Dependency Parsing of Armenian

Hayk Saribekyan (hayks@csail.mit.edu)
Mentors: Yuan Zhang (yuanzh@csail.mit.edu)

November 7, 2015

Abstract

Although significant research has been done recently addressing dependency parsing of natural languages, for many languages there have been results. This project aims to explore parsing of low-resource languages and uses Armenian as an example of such language. The project will try existing methods of parsing. Parciularly, it will focus on cross-lingual parsers, which rely on having some kind of similarity between source and target languages, and a relatively small treebank in the target language.

This project will explore different approaches for parsing of Armenian language, and possibly come up with novel ways to tackle such problem.

1 Background

Linguists classify Armenian as an independent branch of Indo-European Languages. Armenian is a relatively small language (spoken by less than 10 million people), which makes data collection and annotation challenging. The language also has its unique script.

Armenian is a non-projective language. Unlike English, Armenian has 7 noun cases. For example, the phrase "on table" can be translated to "**սեղան**ին" or "**սեղան**ի վրա", where the nominative case of the word "table" is shown in bold.

These features make parsing of Armenian more challenging and interesting to see what are the results we can achieve.

2 Methods and Related Work

The first step of the project is to get large enough annotated treebank in Armenian. For some approaches for training it is enough to have just POS tags for Armenian text. Other approaches require some (small) treebank in target language. For testing and evaluation a dependency treebank is necessary.

Due to lack of enough resources the treebank will be small. So, to achieve reasonable parsing accuracy some kind of transfer of dependency parser from a resource-rich language will be used.

Our initial approach will use only sentences in Armenian with ground truth POS tags for training. The POS tags are used to find a similar language that has a comprehensive treebank. Then a delexicalized parser is transferred from that language to Armenian [3].

Patterns between two languages can be identified using n-gram distributions on POS tag sequences. For example, since in Armenian the order *adjective-noun* is more common than *noun-adjective*, English in this respect should be closer to Armenian than French (where the opposite order is more common).

This approach is very simple and it will give us first preliminary result. However, if small treebank and parallel data is available for Armenian more complex methods can be applied.

The method described in [1] trains a neural network model for both a resource-rich language (e.g. English) and resource poor language (e.g. Armenian). So, the resulting model similar to a universal dependency parser. To avoid having a model dominated by the source language in case the source treebank is much larger than the target, during the training the data batches that are fed to the network have the same number of samples from both languages. In this case too, the model assumes some similarity between languages.

Availability of parallel data is required in [5]. It first uses GIZA++ [4] to project dependencies from source to target. Then a parser is trained on this projected set. This approach however requires both languages to be projective, which is not the case in Armenian. Although Armenian is not a projective language in the majority of the cases sentences have projective dependency.

3 Data

All approaches, some more than the others, require a treebank of some size. Armenian does not have any available one, but I am working on creating a small treebank of few hundred sentences with some help from Armenia.

EANC [6] is a large corpus developed by a team of linguists. It contains POS tags for large texts and I anticipate to get their data in a reasonable format.

Parallel data can be generated relatively easily using simple dynamic programming techniques [2], because Armenian has large amount of translated literature.

References

[1] Long Duong, Trevor Cohn, Steven Bird, and Paul Cook. A neural network model for low-resource universal dependency parsing. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 339–348, Lisbon, Portugal, September 2015. Association for Computational Linguistics.

[2] William A Gale and Kenneth W Church. A program for aligning sentences in bilingual corpora. *Computational linguistics*, 19(1):75–102, 1993.

[3] Ryan McDonald, Slav Petrov, and Keith Hall. Multi-source transfer of delexicalized dependency parsers. pages 62–72, 2011.

[4] Franz Josef Och and Hermann Ney. A systematic comparison of various statistical alignment models. *Computational Linguistics*, 29(1):19–51, 2003.

[5] Mohammad Sadegh Rasooli and Michael Collins. Density-driven cross-lingual transfer of dependency parsers. pages 328–338, 2015.

[6] Corpus Technologies. Eastern armenian national corpus. 2007-2009.