach. The best parse tree is acquired by searching for a spanning tree which maximizes the score on either a partially or a fully connected graph with all words in the sentence as nodes (Eisner, 1996; McDonald et al., 2005). Based on our experience last year, we use the second order setting of the parser, which includes features over pairs of adjacent edges as well as features over single edges in the graph. For the projective or non-projective setting, we compare the results on the development datasets of different languages. According to the parser performance, we decide to use non-projective parsing for German, Japanese, and Czech, and use projective parsing for the rest.

For the Closed Challenge, we first consider whether to use the morphological features. We find that except for Czech, parser performs better without morphological features on other languages (English and Chinese have no morphological features). As for the other features (i.e. lemma and pos) given by the data sets, we also compare the gold standard features and P-columns. For all languages, the performance decreases in the following order: training with gold standard features and evaluating with the gold standard features, training with P-columns and evaluating with P-columns, training with gold standard features and testing with P-columns. Consequently, in the final submission, we take the second combination.

The goal of the Open Challenge is to see whether using external resources can be helpful for the parsing performance. As we mentioned before, our deep parser gives us both the syntactic analysis of the input sentences using the HPSG formalism and also the semantic analysis using MRS as the representation. However, for the syntactic dependency parsing, we only extract features from the syntactic HPSG analyses and feed them into the MST Parser. Although, when parsing with gold standard lemma and POS features, our open system outperforms the closed system on out-domain tests (for English), when parsing with P-columns there is no substantial improvement observed after using the HPSG features. Therefore, we did not include it in the final submission.

5 Semantic Role Labeling

The semantic role labeling component used in the submitted system is similar to the one described by Zhang et al. (2008). Since predicates are indicated in the data, the predicate identification module is removed from this year's system. Argument identification, argument classification and predicate classification are the three sub-components in the pipeline. All of them are MaxEnt-based classifiers. For parameter estimation, we use the open source TADM system (Malouf, 2002).

The active features used in various steps of SRL are fine tuned separately for different languages using development datasets. The significance of feature types varies across languages and datasets.

		ca	zh	cs	en	de	ja	es
SYN	Closed	82.67	73.63	75.58	87.90	84.57	91.47	82.69
	ood	-	-	71.29	81.50	75.06	-	-
SRL	Closed	67.34	73.20	78.28	77.85	62.95	64.71	67.81
	ood	-	-	77.78	67.07	54.87	-	-
	Open	-	-	-	78.13 (↑0.28)	64.31 (↑1.36)	65.95 (↑1.24)	68.24 (↑0.43)
	ood	-	-	-	68.11 (↑1.04)	58.42 (↑3.55)	-	-

Table 2: Summary of System Performance on Multiple Languages

In the *open* challenge, two groups of extra features from HPSG parsing outputs, as described in Section 3.3, were used on languages for which we have HPSG grammars, that is English, German, Japanese, and Spanish.

6 Result Analysis

The evaluation results of the submitted system are summarized in Table 2. The overall ranking of the system is #7 in the closed challenge, and #2 in the open challenge. While the system achieves mediocre performance, the clear performance difference between the closed and open challenges of the semantic role labeler indicates a substantial gain from the integration of HPSG parsing outputs. The most interesting observation is that even with grammars which only achieve very limited coverage, noticeable SRL improvements are obtained. Confirming the observation of Zhang et al. (2008), the gain with HPSG features is more significant on out-domain tests, this time on German as well.

The training of the syntactic parsing models for all seven languages with MST parser takes about 100 CPU hours with 10 iterations. The dependency parsing takes 6 – 7 CPU hours. The training and testing of the semantic role labeler is much more efficient, thanks to the use of MaxEnt models and the efficient parameter estimation software. The training of all SRL models for 7 languages takes about 3 CPU hours in total. The total time for semantic role labeling on test datasets is less than 1 hour.

Figure 2 shows the learning curve of the syntactic parser and semantic role labeler on the Czech and English datasets. While most of the systems continue to improve when trained on larger datasets, an exception was observed with the Czech dataset on the out-domain test for syntactic accuracy. In most of the cases, with the increase of training data, the out-domain test performance of the syntactic parser and semantic role labeler improves slowly relative to the in-domain test. For the English dataset, the SRL learning curve climbs more quickly than those of syntactic parsers. This is largely due to the fact that the semantic role annotation is sparser than the syntactic dependencies. On the Czech dataset which has dense semantic annotation, this effect is not observed.

7 Conclusion

In this paper, we described our syntactic parsing and semantic role labeling system participated in both closed and open challenge of the (Joint) CoNLL 2009 Shared Task. Four hand-written HPSG grammars of a variety of scale have been applied to parse the datasets, and the outcomes were integrated as features into the semantic role labeler of the system. The results clearly show that the integration of HPSG parsing results in the semantic role labeling task brings substantial performance improvement. The conclusion of Zhang et al. (2008) has been reconfirmed on multiple languages for which we handbuilt HPSG grammars exist, even where grammatical coverage is low. Also, the gain is more significant on out-of-domain tests, indicating that the hybrid system is more robust to cross-domain variation.

References

Adolphs, P., Oepen, S., Callmeier, U., Crysmann, B., Flickinger, D., & Kiefer, B. (2008). Some fine points of hybrid natural language parsing. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakech, Morocco.

Burchardt, A., Erk, K., Frank, A., Kowalski, A., Padó, S., & Pinkal, M. (2006). The SALSA corpus: a German corpus resource for lexical semantics. In *Proceedings of the 4th International Conference on Language Resources and Evaluation.* Genoa, Italy.

