Testing Available Mongolian NER Tools and Future Perspectives.

Dr Christian Faggionato

November 2019

Prepared for the project "Named-Entity Recognition in Tibetan and Mongolian Newspapers" for the Mongolian and Inner Asian Studies Unit, Cambridge University
Funded by: Cambridge Language Sciences, Cambridge University
Project PI: Dr Hildegard Diemberger; Coordinator: Dr Robert Barnett

Introduction

Natural Language Processing (NLP) and text mining are fields of research that seek to leverage rich information tools with the goal of understanding, extracting, and retrieving unstructured and structured text. For researchers using computational methods to analyse and study languages, natural language resources are essential. Such resources are needed to help advance the development of NLP tools, machine learning, Named-Entity Recognition (NER) and other procedures that involve the analysis of large quantities of textual data, usually in the form of corpora. Among these procedures, NER is a technique in the information extraction field that recognizes nouns denoting recognizable proper names, titles, dates, times, etc. within a text and classify them into pre-defined categories.

In this paper, we look at NLP tools for use with Mongolian language. Some are designed for use with Cyrillic script, as used in the independent state of Mongolia, while others are intended for use with the traditional vertical script used in the Inner Mongolian Autonomous Region IMAR), part of the People's Republic of China. We want to look at both groups of NLP tools, but in particular we want to see if already available tools that work with Cyrillic Mongolian can be used with Vertical Mongolian.

Two existing NER tools for Cyrillic script

In the present pilot project, we tested two existing, freely available NER tools able to recognize proper nouns for the Mongolian language, targeting three classes: names of people, places and organizations. We used the default PER, ORG and LOC tags to identify individual names, organisations and places. These entity tags are encoded within a B-I-O annotation scheme. The targeted entities labels are prefixed with B- (which denotes the beginning of an entity) or with -I (which denotes the "inside" of an entity). This annotation scheme helps to identify multiword entities. Other words in the text string are labelled with the O tag.

The first NER demo version is provided by the National University of Mongolia and is designed for use with the Cyrillic script. Sampling from different sources, most of them online newspapers, we have been able to measure an accuracy of 80-85% in identifying our categories of interest. The Mongolian NER tool can be found at the following address: http://172.104.34.197/nlp-web-demo/

Another stand-alone version of NER for Mongolian language is available on GitHub: https://github.com/enod/mongolian-bert-ner. We are still in the process of testing the full potential of the GitHub NER for Mongolian. However, according to the results provided on the GitHub repository, the tool can reach a higher accuracy than the online tool previously tested. Here is an example of the output created by the Mongolian NER tool for the string "Мөрдөн шалгах хэлтсийн дарга Д.Батбаяр":

```
#  'Мөрдөн': {'tag': 'B-ORG', 'confidence': 0.9999772310256958},
#  'шалгах': {'tag': 'I-ORG', 'confidence': 0.9999890327453613},
#  'хэлтсийн': {'tag':
 'I-ORG', 'confidence': 0.8935487270355225},
#  'дарга': {'tag': 'O', 'confidence': 0.9999908208847046},
```

```
#  'Д.Батбаяр': {'tag': 'B-PER', 'confidence': 0.9998291730880737},
```

In this example the multiword entity Мөрдөн шалгах хэлтсийн, “Inspection Department”, is correctly identified, assigning the B- and I- tags to each component according to their position within the entity. For entities consisting of only one word, such as 'Д.Батбаяр', a personal name, the tag B-PER is correctly assigned.

Manually Annotated Corpora for Vertical Mongolian

It is worth noting that Mongolian NER systems have been already built and trained using manually annotated corpora in vertical script. These are described in Wang W., Bao F., Gao G., 2016b. However, the data used for the procedures described in that paper have not been made available to the public and do not appear to be accessible.

Refining NER Tools for Mongolian

The actual NER tools for Mongolian have been built based on default NER models which are good enough for basic tasks. However, for the specific objectives of future projects, a custom Mongolian NER model is needed. Additional rules will need to be added in order to identify new entities in our targeted corpora of investigation in order to improve the already good performance of the actual NER. The traditional methods used for Mongolian NER are the Conditional Random Field (CRF) and Long-Short Term Model (LSTM): there has been already some evidence that Bidirectional Recurrent Neural Networks – BLSTMS - outperforms CRF models using manual features (Wang, W., Bao, F., Gao, G. , 2016c). Further improvement could be obtained with NER methods based on attention mechanism (Tan M., Bao F., Gao G., Wang W., 2019).

