《中文深度学习分词是否有必要？》论文报告

张佳敏

目录

Introduction

- 李纪为博士的论文在讲到分词对于中文深度学习是否具有必要性时，提到中文分词相对于英文直接利用空格分词而言是比较麻烦的，其中两种中文分词的模型，word-based models 由于词分布的稀疏性会过度拟合导致大量OOV的产生，分词后语料库太大，使其分词效果并不如char-based models。
- 作者在论文中提到了很多用来得出此结论而做的实验：
  - 语言建模
  - 机器翻译
  - 语义匹配
  - 文本分类
- 并且作者从以下三个角度分析了word-based models表现的比char-based models好的原因：
  - Data Sparsity
  - Out-of-Vocabulary Words
  - Overfitting
  - Visualization

Introduction

- 在Introduction这一部分，作者讲述了中文英文因为单词间隔的原因导致的分词的不同，并用Jieba 对CTB进行分词
- 作者的实验结果显示，使用Jieba分词对CTB进行切词，共得到50266个不同词汇，其中仅出现一次的词语有24458个，占词语比例的48.7%，占总语料库的4.0%,
- 由此可见，很多词出现的频率是很低的，但是这些词却占据了总词数很大的比例，在语料库的占比又特别小。数据稀疏易导致过拟合问题，很多词语会被处理为OOV，进而会影响模型训练效果。

bar	# distinct	prop of vocab	prop of corpus
∞	50,266	100%	100%
4	38,889	77.4%	10.1%
1	24,458	48.7%	4.0%

实验:Language Modeling

- 数据集：CTB6.
  - 作者将数据集按照training, validation 和 test 分为三部分， 分别占比80%, 10%, 10%
- 模型：char-based model 、word- based model 、LSTMs
- 实验结果：

model	dimension	ppl
word	512	199.9
char	512	193.0
word	2048	182.1
char	2048	170.9
hybrid (word+char)	1024+1024	175.7
hybrid (word+char)	2048+1024	177.1
hybrid (word+char)	2048+2048	176.2
hybrid (char only)	2048	171.6

实验:Language Modeling

- 在语言建模这一任务中，是通过给定一定的信息预测后续的词语。数据集使用的是CTB6 来对比两个模型的效果的。在对比了不同纬度下word、char和混合模型的模型效果后，结果显示，维度在512时差距并不明显，但在2048的时候，ppl达到最优的结果差距明显。与此同时，在 CWS 包和 LTP 包下也进行试验，结果相同。
- 作者还对混合模型的效果进行了比较，对比混合模型与char only，结果发现在嵌入词向量之后，效果反而不如char only 模型。

• 实验:Machine Translation

- 语料库： LDC corpora
  - 语料库中包含了1.25M的句子对
- 设置：设置是按照 [1] 中论文设置的，用了30000个英文单词和27500个中文单词。
- 模型： char-based models with word-based models
  - 比较是在SEQ2SEQ +attention 的框架下进行的
  - 训练过程中用到了seq2seq和词袋模型【1】

论文地址【1】：https://www.researchgate.net/publication/325142631_Bag-of-Words_as_Target_for_Neural_Machine_Translation

• 实验:Machine Translation

- 作者对中译英和英译中都进行了实验来对比不同模型下的机器翻译效果。
- 表四的中译英和表五的英译中结果显示，无论是中译英还是英译中，char模型都比word模型的效果好。

TestSet	Mixed RNN	Bi-Tree-LSTM	PKI	Seq2Seq +Attn (word)	Seq2Seq +Attn (char)	Seq2Seq (word) +Attn+BOW	Seq2Seq (char) +Attn+BOW
MT-02	36.57	36.10	39.77	35.67	36.82 (+1.15)	37.70	40.14 (+0.37)
MT-03	34.90	35.64	33.64	35.30	36.27 (+0.97)	38.91	40.29 (+1.38)
MT-04	38.60	36.63	36.48	37.23	37.93 (+0.70)	40.02	40.45 (+0.43)
MT-05	35.50	34.35	33.08	33.54	34.69 (+1.15)	36.82	36.96 (+0.14)
MT-06	35.60	30.57	32.90	35.04	35.22 (+0.18)	35.93	36.79 (+0.86)
MT-08	–	–	24.63	26.89	27.27 (+0.38)	27.61	28.23 (+0.62)
Average	–	–	32.51	33.94	34.77 (+0.83)	36.51	37.14 (+0.63)

Table 4: Results of different models on the Ch-En machine translation task. Results of Mixed RNN (Li et al., 2017), Bi-Tree-LSTM (Chen et al., 2017a) and PKI (Zhang et al., 2018) are copied from the original papers.

