Dependency Parsing by Inference over High-recall Dependency Predictions

Sander Canisius, Toine Bogers,
Antal van den Bosch, Jeroen Geertzen
ILK / Computational Linguistics and AI
Tilburg University, P.O. Box 90153,
NL-5000 LE Tilburg, The Netherlands
{S.V.M.Canisius,A.M.Bogers,
Antal.vdnBosch,J.Geertzen}@uvt.nl

Erik Tjong Kim Sang
Informatics Institute
University of Amsterdam, Kruislaan 403
NL-1098 SJ Amsterdam, The Netherlands
erikt@science.uva.nl

1 Introduction

As more and more treebanks, i.e. syntactically-annotated corpora, become available for a wide variety of languages, machine learning approaches to parsing gain interest as a means of developing parsers without having to repeat such labor-intensive and language-specific activities as grammar development for each new language. In this paper, we describe two different machine learning approaches to the CoNLL-X shared task on multi-lingual dependency parsing. First, we introduce a number of baselines that generate left-branching, right-branching or more complex trees. Next, we present two systems that were submitted to the shared task: 1) an approach that directly predicts all dependency relations in a single run over the input sentence, and 2) a cascade of phrase recognizers. We find that the first approach performs best and conclude with a detailed error analysis of its output for two of the thirteen languages in the task, Dutch and Spanish.

2 Baseline approaches

We developed four different baseline approaches for assigning labeled dependency structures to sentences. All of the baselines produce projective structures. We describe the heuristics for constructing the trees and labeling the nodes separately. The following four baseline structures were constructed:

Binary right-branching trees The first baseline produces right-branching binary trees. The first token in the sentence is marked as the top node with HEAD `0` and DEPREL `ROOT`. For the rest of the tree, token $n-1$ serves as the HEAD of token n. Figure 1 shows an example of the kind of tree this baseline produces.

Binary left-branching trees The binary left-branching baseline mirrors the previous baseline. The penultimate token in the sentence is marked as the top node with HEAD `0` and DEPREL `ROOT` since punctuation tokens can never serve as `ROOT`. For the rest of the tree, the HEAD of token n is token $n+1$. Figure 2 shows an example of a tree produced by this baseline.

Inward-branching trees In this approach, the first identified verb[1] is marked as the `ROOT` node. The part of the sentence to the left of the `ROOT` is left-branching, the part to the right of the `ROOT` is right-branching. Figure 3 shows an example of a tree produced by this third baseline.

Nearest neighbor-branching trees In our most complex baseline, the first verb is is marked as the `ROOT` node and the other verbs (with DEPREL `VC`) point to the closest preceding verb. The other tokens point in the direction of their nearest neighboring verb, i.e. the two tokens at a distance of 1 from a verb have that verb as their HEAD, the two tokens at a distance of 2 have the tokens at a distance of 1 as their head, and so on until another verb is a closer neighbor. Figure 4 clarifies this kind of dependency structure in an example tree.

Labeling is done using a three-fold back-off strategy. From the training set, we collect the most frequent DEPREL tag for each head-dependency

[1] We consider a token a verb if its CPOSTAG tag starts with a 'V'.

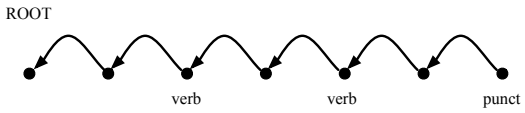

Figure 1: Binary right-branching tree for an example sentence with two verbs.

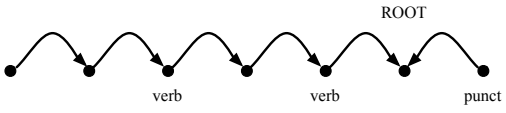

Figure 2: Binary left-branching tree for the example sentence.

FORM pair, the most frequent DEPREL tag for each FORM, and the most frequent DEPREL tag in the entire training set. The unlabeled tokens are labeled in this order: first, we look up if FORM pair of a token and its head was present in the training data. If not, then we assign it the most frequent DEPREL tag in the training data for that specific token FORM. If all else fails we label the token with the most frequent DEPREL tag in the entire training set (excluding `punct` and `ROOT`).

language	baseline	unlabeled	labeled
Arabic	left	58.82	39.72
Bulgarian	inward	41.29	29.50
Chinese	NN	37.18	25.35
Czech	NN	34.70	22.28
Danish	inward	50.22	36.83
Dutch	NN	34.07	26.87
German	NN	33.71	26.42
Japanese	right	67.18	64.22
Portuguese	right	25.67	22.32
Slovene	right	24.12	19.42
Spanish	inward	32.98	27.47
Swedish	NN	34.30	21.47
Turkish	right	49.03	31.85

Table 1: The labeled and unlabeled scores for the best performing baseline for each language (NN = nearest neighbor-branching).

