УДК 004.8

CHALLENGES IN NLP TEXT ANALYSIS IN MANDARIN CHINESE AND LAO: A CASE STUDY OF MEDIA ARTICLES

IUREVA V. (Beijing Language and Culture University), KOSUKHIN P. (Higher School of Economics)
Advisor – IUREV R. (ITMO University)

Introduction. This paper describes specific challenges associated with NLP text analysis in Mandarin Chinese and Lao, on an example of analysis of two media articles, one in Mandarin Chinese and one in Lao. The analysis is conducted using Python. The goal of the analysis is to extract the words repeated most frequently in both texts. In this paper, we present a short description of challenges we encountered while conducting analysis of the media articles, and the possible solutions we came up with to overcome the difficulties existing in NLP for Chinese and Lao.

Main part. There are certain challenges related to the processing of texts in Mandarin Chinese and Lao languages. During preprocessing, a major difficulty for tokenization is that both Chinese and Lao do not have spaces between words, as well as capital letters. Moreover, differentiating specific parts of speech, function words in particular, is quite complicated because of grammatical flexibility performed by many Chinese words, where context and sentence structure plays a crucial role in defining the function of a word. Similarly, due to limited number of Lao letters some words may be written similarly but their meaning is different. Thus, extracting of some words rely only on context. At the same time, Lao and Chinese lexicology has their own peculiarities that require adjusting of the lemmatization algorithms. In Chinese, words can be monosyllabic or polysyllabic; in polysyllabic words, there usually are more than one root, and the meaning of the word often depends on the type of relation between the roots forming the word. In Lao, many words are borrowed from other language such as Pali, Sanskrit, Khmer, English etc, and such words consist fully of one root, which makes lemmatization useless in many cases. Here we present possible practical solutions for these and other challenges encountered during text preprocessing conducted on the two media articles using Python.

Conclusion. In this paper, we present a short description of challenges we encountered while conducting analysis, and the possible solutions we came up with to overcome the difficulties existing in NLP for Chinese and Lao.

References:

1. Chenglei Si, Zhengyan Zhang, et al.; Sub-Character Tokenization for Chinese Pretrained Language Models. *Transactions of the Association for Computational Linguistics* 2023; 11 469–487. https://doi.org/10.1162/tacl_a_00560
2. Zhang, Q., Xue, C., Su, X. et al. Named entity recognition for Chinese construction documents based on conditional random field. *Front. Eng. Manag.* 10, 237–249 (2023). https://doi.org/10.1007/s42524-021-0179-8
3. Enfield, N. J. (2020). Word in Lao. In A. Y. Aikhenvald, R. M. W. Dixon, & N. White (Eds.), *Phonological word and grammatical word: A cross-linguistic typology* (pp. 176–212). Oxford University Press. https://doi.org/10.1093/oso/9780198865681.001.0001