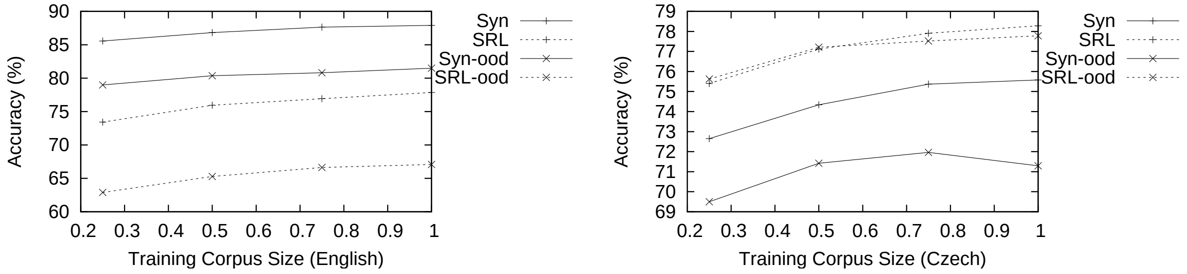

Figure 2: Learning curves of syntactic dependency parser and semantic role labeler on Czech and English datasets

Callmeier, U. (2002). Preprocessing and encoding techniques in PET. In S. Oepen, D. Flickinger, J. Tsujii, & H. Uszkoreit (Eds.), *Collaborative language engineering. A case study in efficient grammar-based processing.* Stanford, CA: CSLI Publications.

Copestake, A., Flickinger, D., Pollard, C., & Sag, I. A. (2005). Minimal Recursion Semantics. An introduction. *Journal of Research on Language and Computation*, *3*(4), 281 – 332.

Crysmann, B. (2005). Relative clause extraposition in German. An efficient and portable implementation. *Research on Language and Computation*, *3*(1), 61 – 82.

Flickinger, D. (2000). On building a more efficient grammar by exploiting types. *Natural Language Engineering*, *6 (1)*, 15 – 28.

Hajič, J., Ciaramita, M., Johansson, R., Kawahara, D., Martí, M. A., Màrquez, L., Meyers, A., Nivre, J., Padó, S., Štěpánek, J., Straňák, P., Surdeanu, M., Xue, N., & Zhang, Y. (2009). The CoNLL-2009 shared task: Syntactic and semantic dependencies in multiple languages. In *Proceedings of the 13th Conference on Computational Natural Language Learning.* Boulder, CO, USA.

Hajič, J., Panevová, J., Hajičová, E., Sgall, P., Pajas, P., Štěpánek, J., Havelka, J., Mikulová, M., & Žabokrtský, Z. (2006). *Prague Dependency Treebank 2.0* (Nos. Cat. No. LDC2006T01, ISBN 1-58563-370-4). Philadelphia, PA, USA: Linguistic Data Consortium.

Kawahara, D., Kurohashi, S., & Hasida, K. (2002). Construction of a Japanese relevance-tagged corpus. In *Proceedings of the 3rd International Conference on Language Resources and Evaluation* (pp. 2008–2013). Las Palmas, Canary Islands.

Malouf, R. (2002). A comparison of algorithms for maximum entropy parameter estimation. In *Proceedings of the 6th conferencde on natural language learning (CoNLL 2002)* (pp. 49–55). Taipei, Taiwan.

Marimon, M., Bel, N., & Seghezzi, N. (2007). Test suite construction for a Spanish grammar. In T. H. King & E. M. Bender (Eds.), *Proceedings of the Grammar Engineering Across Frameworks workshop* (p. 250-264). Stanford, CA: CSLI Publications.

Oepen, S., & Carroll, J. (2000). Performance profiling for parser engineering. *Natural Language Engineering*, *6 (1)*, 81 – 97.

Palmer, M., Kingsbury, P., & Gildea, D. (2005). The Proposition Bank: An Annotated Corpus of Semantic Roles. *Computational Linguistics*, *31*(1), 71–106.

Palmer, M., & Xue, N. (2009). Adding semantic roles to the Chinese Treebank. *Natural Language Engineering*, *15*(1), 143–172.

Siegel, M., & Bender, E. M. (2002). Efficient deep processing of Japanese. In *Proceedings of the 3rd workshop on asian language resources and international standardization at the 19th international conference on computational linguistics.* Taipei, Taiwan.

Surdeanu, M., Johansson, R., Meyers, A., Màrquez, L., & Nivre, J. (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.* Manchester, UK.

Taulé, M., Martí, M. A., & Recasens, M. (2008). AnCora: Multilevel Annotated Corpora for Catalan and Spanish. In *Proceedings of the 6th International Conference on Language Resources and Evaluation.* Marrakesh, Morroco.

Toutanova, K., Manning, C. D., Flickinger, D., & Oepen, S. (2005). Stochastic HPSG parse selection using the Redwoods corpus. *Journal of Research on Language and Computation*, *3*(1), 83 – 105.

Zhang, Y., Oepen, S., & Carroll, J. (2007). Efficiency in unification-based n-best parsing. In *Proceedings of the 10th International Conference on Parsing Technologies* (pp. 48 – 59). Prague, Czech Republic.

Zhang, Y., Wang, R., & Uszkoreit, H. (2008). Hybrid Learning of Dependency Structures from Heterogeneous Linguistic Resources. In *Proceedings of the Twelfth Conference on Computational Natural Language Learning (CoNLL 2008)* (pp. 198–202). Manchester, UK.