1 Purpose and aims

A recent project, Universal Dependencies\(^1\) (henceforth UD) (Nivre, 2015) is seeking to harmonize annotation across treebanks in different languages and has recently released a number of treebanks with consistent annotation. Harmonization has sometimes been done at the potential cost of parsing accuracy. It has been proposed that modifying the treebanks for the purpose of parsing could rectify that problem (de Marneffe et al., 2014). Research has as a matter of fact shown that transforming certain linguistic constructions in treebanks for parsing led to an increased parsing accuracy (Nilsson et al. (2007) among others). The objective of the proposed project is to reproduce research that has been made on monolingual and non-harmonized treebanks and find out whether or not parsing improvements can be obtained in different languages annotated with the same guidelines. The next section will give further detail on the background research underlying the proposal and the last section will give an outline of the proposed plan.

2 Survey of the field

Dependency Parsing

Syntactic parsing is the task of assigning a structure to words in a sentence and it is used in many Natural Language Processing applications. Syntactic parsing has long been dominated by constituency parsing. Constituency parsing organizes the sentence hierarchically into constituents, i.e. groups of words that work as a unit. However, dependency parsing, which assigns head-dependent relationships between lexical items in the sentence, has recently been gaining ground (Kübler et al., 2009). Dependency parsing models have been preferred to constituency parsing models for their transparent encoding of the predicate-argument structure as well as for their usability for languages for which constituency-based account were not satisfying (Kübler et al., 2009). Dependency parsing models are usually inferred from data and thus require annotated treebanks.

Universal Dependencies

Dependency treebanks exist in a number of languages but recent research in cross-lingual parsing development has raised concerns about existing discrepancies in treebank annotation styles. McDonald et al. (2011) use annotation from some languages to parse other languages and while they notice overall positive results, they also find that language families are not a good predictor of the usefulness of a treebank annotation from one language to another. For example, they find that Greek is the best source language for Dutch while German is the worst. They hypothesize that these unexpected results are due to differences in the annotation styles. Several efforts have been made to harmonize Part-of-Speech annotation (Petrov et al., 2012), morphological annotation (Zeman, 2008) and treebank annotation (Zeman et al. (2012) among others) which have all merged into one project, Universal Dependencies (Nivre, 2015). UD is a project that seeks to harmonize annotation across treebanks. UD guidelines have been designed with the objective to maximise parallelism between languages which has led de Marneffe et al. (2014) to make decisions that they predict to be suboptimal for parsing. For example, they favor content words as heads of prepositional phrases which makes sense cross-linguistically although it has been shown that making prepositions the head of prepositional phrases is better for parsing (Schwartz et al., 2012). For that reason, they propose that the treebanks may be modified for the purpose of parsing to ensure a better parsing accuracy.

\(^1\)http://universaldependencies.github.io/docs/
Tree transformations for parsing

Tree transformations have proved to be useful for parsing in the past. The standard procedure for these tree transformations is first to transform the representation of one or several constructions in the treebank before training the parser. When this is done, the test data can be parsed and the parsed output is transformed back to the original representation for comparability with the gold standard. Several researchers have shown that doing this for certain constructions leads to an improved parsing accuracy. This has been done for constituency parsing by Collins (1999) among others but also for dependency parsing by Nilsson et al. (2007) among others. While in constituency parsing, researchers have focussed on one parsing model and one language, in dependency parsing, there has been research that has shown that these improvements can port to several languages (Nilsson et al., 2007) and to several parsing models (Schwartz et al., 2012). Schwartz et al. (2012) observed a consistent increase in accuracy on their tree transformations across 5 parsing models indicating that the accuracy gains are not specific to one parsing model. However, Nilsson et al. (2007) tested their transformations on two parsing models that use very different parsing strategies and found out that the effect of the transformations depends at least to some extent on the parsing model. Nilsson et al. (2007) investigated transformations on non-projective structures (structures that have crossing dependencies), on coordination constructions and on verbal groups. They observed overall improvements but noticed that their verbal transformation was not applicable to the Dutch treebank due to a difference in the annotation style. They were therefore limited in their study by working on treebanks that have different annotation styles. They admit that this makes it impossible in their study to distinguish differences in syntactic properties of the languages investigated with differences in annotation style in the treebanks of these languages.

Research question

Now that we have UD treebanks, it is possible to go beyond the shortcomings of the study by Nilsson et al. (2007) and investigate the effect of tree transformations in more languages and with more reliability because the different annotation style as a factor that can influence the results has been neutralized. This is what I propose to do in this project. So far, only Silveira and Manning (2015)² have attempted to transform trees in a UD treebank for parsing but they focused on English. I will briefly discuss the methodology proposed in the next section as well as the timeline of my plan.

3 Programme description

Methodology

I propose to reproduce the study by (Nilsson et al., 2007) on UD treebanks, i.e. to transform coordination constructions as well as verbal groups in the treebanks. I will follow the standard procedure, i.e. transform the treebanks before parsing, parse the test data and transform the parsing output back to the original representation. When this is done, parsing accuracy can be compared to the original model and the effect of the different transformations on parsing accuracy can be compared across different languages. The standard attachment score can be used for evaluation. If time permits, I could also investigate other constructions. I propose to conduct this experiment on several parsing models since it has been mentioned that different parsing models could make a difference. Most previous studies on tree transformations have used MaltParser (Nivre et al., 2006) but it would be interesting to see if we can also get improved results using the recently developed state-of-the-art parsers: the Neural Network Dependency Parser (Chen and Manning, 2014) and the stack LSTM parser (Dyer et al., 2015).

Timeline

I propose to organise my time as such: the first two weeks will be devoted to further background readings so that I can identify the most sensible transformations to perform and determine which transformation algorithms are applicable in the case of UD. It will be important to determine if transformations can be easily reversed. The next three weeks will be devoted to implementing the proposed project, i.e. designing and implementing the transformation algorithms. One week will be devoted to experimenting with the different algorithms and the different treebanks and the last week will be left for writing up the project and its results.

²To my knowledge.
References