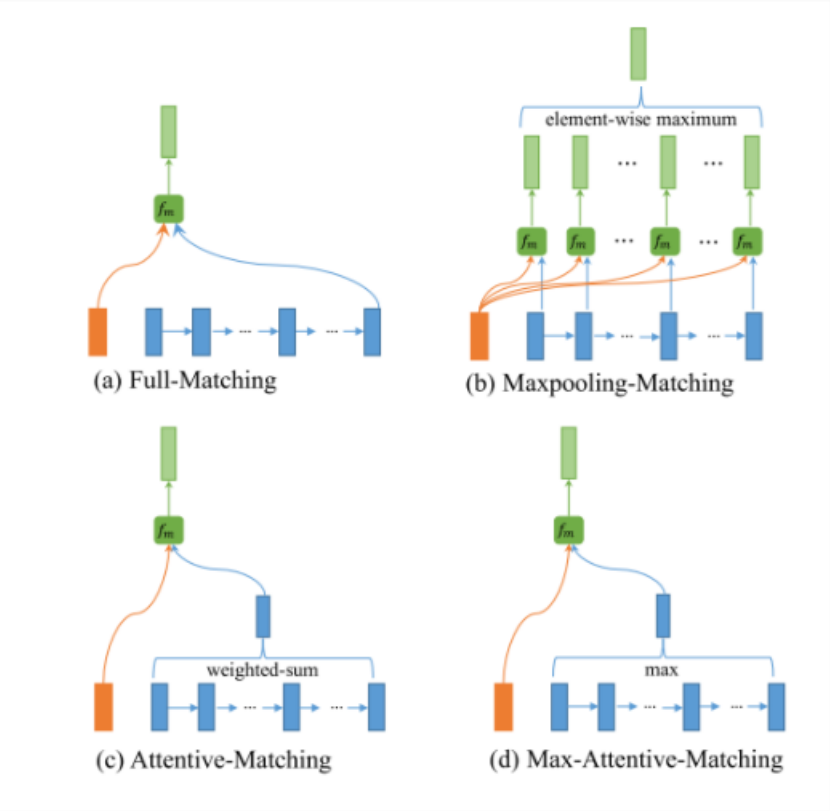
TestSet	Seq2Seq +Attn (word)	Seq2Seq +Attn (char)	Seq2Seq +Attn+BOW	Seq2Seq (char) +Attn+BOW
MT-02	42.57	44.09 (+1.52)	43.42	46.78 (+3.36)
MT-03	40.88	44.57 (+3.69)	43.92	47.44 (+3.52)
MT-04	40.98	44.73 (+3.75)	43.35	47.29 (+3.94)
MT-05	40.87	42.50 (+1.63)	42.63	44.73 (+2.10)
MT-06	39.33	42.88 (+3.55)	43.31	46.66 (+3.35)
MT-08	33.52	35.36 (+1.84)	35.65	38.12 (+2.47)
Average	39.69	42.36 (+2.67)	42.04	45.17 (+3.13)

Table 5: Results on the En-Ch machine translation task.

ers are the other way around. For LCQMC (Liu ... sentence. We conclude that the char-based model

• 实验：Sentence Matching

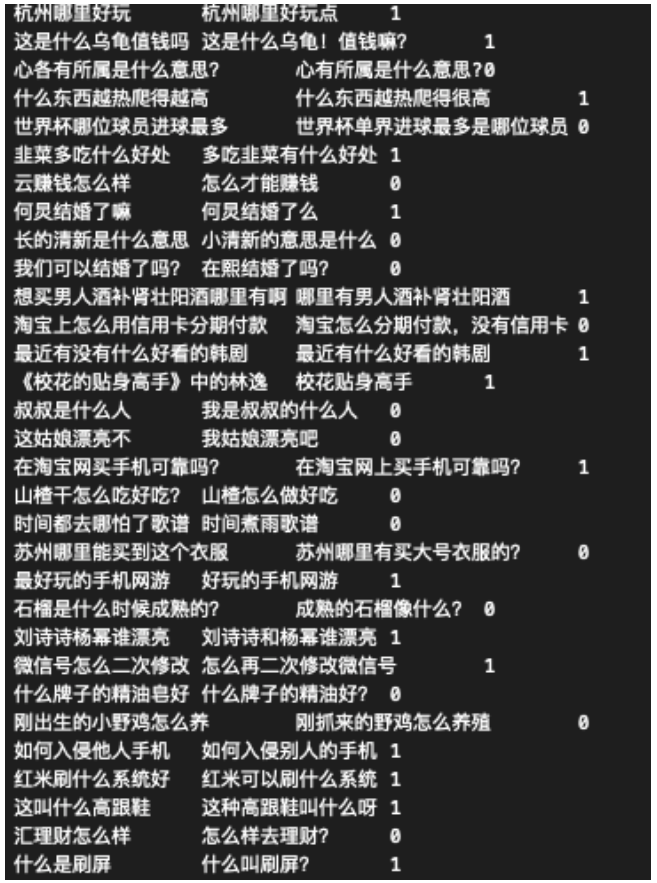
- 实验所基于的模型是BiMPM，论文中引用了王志国研究员的论文，故粗略读了一下论文，了解了一下模型。
  - 除此之外，模型还给了四种匹配策略，分别是full-matching、maxpooling-matching、attentive-matching和max-attentive-matchong，利用多视角余弦匹配函数可比较两个向量。



