

Correlating Natural Language Parser Performance with Statistical Measures of the Text

Yi Zhang and Rui Wang

¹ LT-Lab, German Research Center for Artificial Intelligence and Computational Linguistics, Saarland University

yzhang@coli.uni-sb.de

<http://www.coli.uni-saarland.de/~yzhang>

² Computational Linguistics, Saarland University

rwang@coli.uni-sb.de

<http://www.coli.uni-saarland.de/~rwang>

Abstract. Natural language parsing, as one of the central tasks in natural language processing, is widely used in many AI fields. In this paper, we address an issue of parser performance evaluation, particularly its variation across datasets. We propose three simple statistical measures to characterize the datasets and also evaluate their correlation to the parser performance. The results clearly show that different parsers have different performance variation and sensitivity against these measures. The method can be used to guide the choice of natural language parsers for new domain applications, as well as systematic combination for better parsing accuracy.

1 Introduction

Natural language parsing is not only one of the central tasks in the field of natural language processing, but also widely used in many other AI areas, e.g. human-computer interaction, robotics, etc. While many parsing systems achieve comparably high accuracy from application perspective [1], the robustness of parser performance remains as one of the major problems which is not only unresolved, but also less acknowledged and largely overlooked. The capability of most statistical parsing systems to produce a parse for almost any input does not entail a consistent and robust parser performance on different inputs.

For example, in robotics, the input of the parsers usually comes from an automatic speech recognition (ASR) system, which is error-prone and much worse than the human listeners [2]. More seriously, as one of the earliest components, in many applications, the unsatisfying outputs of the parsers will be propagated and the errors will be amplified through the common pipeline architecture. For example, in a popular task in Bioinformatics, protein-protein interaction extraction, [1] have shown correlation between the parse accuracy and the extraction accuracy.

Through the past decade, there has been development of numerous parsing systems with different approaches using various representations, many of which are available as open source softwares and ready for use off-the-shelf. However, it is a common knowledge now that treebank-trained parsers usually perform much worse when applied onto texts of different genres from the training set. Although it is the nature of human languages to be diverse, the variation of parser performance does not always correspond to the difficulty of the text for human readers.

More recently, the problem has been studied as the task of parser domain adaptation. For instance, [3] generalized the previous approaches using a maximum a posteriori (MAP) framework and proposed both supervised and unsupervised adaptations of statistical parsers. [4] and [5] have shown that out-domain parser performance can be improved with self-training on a large amount of extra unlabeled data. The CoNLL shared task 2007 [6] has a dedicated challenge to adapt parsers trained on the newspaper texts to process chemistry and child language texts. [7] and [8] port their parsers for biomedical texts, while [9] adapts her parser for various Wikipedia biographical texts. In all, most of the studies take a liberal definition of “domain”: the term is used to dub almost any dataset, either slightly or remotely different from the training set. Also, while substantial performance improvements have been reported for different parsing systems, it is not clear whether such methods are equally effective for other parsers (when intuition usually suggests the opposite).

In this paper, we present a series of experiments which correlate the performance of several state-of-the-art parsing systems (Section 3) to three very simple statistical measures of the datasets (Section 2). The result clearly shows that even for a group of datasets of similar genres, parser performance varies substantially. Furthermore, performances of different parsers are sensitive to different statistical measures (Section 4).

2 Statistical Measures for Datasets

There has been a rich literature in text classification on statistical measures that can be used to categorize documents. However, here we are not interested in differentiating the semantic contents of the texts, but in those basic measures which can be potentially correlated with the parser performance. As another related work, [10] focused on annotation differences between datasets and attributed many errors to that; while in this paper, we concern more about the basic statistical distribution of the texts itself within the datasets, without considering the syntactic annotations. When the parsers are tested on datasets with compatible and consistent annotations to the training set, the performance correlation to these measures on unannotated texts reflect the characteristics of the parsers, independent from the annotation scheme adopted.

The following three measures are used for the experiments reported in the this paper.

Average Sentence Length (ASL) is the most simple measure which can be easily calculated for any given dataset without consulting extra resources. Common intuition is that the performance of the parser is relatively worse on longer sentences than shorter ones.

Unknown Word Ratio (UWR) is calculated by counting instances of the unseen words in the training set and deviding it by the total number of word instances in the target dataset.

Unknown Part-of-Speech Trigram Ratio (UPR) is calculated by counting the instances of unseen POS trigram patterns in the training set and deviding it by the total number of trigrams in the target dataset. We also add speical sentence initial and final symbols into the POS patterns, so as to denote the rough position of the trigram in the sentence.