The best baseline performance (labeled and unlabeled scores) for each language is listed in Table 1. There was no single baseline that outperformed

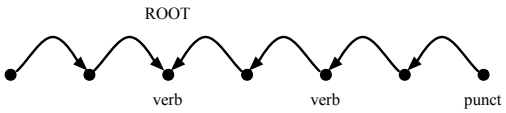

Figure 3: Binary inward-branching tree for the example sentence.

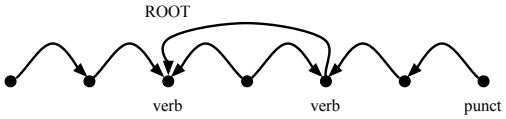

Figure 4: Nearest neighbor-branching tree for the example sentence.

the others on all languages. The nearest neighbor baseline outperformed the other baselines on five of the thirteen languages. The right-branching and inward-branching baselines were optimal on four and three languages respectively. The only language where the left-branching trees provide the best performance is Arabic, due to the fact that Arabic sentences are written and read from right to left.

3 Parsing by inference over high-recall dependency predictions

In our approach to dependency parsing, a machine learning classifier is trained to predict (directed) dependency relations between a head and a dependent. For each token in a sentence, instances are generated where this token is a potential dependent of each of the other tokens in the sentence[2]. The label that is predicted for each classification case serves two different purposes at once: 1) it signals whether the token is a dependent of the designated head token, and 2) if the instance does in fact correspond to a dependency relation in the resulting parse of the input sentence, it specifies the type of this relation, as well.

By considering each potential dependency relation as a separate classification case, inconsistent trees may result, however. For example, one token

[2]To prevent explosion of the number of classification cases to be considered for a sentence, we restrict the maximum distance between a token and its potential head. For each language, we selected this distance so that, on the training data, 95% of the dependency relations is covered.

may be predicted to be a dependent of more than one head. In order to recover a valid dependency tree from the separate pair-wise dependency predictions, a simple inference procedure is performed. Consider an input sentence consisting of n tokens and one of these tokens for which the dependency relation is to be predicted. For this token, a number of classification cases have been processed, each of them indicating whether and if so how the token is related to one of the other tokens in the sentence. Some of these predictions may be negative, i.e. the token is not a dependent of a certain other token in the sentence, others may be positive, suggesting the token is a dependent of some other token. If all classifications are negative, the token is assumed to have no head, and consequently no dependency relation is added to the tree. If one of the classifications is non-negative, suggesting a dependency relation between this token as a dependent and some other token as a head, this dependency relation is added to the tree. Finally, there is the case in which more than one prediction is non-negative. By definition, at most one of these predictions can be correct; therefore, only one dependency relation should be added to the tree. To select the most-likely candidate from the predicted dependency relations, the candidates are ranked according to the classifier confidence of the base classifier that predicted them, and the highest-ranked candidate is selected for insertion into the tree.

We implement our base classifier using a memory-based learner as implemented by TiMBL (Daelemans et al., 2004). The instances processed by this classifier correspond to a rather simple description of the head-dependent pair to be classified. For both the potential head and dependent, there are features encoding a 2-1-2 window of words and part-of-speech tags; in addition, there are two spatial features: a relative position feature, encoding whether the dependent is located to the left or to the right of its potential head, and a distance feature that simply lists the number of tokens between the dependent and its head. The parameters of the memory-based learner have been optimized for accuracy separately for each language by internally sampling training and test data from the training set.

The base classifier in our parser is faced with a classification task with a highly skewed class distribution, i.e. instances that correspond to a dependency relation, are largely outnumbered by those that do not. In practice, such a huge number of negative instances usually results in classifiers that tend to predict fairly conservatively, resulting in high precision, but low recall. In the approach introduced above, however, it is better to have high recall, even at the cost of precision, than to have high precision at the cost of recall. A missed relation by the base classifier can never be recovered by the inference procedure; however, due the constraint that each token can only be a dependent of one head, excessive prediction of dependency relations can still be corrected by the inference procedure. An effective method for increasing the recall of a classifier is down-sampling of the training data. In down-sampling, instances belonging to the majority class (in this case the negative class) are removed from the training data, so as to obtain a more balanced distribution of negative and non-negative instances.