论文地址：https://arxiv.org/abs/1702.03814

• 实验：Sentence Matching

- 论文当中提到两个中文数据集是：BQ和LCQMC，于是尝试去找数据集，找到了LCQMC。
  - 与BQ数据集比较两个句子是否具有相同语义含义不同的是，LCQMC偏向于两个句子是否具有相同的意图。
  - 下载下来的数据集是三个txt文件，以1和0来代表两个句子含义或目的相同与不同。



杭州哪里好玩	杭州哪里好玩点	1
这是什么乌龟值钱吗	这是什么乌龟！值钱嘛?	1
心各有所属是什么意思?	心有所属是什么意思?0	
什么东西越热爬得越高	什么东西越热爬得很高	1
世界杯哪位球员进球最多	世界杯单界进球最多是哪位球员	0
韭菜多吃什么好处	多吃韭菜有什么好处	1
云赚钱怎么样	怎么才能赚钱	0
何炅结婚了嘛	何炅结婚了么	1
长的清新是什么意思	小清新的意思是什么	0
我们可以结婚了吗?	在熙结婚了吗?	0
想买男人酒补肾壮阳酒哪里有啊	哪里有男人酒补肾壮阳酒	1
淘宝上怎么用信用卡分期付款	淘宝怎么分期付款，没有信用卡	0
最近有没有什么好看的韩剧	最近有什么好看的韩剧	1
《校花的贴身高手》中的林逸	校花贴身高手	1
叔叔是什么人	我是叔叔的什么人	0
这姑娘漂亮不	我姑娘漂亮吧	0
在淘宝网买手机可靠吗?	在淘宝网上买手机可靠吗?	1
山楂干怎么吃好吃?	山楂怎么做好吃	0
时间都去哪怕了歌谱	时间煮雨歌谱	0
苏州哪里能买到这个衣服	苏州哪里有买大号衣服的?	0
最好玩的手机网游	好玩的手机网游	1
石榴是什么时候成熟的?	成熟的石榴像什么?	0
刘诗诗杨幂谁漂亮	刘诗诗和杨幂谁漂亮	1
微信号怎么二次修改	怎么再二次修改微信号	1
什么牌子的精油皂好	什么牌子的精油好?	0
刚出生的小野鸡怎么养	刚抓来的野鸡怎么养殖	0
如何入侵他人手机	如何入侵别人的手机	1
红米刷什么系统好	红米可以刷什么系统	1
这叫什么高跟鞋	这种高跟鞋叫什么呀	1
汇理财怎么样	怎么样去理财?	0
什么是刷屏	什么叫刷屏?	1

LCQMC数据集：https://github.com/pengming617/bert_textMatching/tree/master/data

• 实验：Sentence Matching

- 实验所基于的模型是BiMPM，论文中引用了王志国研究员的论文，故粗略读了一下论文，了解了一下模型。
  - 架构：对所要比较的两句话进行匹配预测处理，将匹配结果聚合到一个向量中，在最后再将所得的结果进行最终处理。主要核心在句子之间的匹配。
  - 模型框架如图所示。
  - 其中模型匹配层为模型的核心。这一层的目的是用其中一句话的time step的上下文向量去匹配其中另一句话的上下文向量。