It should be noted that these simple measures are given here as examples to show the performance variation among different parsers. Adding further statistical measures is straightforward, and will be experiment in our future work.

3 Parser Performance Evaluation

Parser evaluation has turned out to be a difficult task on its own, especially in the case of cross-framework parser comparison. Fortunately, in this study we are not interested in the absolute scores of the parser performance, and instead, only the variation of the performance among different datasets. For this reason, we select representative evaluation metrics for each individual parsing system,

We select the following group of representative parsing systems in our experiment. All these parsers are freely available on-line. For those parsers where training is required, we use the Wall Street Journal (WSJ) section 2-21 of the Penn Treebank (PTB) as the training set. This includes both the phrase-structure trees in the original PTB annotation, and the automatically converted word-word dependency representation.

*Dan Bikel's Parser (DBP)*¹ [11] is an open source multilingual parsing engine. We use it to emulate Collins parsing model II [12].

*Stanford Parser (SP)*² [13] is used as an unlexicalized probabilistic Context-Free Grammar (PCFG) parser. It utilizes important features commonly expressed by closed class words, but no use is made of lexical class words, to provide either monolexical or bilexical probabilities.

*MST Parser (MST)*³ [14] is a graph-based dependency parser where the best parse tree is acquired by searching for a spanning tree which maximize the score on an either partially or fully connected dependency graph.

¹ <http://www.cis.upenn.edu/~dbikel/software.html>

² <http://nlp.stanford.edu/software/lex-parser.shtml>

³ <http://sourceforge.net/projects/mstparser/>

Malt Parser (MALT)⁴ [15] follows a transition-based approach, where parsing is done through a series of actions deterministically predicted by an oracle.

ERG+PET (ERG)⁵ is the combination of a large scale hand-crafted HPSG grammar for English [16], and a language independent unification-based efficient parser [17]. The statistical disambiguation model is trained with part of the WSJ data.

Although these parsers adopt different representations, making cross-framework parser evaluation difficult, here we are only interested in the relative performance variation of individual parsers. Hence, different evaluation metrics are used for different parsers. For constituent-based PCFG parsers (DBP and SP), we evaluate the labeled bracketing F-score; and for dependency parsers (MST and MALT), we evaluate the labeled attachment score. Since there is no gold HPSG treebank for our target test set, we map the HPSG parser output into a word dependency representation, and evaluate the unlabeled attachment score against our gold dependency representation.

4 Experiment Results

4.1 Datasets

As test datasets, we use the Brown Sections of the Penn Treebank. The dataset contains in total 24243 sentences with an average sentence length of 18.9 tokens. The dataset has a mixture of genres, ranging from fictions to biographies and memoires, arranged into separate sections. We further split these sections into 97 smaller datasets, and each one contains continuous texts from two adjacent files in the original corpus. The average size of 250 sentences per dataset will provide reliable parser evaluation results.

4.2 Results

All five parsers were evaluated on the 97 datasets. The performance variation is very substantial for all these parsers, although it is hard to compare on concrete numbers due to the different evaluation metrics. Figure 1 shows the correlation between parser performance and the average sentence length (ASL), unknown word ratio (UWR), and unknown POS trigram ratio (UPR)⁶. It is not surprising to observe that all the parsers' performances have negative correlations to these three measures, with some of which more significant than the others. We should note that the correlation reflects the noisiness and direction of the relation between the statistical measure and the parser performance, but not the slope of that relationship.

⁴ <http://w3.msi.vxu.se/~jha/maltparser/>

⁵ <http://lingo.stanford.edu/erg/>

⁶ All these evaluation results and parser outputs will be available online.

