Zhou Y, Zheng XQ, Huang XJ. Chinese named entity recognition augmented with lexicon memory. JOURNAL OF COMPUTER SCIENCE AND TECHNOLOGY 38(5): 1021−1035 Sept. 2023. DOI: 10.1007/s11390-021-1153-y

Chinese Named Entity Recognition Augmented with Lexicon Memory

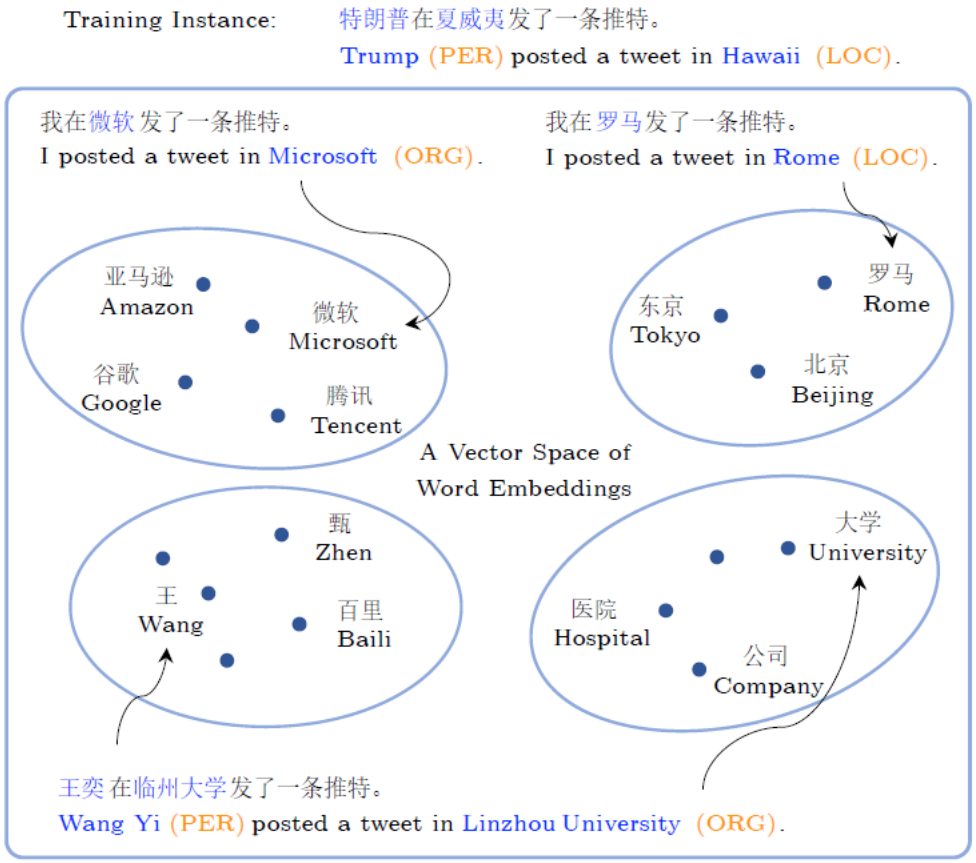
Yi Zhou (周奕), Xiao-Qing Zheng* (郑骁庆),
Xuan-Jing Huang (黄萱菁)

School of Computer Science, Fudan University

Objective

- Named entity recognition (NER) has been a long-standing challenge for the natural language processing community.
- Inspired by the concept of ***content-addressable retrieval*** from the cognitive science, we propose a novel fragment-based Chinese NER model augmented with ***lexicon-based memory***.
- The lexicon-based memory is built to help generate ***position-dependent features***, including prefix and suffix, and deal with the problem of ***out-of-vocabulary words***.
- Experimental results show that the proposed model, called **LEMON**, achieved state-of-the-art on four different data sets.

