Developing accurate dependency parsing tool for Galician as low-resource language

Albina Sarymsakova[1*], Xulia Sánchez-Rodríguez[2*] and Laura Castro[1]

[1]Centro Singular de Investigación en Tecnoloxías Intelixentes (CiTIUS)
Universidade de Santiago de Compostela
[2]Language Variation and Textual Categorisation (LVTC) Universidade de Vigo
[*]Equal contribution

Keywords: Galician, automatic parser, Universal Dependencies, BERT

Automatic syntactic parsing is a fundamental component of natural language processing (NLP). However, to achieve accurate results, robust parsing tools rely on well-annotated treebanks with extensive and high-quality data, as posited by Kondratyuk and Straka. This requirement poses challenges, especially for low-resource languages like Galician, where parsing quality remains insufficient (Gamallo Otero & González López, 2012). Additionally, models trained on specific datasets often struggle to perform well across diverse domains (Vania, Kementchedjhieva, Søgaard, & Lopez, 2019).

In light of this, our project explores various approaches to enhance the automatic syntactic analysis of Galician using the Universal Dependencies (UD) framework. As the previous studies in the context of CoNLL 2017 and 2018 UD Shared Tasks (Zeman et al., 2017, 2018) have shown, the best performance of the Galician TreeGal model achieved the Labeled Attachment Score (LAS) equivalent to 74.34% and Unlabeled Attachment Score (UAS) of 79.17% of accuracy using validation mode from raw text. Our experimental design involves increasing the size of the initial training corpus by incorporating data from a new Galician UD treebank in order to improve the performance of our model, offering higher metrics of LAS and UAS scores. Furthermore, we analyze the advantages of using contextualized vector representations by comparing different BERT models. Additionally, we evaluate the impact of integrating cross-lingual training data from similar language domains on our model[1] performance across various treebanks.

Among others, our key findings outline the following aspects:

a Augmenting training data positively correlates with improved model performance across different treebanks.
b Monolingual BERT models outperform their multilingual analogues.
c Incorporating cross-lingual data allows to enhance overall model performance across our treebanks.
d To our best knowledge, the resulting model achieves superior performance in dependency parsing compared to previous studies on Galician.

Regarding the future of our project, our aim is to expand training datasets by incorporating additional treebanks, not only from the Galician but from other linguistic domains since our project, along with the previous studies, has evidenced enhanced performance of the model, incorporating supplementary data from closely related languages. We also aspire

[1] Both the model and the new treebank will be made freely available.

to augment manually annotated resources to train more resilient parsers. It is worth to mention that our project presents a new treebank and a model, which we believe yield the most promising results to date in UD parsing for Galician.

Participating in the 1st UniDive training school would benefit our project and provide us with more fundamental insights about the automatic syntactic analysis and treebanks development for low-resources languages. Since one of the main objectives of this training school includes sharing of hands-on experience in utilizing new or enhanced Universal Dependencies treebanks, we consider it a unique opportunity to learn about state-of-the-art of manually annotated resources, especially in the context of low-resource languages, such as Galician.

Moreover, the introduction of PARSEME corpus at the training school offers us the opportunity to access additional annotated data specifically related to parsing and analyzing multiword expressions (MWEs). MWEs pose significant challenges in syntactic analysis, particularly for object of our study, low-resources Galician language, and having access to this corpus enables us to develop more robust parsing tools that can accurately handle such linguistic phenomena. Overall, we expect the 1st UniDive training school to equip us with valuable knowledge, skills, and resources essential for advancing our project on automatic syntactic analysis of Galician.

Concerning the questions related to the project that could be addressed during a brainstorming hackathon, we posit as follows:

- What approaches can be employed to improve parsing accuracy for complex syntactic structures, such as long-distance dependencies?
- Are there other approaches, such as incorporation of cross-lingual data into training dataset, that are capable to improve the automatic parsing models' performance?
- What impact would adding a new annotation layer to manually annotated corpora (including Galician one) have on automatic parsing models in context of low-resources languages?

Currently our project is situated in a development phase. We created a new Galician UD treebank and used it in order to improve the existing automatic parsing tools. Additionally, the project is focusing on expanding and refining Galician treebanks to provide comprehensive and high-quality annotated data for training and evaluating parsing models. Through this phase, our goal is to advance the state-of-the-art in automatic syntactic analysis of Galician, ultimately enabling more accurate and effective natural language processing applications for this language.

References

Gamallo Otero, P., & González López, I. (2012). Deppattern: a multilingual dependency parser. In *Proceedings of propor.*

Kondratyuk, D., & Straka, M. (2019). 75 languages, 1 model: Parsing universal dependencies universally. In *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 2779–2795). Hong Kong, China: Association for Computational Linguistics. Retrieved from https://www.aclweb.org/anthology/D19-1279

Vania, C., Kementchedjhieva, Y., Søgaard, A., & Lopez, A. (2019). A systematic comparison of methods for low-resource dependency parsing on genuinely low-resource languages. In K. Inui, J. Jiang, V. Ng, & X. Wan (Eds.), *Proceedings of the 2019 conference on empirical methods in natural language processing and the 9th international joint conference on natural language processing (emnlp-ijcnlp)* (pp. 1105–1116). Hong Kong, China: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/D19-1102` doi: 10.18653/v1/D19-1102

Zeman, D., Hajič, J., Popel, M., Potthast, M., Straka, M., Ginter, F., … Petrov, S. (2018, October). CoNLL 2018 shared task: Multilingual parsing from raw text to Universal Dependencies. In D. Zeman & J. Hajič (Eds.), *Proceedings of the CoNLL 2018 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 1–21). Brussels, Belgium: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/K18-2001` doi: 10.18653/v1/K18-2001

Zeman, D., Popel, M., Straka, M., Hajič, J., Nivre, J., Ginter, F., … Li, J. (2017, August). CoNLL 2017 shared task: Multilingual parsing from raw text to Universal Dependencies. In *Proceedings of the CoNLL 2017 shared task: Multilingual parsing from raw text to universal dependencies* (pp. 1–19). Vancouver, Canada: Association for Computational Linguistics. Retrieved from `https://aclanthology.org/K17-3001` doi: 10.18653/v1/K17-3001