Figure 5 shows the effect of systematically removing an increasingly larger part of the negative instances from the training data. First of all, the figure confirms that down-sampling helps to improve recall, though it does so at the cost of precision. More importantly however, it also illustrates that this improved recall is beneficial for the performance of the dependency parser. The shape of the performance curve of the dependency parser closely follows that of the recall. Remarkably, parsing performance continues to improve with increasingly stronger down-sampling, even though precision drops considerably as a result of this. This shows that the confidence of the classifier for a certain prediction is a reliable indication of the quality of that prediction. Only when the number of negative training instances is reduced to equal the number of positive instances, the performance of the parser is negatively affected. Based on a quick evaluation of various down-sampling ratios on a 90%-10% train-test split of the Dutch training data, we decided to down-sample the training data for all languages with a ratio of two negative instances for each positive one.

Table 2 lists the unlabeled and labeled attachment scores of the resulting system for all thirteen languages.

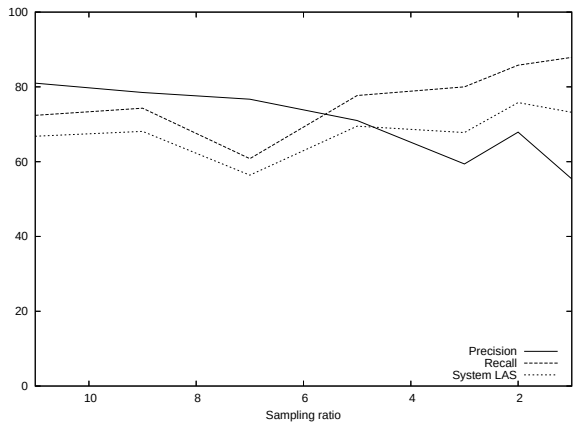

Figure 5: The effect of down-sampling on precision and recall of the base classifier, and on labeled accuracy of the dependency parser. The x-axis refers to the number of negative instances for each positive instance in the training data. Training and testing was performed on a 90%-10% split of the Dutch training data.

language	unlabeled	labeled
Arabic	74.59	57.64
Bulgarian	82.51	78.74
Chinese	82.86	78.37
Czech	72.88	60.92
Danish	82.93	77.90
Dutch	77.79	74.59
German	80.01	77.56
Japanese	89.67	87.41
Portuguese	85.61	77.42
Slovene	74.02	59.19
Spanish	71.33	68.32
Swedish	85.08	79.15
Turkish	64.19	51.07

Table 2: Results of the submitted system.

4 Cascaded dependency parsing

One of the alternative strategies explored by us was modeling the parsing process as a cascade pair of basic learners. In the first phase, each learner had to predict whether each word was a daughter of the preceding or the next word. Dependent words were removed and the remaining words were sent to the learners for further rounds of processing until all words but one had been assigned a parent. Whenever crossing links prevented further assignments of parents to words, the learner was expected to ignore the remaining word requiring the longest dependency link. When the first phase was finished another learner assigned labels to pairs of words present in dependency links.

This approach was based on our earlier work (Tjong Kim Sang, 2002). Because of time constraints we were unable to evaluate different learner configurations. We used two different training files for the first phase: one for predicting the dependency links between adjacent words and one for predicting all other links. As learner, we used TiMBL with its default parameters. We evaluated different feature sets and ended up with using words, lemmas, POS tags and an extra pair of features with the POS tags of the recent attachments to the focus word. With this configuration, this approach achieved a labeled score of 62.99 on our Dutch test data compared to 74.59 of the constraint satisfaction approach.

5 Error analysis

We examined the system output for two languages in more detail: Dutch and Spanish.

5.1 Dutch

With a labeled attachment score of 74.59 and an unlabeled attachment score of 77.79, our submitted Dutch system performs somewhat above the average over all submitted systems (labeled 70.73, unlabeled 75.07). We review the most notable errors made by our system.

From a part-of-speech (CPOSTAG) perspective, a remarkable relative amount of head and dependency errors are made on **conjunctions**. A likely explanation is that the tag "Conj" applies to both coordinating and subordinating conjunctions. A fine-grained part-of-speech tag would likely solve some of these errors.

Left- and right-directed attachment to heads is roughly equally successful. Many errors are made on relations attaching to ROOT; the system appears to be overgenerating attachments to ROOT, mostly in cases when it should have generated rightward attachments. Unsurprisingly, the more distant the head is, the less accurate the attachment; especially recall suffers at distances of three and more tokens.

The most frequent attachment error is generating a ROOT attachment instead of a "mod" (modifier) relation, often occurring at the start of a sentence. Many errors relate to ambiguous adverbs such as *bovendien* (moreover), *tenslotte* (after all), and *zo* (thus), which tend to occur rather frequently at the beginning of sentences in the test set, but less so in the training set. The test set appears to consist largely of formal journalistic texts which typically tend to use these marked rhetorical words in sentence-initial position, while the training set is a more mixed set of texts from different genres and individual sentences.

