UD and POS tagging in Uyghur

Faruk Mardan, University of Leeds, UK

Key words: universal dependency, Uyghur, corpus annotation, POS tagging

Description:

The proposed project for the UniDive Training School aims to build a corpus in Uyghur for Universal Dependency annotation. This proposed project will be instrumental to my PhD research in building a machine translation model between English and Uyghur with the help of resources in other languages.

Uyghur is a widely spoken Turkic language with an estimated 12 million speakers. However, the relatively large number of Uyghur speakers has not translated to an abundance of data available for language model and translation model training. The use of Uyghur in the cyberspace is under severe oppression and censorship. This trend has exacerbated in recent years, with the majority of Uyghur language sites shut down permanently (Wired, 2022).

Parallel data and attention mechanism are essential first steps to the training of translation models (Bahdanau et al., 2016). However, as do all other under-resourced languages, there is a general lack of parallel data between Uyghur and any other language. One approach to somewhat compensate for the lack of parallel data is to use additional monolingual data (Koehn, 2020). The monolingual data can be used to improve fluency of the output and the missing half of the data can also be synthesised using techniques such as back translation (*ibid*). To maximise the use of monolingual data in an under-resourced language, it is important to effectively annotate the data. UD and POS tagging are essential to this process.

How the training will help

The training will provide me the in-depth knowledge of Dependency Syntax and particularly Universal Dependency, as well as a good understanding of the tools and resources needed for corpus annotation. This would help me build a solid foundation to my PhD training. I am also keen to contribute to the UD annotation for Uyghur on universaldependency.org. Having the training on and knowledge of UD would be extremely helpful. Particularly, getting to know and seeking the guidance of Daniel Zeman and Francis Tyers, who are directly involved in the annotation of UD resources in Uyghur, would be valuable to my research.

In addition to its significance to my PhD training, I am keen to attend the training also because I would love to explore how UD annotation in other Turkic languages compares to Uyghur, and how this can be harnessed to enrich the annotation and facilitate building of machine translation systems. Similar work has been done by many researchers, such as (Sharoff, 2019), who created cross-lingual spaces for word embedding in Slavic languages. (Chen & Abdul-Mageed, 2021) explored transferring knowledge of richer Indo-European languages to their low-resourced counterparts and compared the effects of linguistic similarity on such knowledge transfer. UniDive training school would be a great opportunity for me to explore the potential of transferring knowledge of all related languages to compensate for the lack of data in Uyghur.

Question for the hackathon

Is there a potential for AI-powered Optical Character Recognition technology to be used to convert scanned documents into digital texts (Rijhwani et al., 2020) to expand language data?

Project phase

I have only recently started my PhD research and the project is at its initial planning stages.

References

Bahdanau, D., Cho, K., & Bengio, Y. (2016). *Neural Machine Translation by Jointly Learning to Align and Translate*.

Chen, W.-R., & Abdul-Mageed, M. (2021). Machine Translation of Low-Resource Indo-European Languages. *CoRR*, *abs/2108.03739*. https://arxiv.org/abs/2108.03739

Koehn, P. (2020). *Neural Machine Translation*. Cambridge University Press. https://doi.org/10.1017/9781108608480

Rijhwani, S., Anastasopoulos, A., & Neubig, G. (2020). *OCR Post Correction for Endangered Language Texts*.

Sharoff, S. (2019). Finding next of kin: Cross-lingual embedding spaces for related languages. *Natural Language Engineering*, *26*(2), 163–182. https://doi.org/10.1017/S1351324919000354

Wired. (2022). *The Strange Death of the Uyghur Internet | WIRED*. Wired. https://www.wired.com/story/uyghur-internet-erased-china/