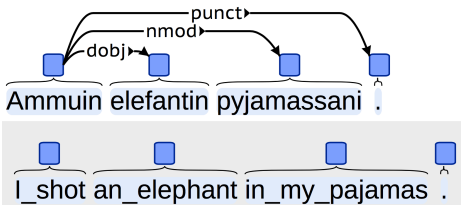
A better evaluation scheme for multilingual parsing

Jimmy Callin, October 2015

As the use of syntactic parsing keeps increasing in applications of computational linguistics, the importance of reliable evaluation methods becomes more and more critical. Today, two evaluation metrics have been more commonly applied than others: Labeled Attachment Score (LAS) and Unlabeled Attachment Score (UAS). These have in their simplicity and intuitiveness served the research area well, but their design relies on a number of assumptions that we argue no longer holds.

Firstly, the design of attachment scores assumes that we know next to nothing about the taxonomy and design choices of the parsing framework. This has historically been necessary since there have not been a consistently adapted framework for dependency parsing across many languages. We argue that recent initiatives in Universal Dependencies (UD) (Nivre 2015) have made this assumption invalid. UD has a carefully specified framework that all UD treebanks has to adapt, and we can exploit this fact to learn more about the performance of a model.

Secondly, parsing results are becoming increasingly juxtaposed in a cross-linguistic manner, and this becomes especially true with the influence of UD. It is not uncommon to compare the output of e.g. English and Finnish under the assumption that equal evaluation scores is equivalent to equal parsing performance.



In this example, we have a Finnish sentence with three edges, while the equivalent sentence in English requires a total of seven edges. This increase does not necessarily mean that the English sentence is harder to parse, since the additional edges are function words which are highly regular in their appearance in treebanks. The problem appears as soon as we introduce a parsing error into the trees: one faulty edge in the Finnish sentence would mean a performance reduction of 33%, while the same error in the English sentence only would reduce the accuracy by 14%.

Our hypothesis is that languages with a comparatively large number of lexicalized grammar particles, like English, receive a “free” performance boost when using classic evaluation metrics.

The purpose of this project is to answer the following questions:

- How much does languages with high lexicalization of grammar particles benefit from current evaluation schemes?
- Would focusing on correct classification of content word dependencies be a better evaluation scheme for cross-linguistic parsing performance?

By studying these topics, we aim to develop a new evaluation scheme that is better suited for cross-linguistic evaluation.

Related work

- M. Collins. 1999. Head-Driven Statistical Models for Natural Language Parsing. Ph.D. thesis, Uni- versity of Pennsylvania.
- J. M. Eisner. 1996. Three new probabilistic models for dependency parsing: An exploration. In Pro- ceedings of COLING, Copenhagen, Denmark
- J. Nivre, J. Hall, and J. Nilsson. 2004. Memory- based dependency parsing. In Proceedings of CoNLL, pages 49–56.
- Carroll, John, Ted Briscoe, and Antonio Sanfilippo. 1998. Parser Evaluation: A Survey and a New Proposal. In Proceedings of the 1st International Conference on Language Resources and Evaluation, 447–54.
- Tsarfaty, Reut, Joakim Nivre, and Evelina Andersson. 2011. "Evaluating Dependency Parsing: Robust and Heuristics-Free Cross-Nnotation Evaluation." In Proceedings of the Conference on Empirical Methods in Natural Language Processing, 385–96. Association for Computational Linguistics.

Methodology

The work is expected to be concluded and sent in for review before December 11. This is the expected time plan:

1. Study the taxonomy of UD relations and make an initial manual classification of content and function dependencies. Having a deep understanding of the framework's taxonomy and its specified dependency relations is critical before continuing onwards. A manual classification gives us the necessary resources for later analyses and development data for the next step. (2 weeks)

2. Further study the distribution of UD relations to see if it is possible to come up with a statistically motivated separation. This would allow us to easily study the effect of varying granularity of content and function dependency separation. The manual classification in the previous step serves as verification of whether the statistical method works or not. (2 weeks)

3. Implement and use the evaluation metric for analysis on an appropriate parser with a suitable number of different languages. We expect the performance gap of analytic languages such as English to drop when compared to more agglutinative languages such as Turkish or Finnish. (2 weeks)

4. Compile results and write paper. After passing review, an implementation of the evaluation scheme will be made openly available online. (1 week)

Further potential experiments, if time allows, include analyzing the effect on performance when varying the amount of training data. Previous experiments have shown weak correlation between data size and performance, which could potentially be explained by not taking into account the relative ease of identifying function word relations.

While we expect the final evaluation scheme to be similar in algorithmic design to the widely used attachment scores, we still want to emphasize what we believe to be essential qualities to keep in mind when developing new metrics:

- *Keep it intuitive* – Evaluation should as little as possible be a black box of magic. If researchers feel like they easily can get an intuitive understanding of its process as well as its output, there is a much higher chance of the metric to get widely adopted.
- *Keep it interpretable* – While an evaluation score is best understood in the context of competing systems, it should ideally be possible to adapt it for getting a deeper understanding of one's system.

We hope to adhere to these qualities in our project.

Significance

We have expressed what problems we see with current evaluation schemes when it comes to cross-lingual evaluation of dependency parsing. Through this project we plan to shed light on these issues, as well as present a viable alternative evaluation scheme that could potentially be adopted in a wider setting.

References

Nivre, Joakim. 2015. "Towards a Universal Grammar for Natural Language Processing." In *Computational Linguistics and Intelligent Text Processing*, 3–16. Springer. http://link.springer.com/chapter/10.1007/978-3-319-18111-0_1.