Title: Larger Treebank for Amharic Annotated with Semi-automatic Method

Applicant:
Abnet Shimeles Ibssa
PhD student at Addis Ababa University, Addis Ababa, Ethiopia (Africa)

Keywords: Larger Amharic Treebank, Universal dependency, neural parsers

1. Overview of the Project

UD Treebanks are non-existent for most low-resourced languages, and if they exist, they are typically small in size. Amharic, an Ethio-semitic morphologically rich and low resourced language spoken by more than 58 million speakers in Ethiopia, Africa, is no different. The first and only Amharic UD Treebank[1] was released in UD 2.2[2] and contains about 10K words [1], [2]. Motivated by the Universal Dependency project open call for contributors and Amharic NLP hunger for more data, we opt to seek for a faster way of building an Amharic Treebank that fits the UD framework in this project. We propose to employ a semi-automatic methodology that is faster than the manual process to create an Amharic UD Treebank by taking advantages of existing resources available as a result of prior efforts. In this regard, we propose a methodology to minimize the manual intervention of creating treebanks by systematically making use of the resources at hand, which is briefly described below.

The primary source of data for the Amharic treebank would be CACO, a morpho-syntactically annotated Amharic dataset, which has more than 1M sentences and 24M orthographic words [3]. The researchers in this work used an improved version of the existing Amharic morphological analyzer, HornMorpho [4] to do the segmentation and analysis automatically. We planned to further improve HornMorpho to make the segmentation produce half-processed sentences, which consist of segmented syntactic words, which are annotated with POS tags, morphological features and within-word syntactic relations that abide by the UD guidelines.

Inspired by the reasonable performance achieved by the off-the-shelf neural network parsers when tested on gold segmented data (LAS score of 85.26, and 93%-95% for POS tagging) reported in [5], we propose to begin with exploring state-of-the-art neural network parsers that are easily adaptable to Amharic language. Then, we plan to train the models with manually annotated Treebank (We call it "Starter Treebank) and enhance them with contextualized pre-trained word embeddings. The best performing neural parsers will then be applied to parse the segmented sentences to produce the Treebank. We also plan to involve experts (at least two) to check the output for errors and fixed them. This human intervention would be controlled by measuring the agreement between the experts and seeking for validation from a third expert. Since we are dealing with a large corpus, we explore various techniques to reduce the time required for this task. These include iterative processing, error analysis and systematic error detection.

[1] https://github.com/UniversalDependencies/UD_Amharic-ATT/blob/master/README.md
[2] https://universaldependencies.org/

2. Phase of the Project

The project has been started since the beginning of March 2024, and we are currently annotating the Starter Treebank manually. We have created about 100 reference sentences that represent the various UD relations, and annotated them to use them as a reference when annotating more sentences (we expect to annotate about 1000 sentences), which will be used to train the neural parsers.

The rich morphology of Amharic allows the possibility of binding content words, inflectional morphemes and function words together to create an orthographic word. Thus, an orthographic word may function as a phrase or full-fledged sentence and may possibly contain elements representing syntactic relations in addition to morphological information. Thus, the analysis of the reference sentences took more time than we expected, and some relations which seem to be unique in Amharic are found to be difficult to analyze. Participating in the training offered by UniDive would help to understand the UD framework better, get acquainted with the standard techniques of annotating UD treebanks, and would create an opportunity to discuss various issues that we found peculiar to Amharic. Some of the open questions are listed below.

3. Open Questions

1. How do you identify what relations are relevant to a new language in UD?
2. How do you analyze the objects of "experiencer/possessor" verbs?
3. How do you analyze inverted and uninverted clefts in Amharic?
4. How to choose between two relations (obl and advmod) for derived words that serve as adverbs (almost lexicalized)?

References

[1] B. E. Seyoum, Y. Miyao, and B. Y. Mekonnen, “Universal Dependencies for Amharic,” p. 8, 2018.

[2] E. B. Seyoum, B. Yimam, and Miyao, Yusuke, “Morpho-syntactically Annotated Amharic Treebank,” p. 11, 2016.

[3] A. M. Gezmu, B. E. Seyoum, M. Gasser, and A. Nürnberger, “Contemporary Amharic Corpus (COCA),” 2018, doi: 10.24352/UB.OVGU-2018-144.

[4] M. Gasser, “HornMorpho: a system for morphological processing of Amharic, Oromo, and Tigrinya,” p. 7, 2011.

[5] B. E. Seyoum, Y. Miyao, and B. Y. Mekonnen, “Comparing Neural Network Parsers for a Less-resourced and Morphologically-rich Language: Amharic Dependency Parser,” p. 6, May 2020.