Is Word Segmentation Necessary for Deep Learning of ChineseRepresentations

1
Conclusion

Through direct comparisons between these two types of models, we find that char- based models consistently outperform word- based models.

1
Reasons

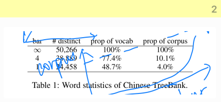
it is because word-based models are more vulnerable to data sparsity and the presence of out-of-vocabulary (OOV) words, and thus more prone to overfitting.

1
1 Introduction

1

Firstly, word data sparsity inevitably leads to overfitting and the ubiquity of OOV words limits the model's learning capacity

2

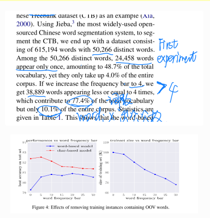


bar	# distinct	prop of vocab	prop of corpus
∞	50,266	100%	100%
4	38,889	77.4%	10.1%
1	24,458	48.7%	4.0%

Table 1: Word statistics of Chinese TreeBank.

corpus

1



nese Treebank dataset (CTB) as an example (Xia, 2000). Using Jieba,[3] the most widely-used open-sourced Chinese word segmentation system, to segment the CTB, we end up with a dataset consisting of 615,194 words with 50,266 distinct words. Among the 50,266 distinct words, 24,458 words appear only once, amounting to 48.7% of the total vocabulary, yet they only take up 4.0% of the entire corpus. If we increase the frequency bar to 4, we get 38,889 words appearing less or equal to 4 times, which contribute to 77.4% of the total vocabulary but only 10.1% of the entire corpus. Statistics are given in Table 1. This shows that the word-based

First experiment

> 4

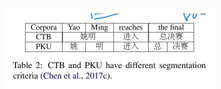
performances vs word frequency bar
word-based model
char-based model
best accuracy on test set
word frequency bar
trainset size vs word frequency bar
size of training set (K)
word frequency bar

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

2

Secondly, the state-of-the-art word segmenta- tion performance is far from perfect, the errors of which would bias downstream NLP tasks.

2



Corpora	Yao	Ming	reaches	the final	
CTB	姚明		进入	总决赛	
PKU	姚	明	进入	总	决赛

Table 2: CTB and PKU have different segmentation criteria (Chen et al., 2017c).

2

Thirdly, if we ask the fundamental problem of how much benefit word segmentation may provide, it is all about how much additional semantic infor- mation is present in a labeled CWS dataset.

2

the answer to this question remains unclear

2

Before neural network models became popular, there were discussions on whether CWS is nec- essary and how much improvement it can bring about.

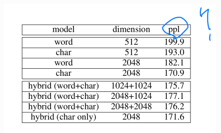
3
2 Related Work

3
3 Experimental Results

3

3.1 Language Modeling

3

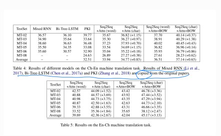


model	dimension	ppl
word	512	199.9
char	512	193.0
word	2048	182.1
char	2048	170.9
hybrid (word+char)	1024+1024	175.7
hybrid (word+char)	2048+1024	177.1
hybrid (word+char)	2048+2048	176.2
hybrid (char only)	2048	171.6

4

3.2 Machine Translation

5



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.

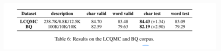
TestSet	Seq2Seq +Attn (word)	Seq2Seq +Attn (char)	Seq2Seq +Attn+BOW	Seq2Seq (char) +Attn+BOW
MT-02	42.57	44.09 (+1.52)	43.42	46.78 (+3.36)
MT-03	40.88	44.57 (+3.69)	43.92	47.44 (+3.52)
MT-04	40.98	44.73 (+3.75)	43.35	47.29 (+3.94)
MT-05	40.87	42.50 (+1.63)	42.63	44.73 (+2.10)
MT-06	39.33	42.88 (+3.55)	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.

4

3.3 Sentence Matching/ Paraphrase

6



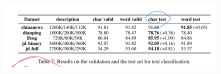
Dataset	description	char valid	word valid	char test	word test
LCQMC	238.7K/8.8K/12.5K	84.70	83.48	**84.43** (+1.34)	83.09
BQ	100K/10K/10K	82.59	79.63	**82.19** (+2.90)	79.29

Table 6: Results on the LCQMC and BQ corpus.