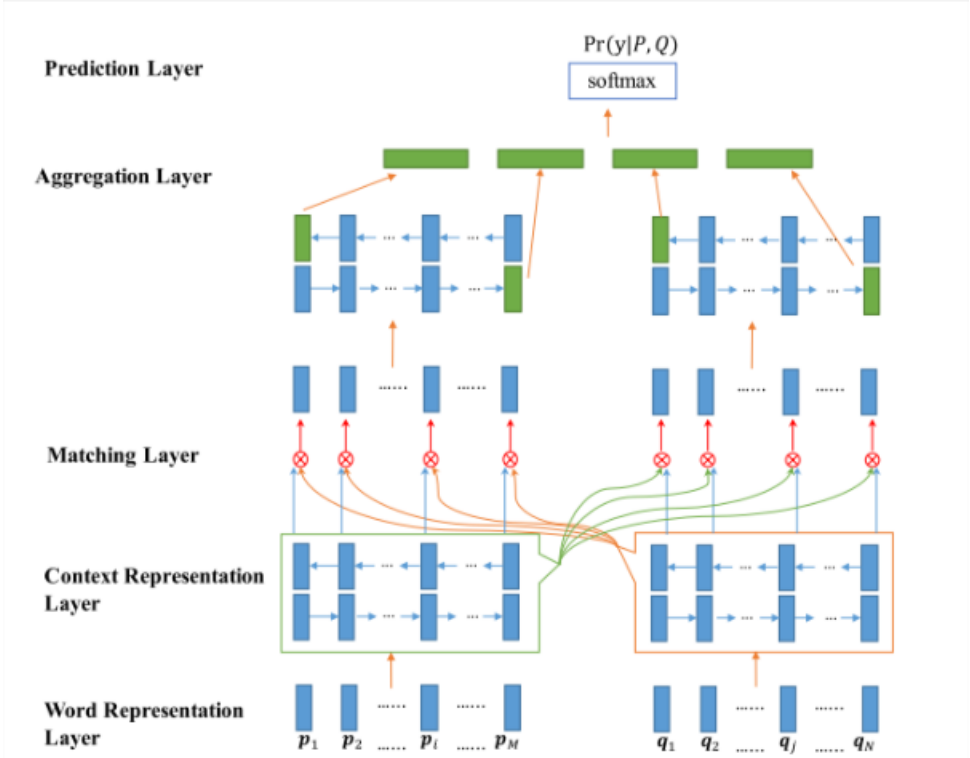


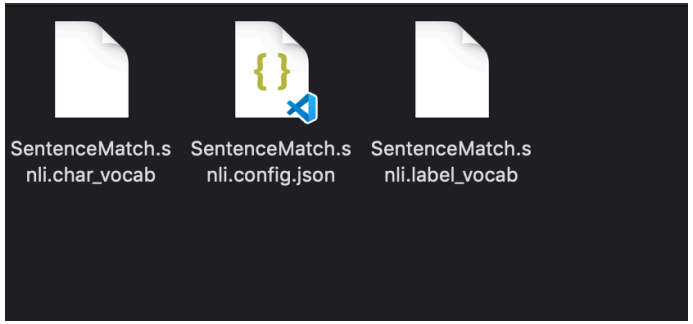
Figure 1: Architecture for Bilateral Multi-Perspective Matching (BiMPM) Model, where ⊗ is the multi-perspective matching operation described in sub-section 3.2.

论文地址：https://arxiv.org/abs/1702.03814

• 实验：Sentence Matching

- 在Github上找到了王志国研究员关于实现Sentence Matching的代码[1]，于是尝试去跑
  - 编辑好配置文件后使用数据集去训练，在训练中生成Testing所要用的log文件。

```
"train_path": "/Users/zhangjiamin/Desktop/BiMPM-master/snli/train.tsv",
"dev_path": "/Users/zhangjiamin/Desktop/BiMPM-master/snli/dev.tsv",
"word_vec_path": "/Users/zhangjiamin/Desktop/BiMPM-master/snli/wordvec.txt",
"model_dir": "/Users/zhangjiamin/Desktop/BiMPM-master/snli/logs",
"suffix": "snli",
```