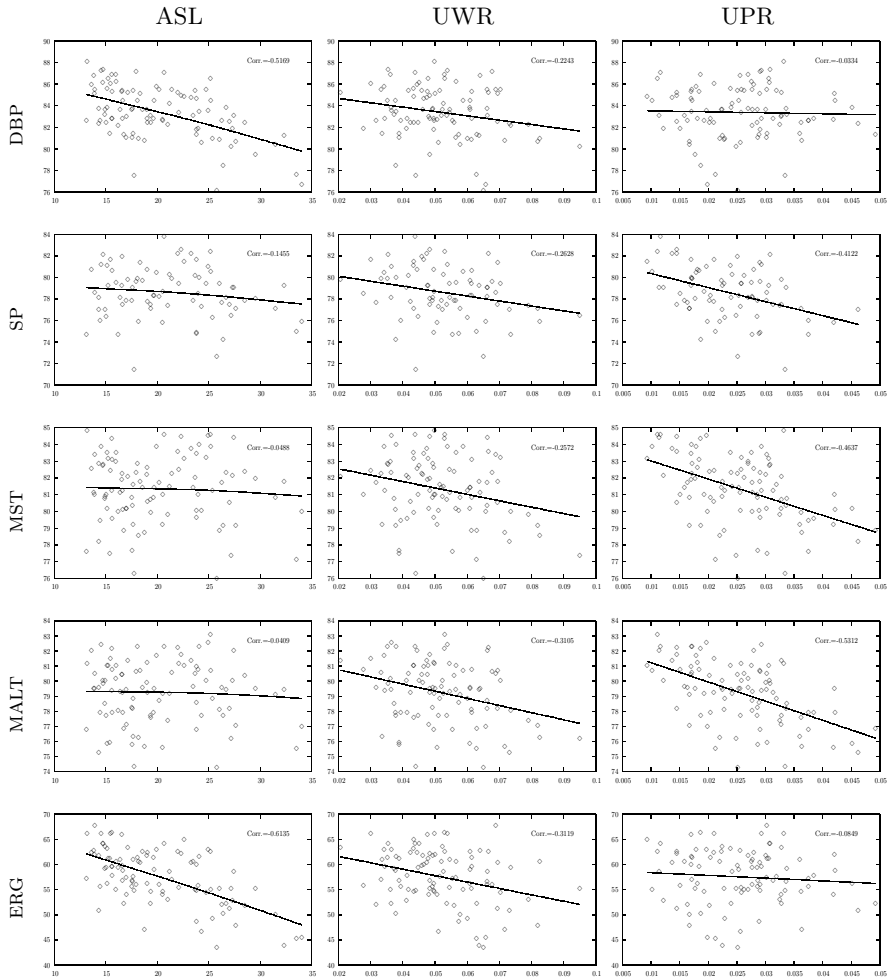


Fig. 1. Parser performance against three statistical measures, and their correlation coefficients

Table 1. Correlation coefficient of the linear regression models using all three measures

	DBP	SP	MST	MALT	ERG
Corr.	0.6509	0.5980	0.6124	0.6961	0.8102

Among all the parsers, ERG has the highest correlation with ASL. This is because the longer sentences lead to a sharp drop in parsing coverage of ERG. Between the two PCFG parsers, the unlexicalized SP parser appears to be more robust against the sentence length. Both dependency parsers appear to be robust to ASL.

With UWR, all parsers shows certain degree of correlation (coefficient from -0.22 to -0.31), with ERG and MALT being the most sensitive ones. An interesting observation is that unexpectedly the unlexicalized parser does not show to be more robust to unknown words than the lexicalized counterpart. SP’s performance not only has higher negative correlation to UWR than DBP does, but also suffers a sharper performance drop with increasing UWR. Both dependency parsers also shows clear performance degradation, indicating the parser is missing critical lexical information. It should be noted that UWR as calculated here does not directly correspond to the unknown words for ERG, which contains a hand-crafted lexicon built independently from the training set (WSJ). But the UWR still reflects how often infrequent words are observed in the dataset. And it is known that most of the ERG parsing errors are caused by missing lexical entries [18].

With UPR, the performances of both dependency parsers show strong negative correlation. MALT has stronger correlation than MST because the transition-based approach is more likely to suffer from unknown sequence of POS than the graph-based approach. The unlexicalized SP shows much more significant correlation to the UPR than the lexicalized DBP does, for the POS trigrams reflect the syntactic patterns on which the unlexicalized parser depends most. ERG performs very robustly to the variation of UPR. This is because that the syntactic constructions in ERG is carefully hand-crafted, and not biased by the training set.

With all parser performance data points, we further built linear regression models using all three measures for each parser, and the correlation coefficient of the models are shown in Table 1. The high levels of correlation indicate that the performance of a parser is largely predictable for a given dataset with these three very simple statistical measures. We expect to achieve even higher correlation if more informative dataset measures is used.

5 Conclusion and Future Work

The method we proposed in this paper can be adapted to various other parsers and datasets, and potentially for different languages. The varying correlation of the proposed statistical measures with parsers’ performance suggests that different parsing models are sensitive to different characteristics of the datasets.

Since all the measures are obtained from the unannotated texts, the method is not committed to specific linguistic framework or parsing algorithm. In the future we plan to experiment with more statistical measures and their combinations.

The linear regression model we built suggests one way of predicting the parser performance on unseen datasets, so that an estimation of the parser performance can be achieved without any gold-standard annotations on the target datasets. The result analysis also shows the possibility of parser combination (by either parse reranking or feature stacking) to achieve more robust performances. Furthermore, the variance of the performance and its correlation to the statistical measures can be viewed as an alternative parser evaluation metrics revealing the robustness of the parser performance, in addition to the standard accuracy measures.

