Application of multilingual models in POS tagging for Alsatian and Dagur

Joanna Dolińska & Delphine Bernhard
(University of Warsaw & University of Strasbourg)

Keywords: Dagur, Alsatian, multilingual models, endangered languages, POS tagging

This presentation is dedicated to two extremely low-resource languages from typologically distant language families: Alsatian, a group of Germanic dialects spoken in Northeastern France and Dagur, a Mongolic language mostly used in Northeast China. Both Alsatian and Dagur are characterized by the absence of a unified spelling system and the small size of available corpora manually annotated with part-of-speech (POS) tags.

We address the question of whether automatic POS tagging is feasible for Alsatian and Dagur when working with multilingual models fine-tuned for POS tagging (De Marneffe et al. 2021) on the Universal Dependencies (UD) corpora by de Vries et al. (2022). Our goal is to assess the potential for transfer learning from other related and unrelated languages and to identify the settings and data transformations that perform best.

The corpora used in this work are an Alsatian corpus comprising 12,582 tokens in Latin script (Bernhard et al. 2018) and a Dagur corpus comprising 4,502 tokens in Cyrillic (Dolińska and Bernhard 2024). Both have been manually annotated with UD POS tags. We carried out a comparison of several different zero-shot methodologies for automatic POS tagging and investigated the effects of linguistic proximity with the source language used for training (language family and script), as well as spelling variation reduction techniques.

For Alsatian, we used several procedures based on bilingual German-Alsatian lexicons to transpose Alsatian lexical items into their German translation, taking advantage of the proximity between the Alsatian dialects and German. We found that these simple transformations had a large positive impact on POS tagging accuracy, without the need to retrain the models. In contrast to Alsatian, none of the models by de Vries et al. (2022) were trained on a language from the same linguistic family. We therefore trained other models in order to include Buryat, which is currently the only Mongolic language represented in UD corpora (Badmaeva and Tyers 2017, and Badmaeva and Tyers 2023). This resulted in three different zero-shot settings: (1) models trained on an unrelated language, (2) a model trained on the related Buryat language and (3) a combination of training on an unrelated language + Buryat. We observed the best performance by training on Buryat, while Uyghur and Turkish are among the best performing unrelated languages. This confirms that the linguistic proximity of languages belonging either to the same family or to closely related families has a positive effect on performance.

Acknowledgements

This work has been conducted in part within the framework of the ANR-21-CE27-0004 DIVITAL project supported by the French National Research Agency and of the Miniatura-6 grant 2022/06/X/HS2/01374 financed by the National Science Centre in Poland.

References

Badmaeva, Elena and Tyers, Francis M. (2017), Dependency Treebank for Buryat. In *Proceedings of the 15th International Workshop on Treebanks and Linguistic Theories* (TLT15), 1–12.

Badmaeva, Elena and Tyers, Francis M. (2023), *UD Buryat-BDT Treebank*. Universal Dependencies v2.12.

Bernhard, Delphine, Ligozat, Anne-Laure, Martin, Fanny, Bras, Myriam, Magistry, Pierre, Vergez-Couret, Marianne, Steiblé, Lucie, Erhart, Pascale, Hathout, Nabil, Huck, Dominique, Rey, Christophe, Reynés, Philippe, Rosset, Sophie, Sibille, Jean and Lavergne, Thomas (2018), Corpora with Part-of-Speech Annotations for Three Regional Languages of France: Alsatian, Occitan and Picard, in N. Calzolari, K. Choukri, C. Cieri, T. Declerck, S. Goggi, K. Hasida, H. Isahara, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis and T. Tokunaga (eds), *Proceedings of the 11th Edition of the Language Resources and Evaluation Conference*, Miyazaki, 3917– 3924.

De Marneffe, Marie-Catherine, Manning, Marie-Catherine, Nivre, Joakim and Zeman, Daniel (2021), Universal dependencies. *Computational linguistics* 47(2), 255–308.

De Vries, Wietse, Wieling, Martijn and Nissim, Malvina (2022), Make the Best of Cross-lingual Transfer: Evidence from POS Tagging with over 100 Languages, in *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics*, vol.1 (Long Papers), Dublin: Association for Computational Linguistics, 7676–7685.

Dolińska, Joanna and Bernhard, Delphine (2024), POS Tagging for the endangered Dagur language, in *Proceedings of the 2024 Joint International Conference on Computational Linguistics, Language Resources and Evaluation (LREC-COLING 2024)*, 12906-12916.