Supplemental Material of Unsupervised Neural Word Segmentation for Chinese via Segmental Language Modeling

Zhiqing Sun
Peking University
1500012783@pku.edu.cn

Zhi-Hong Deng
Peking University
zhdeng@pku.edu.cn

A Rule-based Post-processing Module

We apply the following rules on the results of SLM to introduce the empirical ad hoc guideline for each benchmarks. Specifically, we use the following two rules:

A.1 Splitting Rule

Figure 1 illustrate the splitting rule we use. In this rule, we split the "Splitted Character" with both consecutive character before it and after it, except the exception cases. For example, if we apply the splitting rule to the following sentences

- 我们的 心 连 心
- 我 即将 把你 将军

We will have

- 我 们 的 心 连 心
- 我 即将 把 你 将 军

Splitted Character	Exception
的, 和, 也, 都	
了	了解, 为了, 除了
与	与其, 与否, 参与
就	成就, 就要
在	现在, 正在, 存在, 所在, 在于
很	很多, 很难, 很快
将	即将, 必将, 将来
你, 我, 他, 她, 它	你们, 他们, 其他, 其它, 它们, 她们, 我们, 我国, 自我
要	主要, 需要, 要求, 重要, 只要, 还要
这	这里, 这次, 这样, 这种, 这是, 这些, 这个
上	上午, 上年, 上海, 上市, 以上

Figure 1: The splitting rule we use for post-processing.

A.2 Date Rule

It can be found that the date is always regarded as the same word. Therefore, we concatenate the characters that form a date. For example, given

- 2018 年 10月 6号

We transform it into

- 2018年10月6号