A better evaluation scheme for multilingual parsing

Jimmy Callin

Purpose and aims

The role of language processing is becoming increasingly multi-lingual, which is reflected in recent efforts into providing dependency parsing frameworks that can reliably be applied on an multitude of languages. One of the most ambitious projects in this area is called Universal Dependencies (UD) (Nivre, 2015), where the goal is to create a parsing framework with a cross-linguistically consistent grammatical annotation. The goal is to remove the requirement of language specific components which has up to this point been a necessity because of inconsistent annotation standards.

To evaluate statistical parsing models, two of the most ubiquitous evaluation metrics are 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 do not hold in the context of UD.

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 progress in UD has made this assumption invalid. UD has a carefully specified framework to which all treebanks has to adapt, and we can exploit this constraint 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. The reason why this is problematic becomes apparent when studying grammatical morphemes in languages where these may be unbounded (i.e. function words) with languages where they typically are bounded on content words (i.e. affixes).

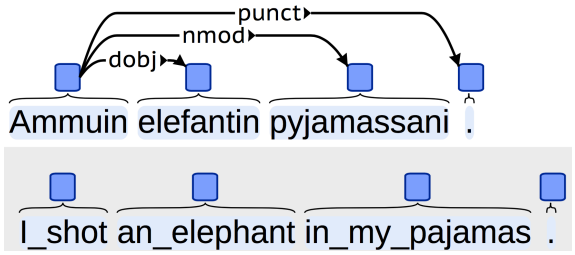

Figure 1: A dependency tree in Finnish.

In figure 1, we have a Finnish sentence with three edges, while the equivalent sentence in English requires a total of seven edges[1]. 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

[1]Example taken from https://universaldependencies.github.io/docs/fi/overview/specific-syntax.html

trees: one faulty edge in the Finnish sentence would result in a performance reduction of 33%, while the same error in the English sentence only would reduce the accuracy by 14%.

Function words are regular in their appearance, and should therefore be relatively easy for a model to parse correctly. Our hypothesis is that languages with a comparatively large number of function words, like English, receive a more or less "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 and implement a new evaluation scheme that is better suited for cross-linguistic evaluation.

Survey of the field

Data-driven evaluation metrics have been used as long as treebanks have been available (see Collins (1999) chap. 4 for a survey on early results). These mainly used unlabeled precision, recall, and accuracy for evaluation, where *accuracy* is equivalent to UAS. The earliest mention of (unlabeled) attachment score we could find was in Eisner (1997), which refers to work done by Lin (1995). Introducing labeled attachment score is credited to Nivre et al. (2004). Carroll et al. (1998) go deeper into parser evaluation methodologies and give a thorough overview of available metrics each with their pros and cons.

Not much work has been done in cross-linguistic evaluation, and papers presenting evaluation scores on several languages simply use previously available metrics without analyzing their shortcomings in such a context. In light of recent work on cross-linguistically consistent annotation frameworks, Tsarfaty et al. (2011) take a separate approach with cross-framework evaluation, where they suggest an evaluation technique that is robust towards differing annotation criteria. Before UD was publicly available, there have been several attempts at automatic normalization of dependency treebanks into a common format for a more robust evaluation (Zeman et al., 2012).

Programme description

We expect the work 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.

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

John Carroll, 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*, pages 447–454.

Michael Collins. 1999. *Head-Driven Statistical Models for Natural Language Parsing*. PhD thesis, University of Pennsylvania.

Jason Eisner. 1997. An empirical comparison of probability models for dependency grammar. *arXiv preprint cmp-lg/9706004*.

Dekang Lin. 1995. A Dependency-based Method for Evaluating Broad-Coverage Parsers. In *In Proceedings of IJCAI-95*.

Joakim Nivre. 2015. Towards a Universal Grammar for Natural Language Processing. In *Computational Linguistics and Intelligent Text Processing*, pages 3–16. Springer, editions.

Joakim Nivre, Johan Hall, and Jens Nilsson. 2004. Memory-Based Dependency Parsing. *Proceedings of the Eighth Conference on Computational Natural Language Learning (CoNLL)*, s. 49-56.

Reut Tsarfaty, Joakim Nivre, and Evelina Andersson. 2011. Evaluating dependency parsing: robust and heuristics-free cross-annotation evaluation. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, pages 385–396. Association for Computational Linguistics.

Daniel Zeman, David Marecek, Martin Popel, Loganathan Ramasamy, Jan Stepánek, Zdenek Zabokrtskỳ, and Jan Hajic. 2012. HamleDT: To Parse or Not to Parse? In *LREC*, pages 2735–2741.