Acknowledgments

The first author is grateful to DFKI and the German Excellence Cluster of Multimodal Computing and Interaction for their support of the work. The second author is funded by the PIRE PhD scholarship program.

References

1. Miyao, Y., Sagae, K., Sætne, R., Matsuzaki, T., Tsujii, J.: Evaluating Contributions of Natural Language Parsers to Protein-Protein Interaction Extraction. *Journal of Bioinformatics* 25(3), 394–400 (2009)
2. Moore, R.K.: Spoken language processing: piecing together the puzzle. *Speech Communication: Special Issue on Bridging the Gap Between Human and Automatic Speech Processing* 49, 418–435 (2007)
3. Bacchiani, M., Riley, M., Roark, B., Sproat, R.: Map adaptation of stochastic grammars. *Computer speech and language* 20(1), 41–68 (2006)
4. McClosky, D., Charniak, E., Johnson, M.: Reranking and self-training for parser adaptation. In: *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*, Sydney, Australia, pp. 337–344 (2006)
5. McClosky, D., Charniak, E., Johnson, M.: When is self-training effective for parsing? In: *Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008)*, Manchester, UK, pp. 561–568 (2008)
6. Nivre, J., Hall, J., Kübler, S., McDonald, R., Nilsson, J., Riedel, S., Yuret, D.: The CoNLL 2007 shared task on dependency parsing. In: *Proceedings of EMNLP-CoNLL 2007*, Prague, Czech Republic, pp. 915–932 (2007)
7. Hara, T., Miyao, Y., Tsujii, J.: Adapting a probabilistic disambiguation model of an HPSG parser to a new domain. In: Dale, R., Wong, K.-F., Su, J., Kwong, O.Y. (eds.) *IJCNLP 2005*. LNCS (LNAI), vol. 3651, pp. 199–210. Springer, Heidelberg (2005)
8. Rimell, L., Clark, S.: Porting a Lexicalized-Grammar Parser to the Biomedical Domain. *Journal of Biomedical Informatics* (in press, 2009)
9. Plank, B.: Structural Correspondence Learning for Parse Disambiguation. In: *Proceedings of the Student Research Workshop at EACL 2009*, Athens, Greece, pp. 37–45 (2009)

10. Dredze, M., Blitzer, J., Pratin Talukdar, P., Ganchev, K., Graca, J.a., Pereira, F.: Frustratingly hard domain adaptation for dependency parsing. In: Proceedings of the CoNLL Shared Task Session of EMNLP-CoNLL 2007. Association for Computational Linguistics, Prague, June 2007, pp. 1051–1055 (2007)
11. Bikel, D.M.: Intricacies of Collins' parsing model. *Computational Linguistics* 30, 479–511 (2004)
12. Collins, M.: Three Generative, Lexicalised Models for Statistical Parsing. In: Proceedings of the 35th annual meeting of the association for computational linguistics, Madrid, Spain, pp. 16–23 (1997)
13. Klein, D., Manning, C.D.: Accurate Unlexicalized Parsing. In: Proceedings of the 41st Meeting of the Association for Computational Linguistics, Sapporo, Japan, pp. 423–430 (2003)
14. McDonald, R., Pereira, F., Ribarov, K., Hajic, J.: Non-Projective Dependency Parsing using Spanning Tree Algorithms. In: Proceedings of HLT-EMNLP 2005, Vancouver, Canada, pp. 523–530 (2005)
15. Nivre, J., Nilsson, J., Hall, J., Chanev, A., Eryigit, G., Kübler, S., Marinov, S., Marsi, E.: Maltparser: A language-independent system for data-driven dependency parsing. *Natural Language Engineering* 13(1), 1–41 (2007)
16. Flickinger, D.: On building a more efficient grammar by exploiting types. In: Oepen, S., Flickinger, D., Tsujii, J., Uszkoreit, H. (eds.) *Collaborative Language Engineering*, pp. 1–17. CSLI Publications, Stanford (2002)
17. Callmeier, U.: Efficient parsing with large-scale unification grammars. Master's thesis, Universität des Saarlandes, Saarbrücken, Germany (2001)
18. Baldwin, T., Bender, E.M., Flickinger, D., Kim, A., Oepen, S.: Road-testing the English Resource Grammar over the British National Corpus. In: Proceedings of the 4th International Conference on Language Resources and Evaluation (LREC 2004), Lisbon, Portugal (2004)