【1】代码地址：https://github.com/zhiguowang/BiMPM

• 实验：Sentence Matching

- 在Github上找到了王志国研究员关于实现Sentence Matching的代码[1]，于是尝试去跑
  - 在测试过程中，由于训练过程可能是配置和路径没设置好，开始浪费了一些时间，最后test的时候还是没有成功实现，最后提示没有注册的内核，是系统原因未用GPU用的CPU跑的，之后会继续换台电脑跑一遍。

```
  File "/Users/zhangjiamin/Library/Python/2.7/lib/python/site-packages/tensorflow/python/client/session.py", line 1370, in _do_call
    raise type(e)(node_def, op, message)
tensorflow.python.framework.errors_impl.InvalidArgumentError: No OpKernel was registered to support Op 'CudnnRNNCanonicalToParams' used by
bi_lstm/CudnnRNNCanonicalToParams (defined at /Desktop/BiMPM-master/src/layer_utils.py:21) with these attrs: [num_params=16, T=DT_FLOAT, ir
"lstm", seed2=0, seed=0, dropout=0]
Registered devices: [CPU]
Registered kernels:
  <no registered kernels>

        [[Model/char_lstm/char_lstm_cudnn_bi_lstm/char_lstm_cudnn_bi_lstm/CudnnRNNCanonicalToParams]]
```

【1】代码地址：https://github.com/zhiguowang/BiMPM

• 实验：Text classification

- 利用不同的benchmarks 来对基于char和基于word的模型进行实验。

Table 6: Results on the LCQMC and BQ corpus.

Dataset	description	char valid	word valid	char test	word test
chinanews	1260K/140K/112K	91.81	91.82	91.80	**91.85** (+0.05)
dianping	1800K/200K/500K	78.80	78.47	**78.76** (+0.36)	78.40
ifeng	720K/80K/50K	86.04	84.89	**85.95** (+1.09)	84.86
jd_binary	3600K/400K/360K	92.07	91.82	**92.05** (+0.16)	91.89
jd_full	2700K/300K/250K	54.29	53.60	**54.18** (+0.81)	53.37

Table 7: Results on the validation and the test set for text classification.

- 使用双向 LSTM 模型对基于word 和基于char 的模型分别进行训练用于评测，表七结果显示除了China news 以外，基于char的模型得出的结果都要优于基于word的模型。

- 实验：Text classification

4. 文本分类

表八展示了模型基于对已有数据分布（源领域）的训练，学习新数据分布（目标领域）的能力。作者基于不同的情感分析数据库对两种模型进行了评测，结果发现，基于char的模型在领域适应能力更强且具有更好的表现。

train_dianping_test_jd		
model	acc	proportion of sen containing OOV
word-based	81.28%	11.79%
char-based	83.33%	0.56%

train_jd_test_dianping		
model	acc	proportion of sen containing OOV
word-based	67.32%	7.10%
char-based	67.93%	46.85%

分析:Data Sparsity

- 设置Frequency Bar来探究词语稀疏性对于两种模型的影响
  - 在Analysis这一部分当中，作者通过实验利用设置Frequency Bar来探究词语稀疏性对于两种模型的影响。结果显示，两种模型表现得最好的时候，词规模是差不多的，但是Frequency Bar却相差不小，对于word的模型来说，低频词的学习难度是比较高的，相应的准确度也远不如基于model的模型。
  - 对于已知的数据集，若要语义学习表现较好，模型是需要足够的word/character。对于word-based model，由于数据稀疏性是很难达到要求的

