8.12

阅读两遍李纪为博士的论文，总结：

中文分词比较麻烦，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，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结论的实验：

1. 用Jieba分词对CTB进行分词

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%

实验结果显示，仅出现一次的词有24，458个，占词语比例的48.7%，占总语料库的4%。很多词语会被处理为OOV，会影响模型训练与结果。

2.Language modeling

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

在不同维度下，对比单独的word、char和混合模型的效果，发现char模型始终优于word模型

3，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。

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

结果显示无论是中译英还是英译中都是char模型优于word模型。

4.Text Classification （文本分类）

作者利用不同的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.

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%

表7得出实验结论：除了Chinanews以外，基于char的模型得出的结果都要优于基于word的模型。
表8得出实验结论：基于char的模型具有更强的领域适应能力。

5.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

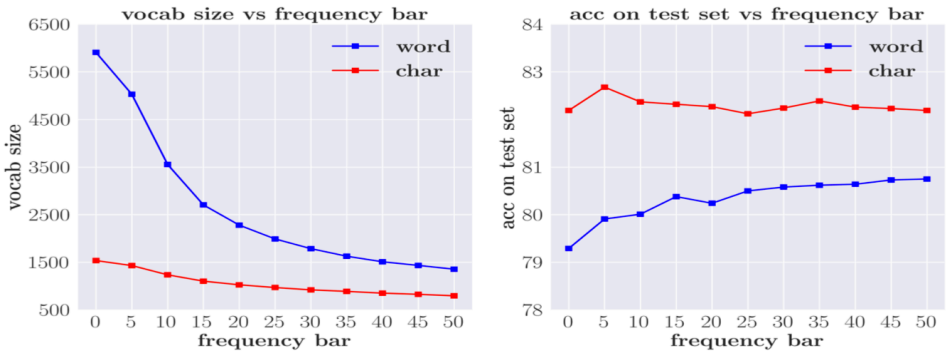

Figure 2: Effects of data sparsity on the char-based model and the word-based model.

结果显示，两种模型表现得最好的时候词规模和词频是差不多的，只不过对于基于word的模型来说，对于低频词的学习难度是比较高的。

6.OOV对模型效果的影响。

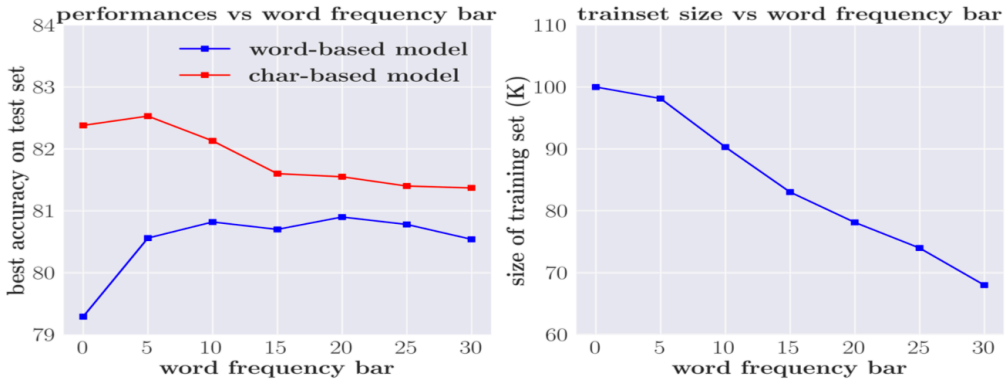

Figure 4: Effects of removing training instances containing OOV words.

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.

作者在这篇论文探究了中文分词必要性的这个问题，从四类自然语言处理任务的实验中得出char-based model是优于word-based model的，并认为word-based model结果不佳的原因是数据稀疏和OOV导致的。

接下来的任务：
对论文中提到的实验尽量进行复现，并寻找相关论文继续深入了解中文分词的必要性。