Motivation



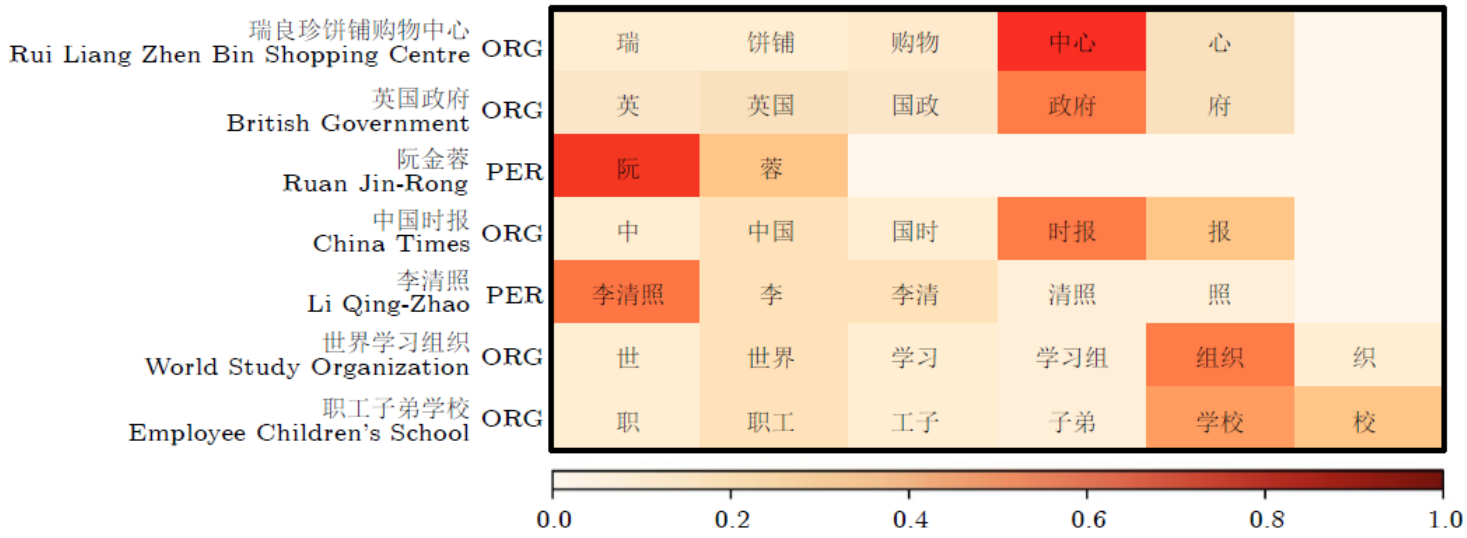
Illustrative instances demonstrate that the position-dependent features can benefit the NER task. For the instances involving "Microsoft" and "Rome", the semantic information of the words alone is enough for the recognition because their word embeddings are close to those of their kinds. Conversely, lesser-known individuals such as "Wang Yi" or organizations like "Linzhou University" often encounter out-of-vocabulary issues. The position-dependent information (e.g., the prefix "Wang" or the suffix "University") could be useful for the recognition.

Method

- We present a ***fragment-based*** model for Chinese NER augmented with a ***lexicon-based memory*** which combines features at different levels of granularity.
- The fragment-based model conforms to the way how human beings recognize entity names, consisting of three major submodules: a ***character encoder*** that imitates the process of scanning each character in an input sentence to grasp the global semantics, a ***fragment-encoder*** that simulates the procedures of reading a word or a phrase, and a ***memory*** that stores plenty of words that have ever seen. A ***ranking algorithm*** is used to determine whether a fragment is a valid name and which category it belongs to by taking its ***prefix***, ***suffix***, and ***infix*** features into account.

Results

- Experimental results show that the proposed model achieved ***state-of-the-art*** results on four different benchmark datasets.
- The ***source code*** of LEMON can be downloaded from the website of “github.com/dugu9sword/LEMON”.



An heat map illustrates which words will be given more weights calculated by the attention operations over the lexicon memory, and show that LEMON can learn to assign more weights on the keywords particularly informative for NER.

Conclusions

- Observing that Chinese entity names are usually formed in some distinct patterns and the features derived from their ***prefix*** and ***suffix*** are particularly useful to identify them, a ***fragment-based*** model augmented with ***position-dependent*** features derived from a lexicon has been proposed for the Chinese NER task.
- To address the problem of ***out-of-vocabulary words*** in NER, a ***lexicon-based memory*** is designed to produce such position-dependent features.
- The experimental results showed that the model using position-dependent features and lexicon-based memory achieved ***state-of-the-art*** performance with an increase in the *F*1-score up to 3.2% over the existing models on four different NER datasets.