Cuneiform Language Identification (CLI) shared task at VarDial Evaluation Campaign 2019

Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, Krister Lindén

We will present the Cuneiform Language Identification (CLI) shared task that we organized as part of the VarDial Evaluation Campaign in 2019.[1] A shared task is a competition type of event where teams of data scientists from all around the world are able to test their methods in a comparable way. Our dataset consisted of individual lines in Sumerian and six Akkadian dialects. We will introduce the dataset, as well as the results attained, and the methods used by the participating teams.

Automatic language identification is the task of determining the language a given text is written in by using only the clues found in the text itself. The first ever language identification experiments in the languages using the cuneiform script were carried out by us in autumn 2018 and presented at the ASOR/EPHE symposium in Paris.

Using the transliterated cuneiform texts from the Oracc database, we created a dataset that could be used in a shared task.[2] We published the training and the development data in January 2019 and the participating teams had around a month to experiment and investigate how to distinguish between these languages and dialects using the cuneiform signs as encoded in Unicode. The training and the development set included the language labels associated with each line. In February, we gave the participants the test set, which did not include the language labels and the participants had two days in which to submit their results. No information outside the dataset provided by us was to be used.

In the end, eight teams provided results on the shared task. The team from the National Research Council Canada won the shared task, establishing new state-of-the-art for cuneiform language identification using deep neural networks.[3]

---

[1] Tommi Jauhiainen, Heidi Jauhiainen, Tero Alstola, and Krister Lindén. 2019a. Language and Dialect Identification of Cuneiform Texts. arXiv preprint, arXiv:1903.01891.

[2] Marcos Zampieri, Shervin Malmasi, Yves Scherrer, Tanja Samardzic, Francis Tyers, Miikka Silfverberg, Natalia Klyueva, Tung-Le Pan, Chu-Ren Huang, Radu Tudor Ionescu, Andrei Butnaru, and Tommi Jauhiainen. 2019. A Report on the Third VarDial Evaluation Campaign. In Proceedings of the Sixth Workshop on NLP for Similar Languages, Varieties and Dialects (VarDial). Association for Computational Linguistics.

[3] Gabriel Bernier-Colborne, Cyril Goutte, and Serge Léger. 2019. Improving Cuneiform Language Identification with BERT. In Proceedings of VarDial.