Supplementary Material for "A Neural Multi-digraph Model for Incorporating Gazetteers in Chinese NER"

Ruixue Ding[1], **Pengjun Xie**[1], **Xiaoyan Zhang**[2], **Wei Lu**[3], **Linlin Li**[1] and **Luo Si**[1]
[1]Alibaba Group
[2]Beihang University, China
[3]Singapore University of Technology and Design
{ada.drx,chengchen.xpj,linyan.lll,luo.si}@alibaba-inc.com
xiaoyan.loic@gmail.com, luwei@sutd.edu.sg

Abstract

We present the e-commerce dataset information as well as gazetteers used in our model. The details of experiments are further discussed.

1 E-commerce Dataset

The E-commerce dataset is created by crawling and annotating product titles from the Taobao which is a Chinese e-commerce site with various types of products. Entity types including the Product name and the Brand name. Details of this dataset are shown in 1 and in 2.

Entity Number	Product	Brand
Train	10479	1630
Test	1345	222
Dev	1340	200

Table 1: The Entity Information

	Utterances	Tokens	Avg. Tokens
Train	3989	2956	29.9
Test	498	1706	29.5
Dev	500	1685	29.8

Table 2: Statistics of Dataset

2 Gazetteers

For general gazetteers, we collect gazetteers of 4 categories (PER, GPE, ORG, LOC). Each category has 3 gazetteers with different sizes, selected from multiple sources including Sougou (https://pinyin.sogou.com/dict/), HanLP (https://github.com/hankcs/HanLP) and Hankcs (http://www.hankcs.com/nlp/corpus). Sougou is a popular Chinese IME with a crowd source platform containing a huge number of gazetteers. HanLP is a widely used open-source Chinese NLP toolkit with many lexicons provided. Hankcs provides collection of lexicons of a ten million level volume.

For domain-specific gazetteers, We collect a list of person names from Weibo which is a Chines microblog site. The gazetteers in the e-commerce domain are obtained by crawled product catalogues from Taobao.

3 Experimental Details

3.1 Hyper-parameter tuning

As shown in Table 3, parameters of NCRFPP are tuned on the OntoNotes development set by grid-search without gazetteers. We setup our model and compared models with the same configuration. The parameters of graph embedding are tuned on the OntoNotes development set by grid-search with one ORG gazetteer added.

3.2 Models for comparison

Wang et al. (2018) propose detailed description for constructing the following methods. We follow the same constructing method as them. These methods are the same as (Qi et al., 2019; Chiu and Nichols, 2016).

N-gram Given the input sentence S with characters $c_1 \dots c_n$, the feature f_{c_i} of c_i is composed of 0-1 vectors (i.e., each entry of such vectors is either 0 or 1) for forward N-grams segments (e.g., $c_i c_{i+1}, c_i c_{i+1} c_{i+2}, \dots$) and 0-1 vectors for backward N-grams segments (e.g., $c_i c_{i-1}, c_i c_{i-1} c_{i-2}, \dots$). The 0-1 vector indicates whether the segment can be found in gazetteers of a certain category (PER, GPE, ORG, LOC). For example, if $c_i c_{i+1}$ can be found in a PER gazetteer and a ORG gazetteer, its 0-1 vector should be

Parameter	Value	Parameter	Value
Char emb size	200	Learning rate	0.001
Bigram emb Size	200	Batch size	10
LSTM hidden	600	Graph state	300
LSTM layers	2	T steps	2

Table 3: Hyper-parameter values

$[1, 0, 1, 0]$. Finally, f_{c_i} is the concatenation of all these 0-1 vectors.

PIET Given a sentence X and a gazetteer G, we first select non-overlapping matches entities in segment X by maximizing the total number of matched tokens in X. Then each character x_i is labeled as the gazetteer of the entity which x_i belongs to. The feature can be further represented in the format of one-hot encoding or feature embedding.

PDET PIET feature only considers the type of the entity which a character belongs to. Different from PIET feature, PDET feature also takes the position of a character in an entity into account: If the character is merely a single-character entity, we add a flag S before the PIET feature. Otherwise, for the first character of an entity, we add a flag B before the PIET feature; For the last character of an entity, we add a flag E before the PIET feature; For the middle character(s) of an entity, we add a flag I before the PIET feature. Similar to PIET feature, PDET feature can also be represented in the format of one-hot encoding or feature embedding

References

Jason P C Chiu and Eric Nichols. 2016. Named Entity Recognition with Bidirectional LSTM-CNNs. In *Transactions of the Association for Computational Linguistics (TACL)*, volume 4, pages 357–370.

Zhang Qi, Liu Xiaoyu, and Fu Jinlan. 2019. Neural Networks Incorporating Dictionaries for Chinese Word Segmentation. In *Proceedings of the AAAI Conference on Artificial Intelligence (AAAI)*.

Qi Wang, Yuhang Xia, Yangming Zhou, Tong Ruan, Daqi Gao, and Ping He. 2018. Incorporating Dictionaries into Deep Neural Networks for the Chinese Clinical Named Entity Recognition. *arXiv preprint arXiv:1804.05017*.