5

3.4 Text Classification

6



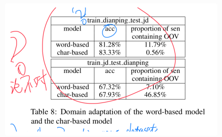
Dataset	description	char valid	word valid	char test	word test
chinanews	1260K/140K/112K	91.81	91.82	91.80	**91.85** (+0.05)
dianping	1800K/200K/500K	78.80	78.47	**78.76** (+0.36)	78.40
ifeng	720K/80K/50K	86.04	84.89	**85.95** (+1.09)	84.86
jd_binary	3600K/400K/360K	92.07	91.82	**92.05** (+0.16)	91.89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

6

Domain Adaptation Ability

6



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%

Table 8: Domain adaptation of the word-based model and the char-based model

6
4Analysis

6

4.1 Data Sparsity

7

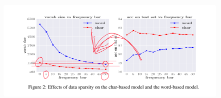
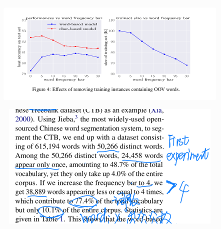


Figure 2: Effects of data sparsity on the char-based model and the word-based model.

7

4.2 Out-of-Vocabulary Words

8



performances vs word frequency bar
word-based model
char-based model
best accuracy on test set
word frequency bar
trainset size vs word frequency bar
size of training set (K)
word frequency bar

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

nese Treebank dataset (CTB) as an example (Xia, 2000). Using Jieba,[3] the most widely-used open-sourced Chinese word segmentation system, to segment the CTB, we end up with a dataset consisting of 615,194 words with 50,266 distinct words. Among the 50,266 distinct words, 24,458 words appear only once, amounting to 48.7% of the total vocabulary, yet they only take up 4.0% of the entire corpus. If we increase the frequency bar to 4, we get 38,889 words appearing less or equal to 4 times, which contribute to 77.4% of the total vocabulary but only 10.1% of the entire corpus. Statistics are given in Table 1. This shows that the word-based

First experiment

> 4

8

4.3 Overfitting

7

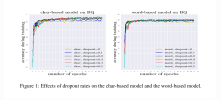


Figure 1: Effects of dropout rates on the char-based model and the word-based model.

8

4.4 Visualization

7

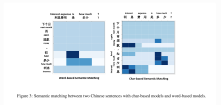


Figure 3: Semantic matching between two Chinese sentences with char-based models and word-based models.

8
5 Conclusion

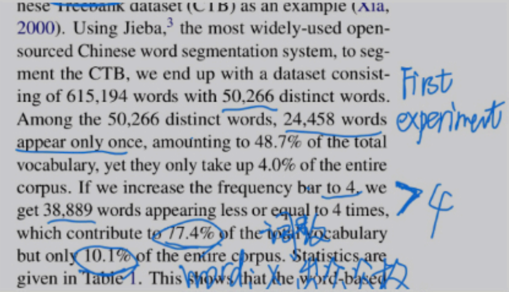
1. 目前很多在用 subword model 和 char-based 和 word-based 可以作一下对比（在不同任务上的）

2 分字可以用 transformer

3. 评价的标准还是不够了解.
例如: acc 是 accuracy 吗?
ppl 是 perplexity

4. CTB6.0 Chinese TreeBank.
segmented, POS targeted, bracketed Chinese corpus.
来自于 新闻. 文章杂志等

最后提到了对于OOV不同的frequency bar，实验结果先升后降.
因为 frequency bar 小，那么 infrequent 但接近 frequency bar
的词会被归于词库中，亦不视为是 OV。同时这种词的进入，
会加重 data sparity 的问题，使很多有可能有词库代表性的词，
即（频率低 有特征 ）的词，~~进入词库~~ 无法进入词库，影响参数拟合效果。



nese Treebank dataset (CTB) as an example (Xia, 2000). Using Jieba,[3] the most widely-used open-sourced Chinese word segmentation system, to segment the CTB, we end up with a dataset consisting of 615,194 words with 50,266 distinct words. Among the 50,266 distinct words, 24,458 words appear only once, amounting to 48.7% of the total vocabulary, yet they only take up 4.0% of the entire corpus. If we increase the frequency bar to 4, we get 38,889 words appearing less or equal to 4 times, which contribute to 77.4% of the total vocabulary but only 10.1% of the entire corpus. Statistics are given in Table 1. This shows that the word-based

First experiment

> 4

词汇

word:x 出现次数