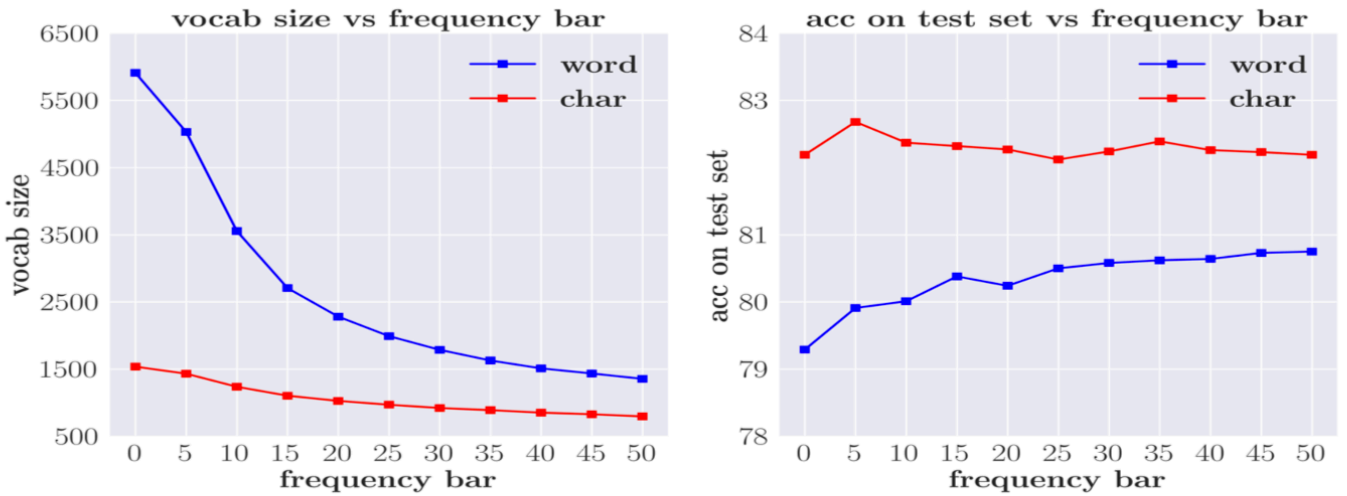


Figure 2: Effects of data sparsity on the char-based model and the word-based model.

分析:Out-of-Vocabulary Words

- OOV对模型效果的影响。
  - 作者在论文中阐述影响word model的另一大因素是OOV，但降低Frequency bar虽然会减少OOV,但是会出现数据稀疏问题。作者在实验中基于不同的Frequency bar 分别移除对应包含OOV的句子，结果显示，随着 frequency bar的增加，两种模型的效果差距在减小，不过char-based model始终优于word-based model.

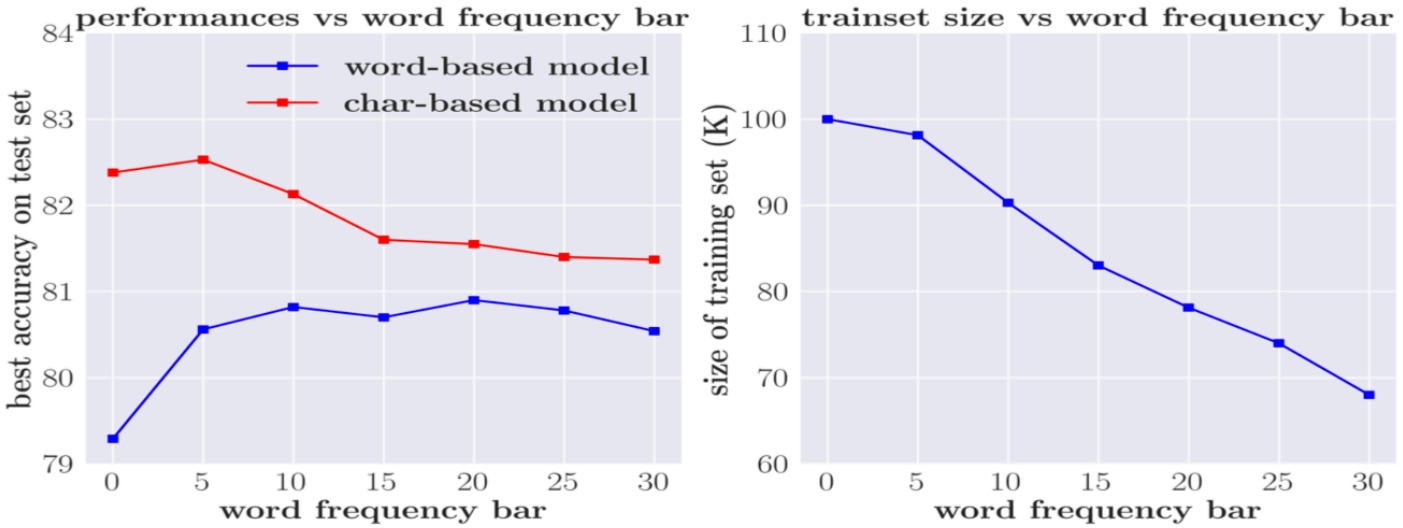


Figure 4: Effects of removing training instances containing OOV words.

• 分析: Overfitting

- Data Sparsity导致基于单词的模型需要学习更多的参数，因此更容易过度拟合。
- 使用BQ数据集 [1] 进行实验验证，基于字的模型需要比基于字符的模型虚啊哟更大的dropout rate以达到更好的实验结果

Figure 1: Effects of dropout rates on the char-based model and the word-based model.

【1】地址： https://www.aclweb.org/anthology/D18-1536

论文总结

- 李纪为博士在这项研究探究了中文分词必要性的问题，并从四类NLP任务的实验中得出char-based model是优于word-based model的这个结论。
- 作者将word模型效果不佳的原因归结于词分布的稀疏性导致更多OOV的产生、过拟合以及领域转化能力差。

感谢漆老师的观看！