A recent study showed that by adding a number of categories or features of the Mongolian language - orthographic features, morphological features, gazetteer features, syllable features, word-clusters features and so on - the overall system performance is improved (Wang W., Bao F., Gao G., 2016a). Through this approach the features that benefit the most are the stem features. Optimal performance results from a combination of all the features.

Not only are new rules needed to improve NER for Mongolian, but a deeper look into the morphosyntactic structure of the language is also essential. Considering the agglutinative nature of Mongolian, different morphological processing methods have been investigated in a combined study of syllable features, lexical features, context features, morphological features, and semantic features. The results showed that, to improve NER performance, it is better to segment each suffix into an individual token than to delete suffixes or use suffixes as a feature (Wang W., Bao F., Gao G., 2016b).

Conclusions and Feasibility of Applying Cyrillic Tools to Vertical Script

The tested tools work with Mongolian Cyrillic scripts. However, this should not represent an obstacle in handling sources written in traditional vertical (also known as Uyghur-Mongolian) script, as long as a suitable conversion tool exists for converting Cyrillic to vertical Mongolian. We therefore extensively tested the online converter provided by the Inner Mongolian University (http://trans.mglip.com/EnglishC2T.aspx). We found the following issues occurring during the conversion process:

- When Cyrillic is written in all capital letters, the converter does not recognize the words.
- The converter does not recognize text abbreviations well.
- The converter sometimes does not recognize English letters that are used in Cyrillic Mongolian (e.g. Km, Kg)
- The converter makes many mistakes when it comes to prepositions (connecting words).
- The converter makes mistakes with foreign proper nouns.

- Sometimes words in Cyrillic are converted into words in vertical script with a similar pronunciation but a different meaning (often such words are obsolete or do not make sense).

The NLP tools that we found and that are publicly available have been developed to work with Mongolian Cyrillic script. The Vertical Mongolian to Cyrillic Mongolian converters are not sufficiently developed at the time of writing to make possible the use of the Cyrillic based NLP tools for vertical Mongolian. It is therefore clear that major improvements need to be made to the converters to improve their overall accuracy.

Bibliography

Wang W., Bao F., Gao G (2016a). Cyrillic Mongolian Named Entity Recognition with Rich Features. In: Lin CY., Xue N., Zhao D., Huang X., Feng Y. (eds) Natural Language Understanding and Intelligent Applications. ICCPOL 2016, NLPCC 2016. Lecture Notes in Computer Science, vol 10102. Springer, Cham.

Wang, W., Bao, F., Gao, G. (2016b). Mongolian Named Entity Recognition System with Rich Features. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. 505–512. The COLING 2016 Organizing Committee, Osaka, Japan.

Wang, W., Bao, F., Gao, G. (2016c). Mongolian named entity recognition with bidirectional recurrent neural networks. In: Proceedings of the 28th IEEE International Conference on Tools with Artificial Intelligence (ICTAI), pp. 495–500 (2016).

Tan M., Bao F., Gao G., Wang W. (2019). An Attention-Based Approach for Mongolian News Named Entity Recognition. In: Sun M., Huang X., Ji H., Liu Z., Liu Y. (eds) Chinese Computational Linguistics. CCL 2019. Lecture Notes in Computer Science, vol 11856. Springer, Cham.

Available literature on Mongolian NER

Banu Diri (2012). Mongolian Named Entity Recognition, Yildiz Technical University, Faculty of Electrical and Electronics, Computer Engineering Department, Istanbul.

Zoljargal Munkhjargal, Gabor Bella, Altangerel Chagnaa, Fausto Giunchiglia (2015). Named Entity Recognition for Mongolian Language. TSD 2015 Proceedings of the 18th International Conference on Text, Speech, and Dialogue - Volume 9302, 243-251, Springer-Verlag New York.

Wang, Weihua & Bao, Feilong & Gao, Guanglai. (2015). Mongolian Named Entity Recognition using suffixes segmentation. 169-172. 10.1109/IALP.2015.7451558.

Wang, W., Bao, F. & Gao, G. (2019). Learning Morpheme Representation for Mongolian Named Entity Recognition. Neural Process Letter, pp 1–18.