5.2 Spanish

The Spanish test data set was the only data set in which the alternative cascaded approach (72.15) outperformed our main constraint satisfaction approach (68.32). A detailed comparison of the output files of the two system has not revealed a unique cause for the performance difference. We have examined a possible sentence length influence, scores obtained for different POS tags and scores related to links spanning a different number of words. As expected the cascaded approach, without a link length limit, outperforms the constraint satisfaction approach, with a 15-word span maximum, when predicting links of longer lengths. But the first method also outperformed the second for shorter spans.

Acknowledgements

This research is funded by NWO, the Netherlands Organization for Scientific Research under the IMIX programme, and the Dutch Ministry for Economic Affairs' IOP-MMI programme.

References

A. Abeillé, editor. 2003. *Treebanks: Building and Using Parsed Corpora*, volume 20 of *Text, Speech and Language Technology*. Kluwer Academic Publishers, Dordrecht.

S. Afonso, E. Bick, R. Haber, and D. Santos. 2002. "Floresta sintá(c)tica": a treebank for Portuguese. In *Proc. of the Third Intern. Conf. on Language Resources and Evaluation (LREC)*, pages 1698–1703.

N. B. Atalay, K. Oflazer, and B. Say. 2003. The annotation process in the Turkish treebank. In *Proc. of the 4th Intern. Workshop on Linguistically Interpreteted Corpora (LINC)*.

A. Böhmová, J. Hajič, E. Hajičová, and B. Hladká. 2003. The PDT: a 3-level annotation scenario. In Abeillé (Abeillé, 2003), chapter 7.

S. Brants, S. Dipper, S. Hansen, W. Lezius, and G. Smith. 2002. The TIGER treebank. In *Proc. of the First Workshop on Treebanks and Linguistic Theories (TLT)*.

K. Chen, C. Luo, M. Chang, F. Chen, C. Chen, C. Huang, and Z. Gao. 2003. Sinica treebank: Design criteria, representational issues and implementation. In Abeillé (Abeillé, 2003), chapter 13, pages 231–248.

M. Civit Torruella and M^{a} A. Martí Antonín. 2002. Design principles for a Spanish treebank. In *Proc. of the First Workshop on Treebanks and Linguistic Theories (TLT)*.

W. Daelemans, J. Zavrel, K. Van der Sloot, and A. Van den Bosch. 2004. TiMBL: Tilburg memory based learner, version 5.1, reference guide. Technical Report ILK 04-02, ILK Research Group, Tilburg University.

S. Džeroski, T. Erjavec, N. Ledinek, P. Pajas, Z. Žabokrtsky, and A. Žele. 2006. Towards a Slovene dependency treebank. In *Proc. of the Fifth Intern. Conf. on Language Resources and Evaluation (LREC)*.

J. Hajič, O. Smrž, P. Zemánek, J. Šnaidauf, and E. Beška. 2004. Prague Arabic dependency treebank: Development in data and tools. In *Proc. of the NEMLAR Intern. Conf. on Arabic Language Resources and Tools*, pages 110–117.

Y. Kawata and J. Bartels. 2000. Stylebook for the Japanese treebank in VERBMOBIL. Verbmobil-Report 240, Seminar für Sprachwissenschaft, Universität Tübingen.

M. T. Kromann. 2003. The Danish dependency treebank and the underlying linguistic theory. In *Proc. of the Second Workshop on Treebanks and Linguistic Theories (TLT)*.

J. Nilsson, J. Hall, and J. Nivre. 2005. MAMBA meets TIGER: Reconstructing a Swedish treebank from antiquity. In *Proc. of the NODALIDA Special Session on Treebanks*.

K. Oflazer, B. Say, D. Zeynep Hakkani-Tür, and G. Tür. 2003. Building a Turkish treebank. In Abeillé (Abeillé, 2003), chapter 15.

K. Simov and P. Osenova. 2003. Practical annotation scheme for an HPSG treebank of Bulgarian. In *Proc. of the 4th Intern. Workshop on Linguistically Interpreteted Corpora (LINC)*, pages 17–24.

K. Simov, P. Osenova, A. Simov, and M. Kouylekov. 2005. Design and implementation of the Bulgarian HPSG-based treebank. In *Journal of Research on Language and Computation – Special Issue*, pages 495–522. Kluwer Academic Publishers.

Erik F. Tjong Kim Sang. 2002. Memory-based shallow parsing. *Journal of Machine Learning Research*, 2(Mar):559–594.

L. van der Beek, G. Bouma, R. Malouf, and G. van Noord. 2002. The Alpino dependency treebank. In *Computational Linguistics in the Netherlands (CLIN)*.