**综述：**

本文章探讨的问题是：在基于深度学习的中文语言处理（deep learning-based Chinese

Natural Language Processing）中，将文本分割（segmentation）成word是否必要（因为分割成word常常被认为是必有的第一步操作）

在四个端到端NLP基准测试任务中对**依赖于词neural word分割的基于神经词的模型**进行基准测试，对比**不涉及词分割的基于神经字符neural char的模型**进行基准测试来说明问题：language modeling, machine translation, sentence matching/paraphrase and text classification

结果：char based(未分割)的结果一致高于word based（有分割）

结果的原因：word-based 模型在面对data sparsity和out-of-vocabulary (OOV)的情况下更脆弱，因而更容易过拟合

**segmentation处理基本方法：**

segmented word 被表示为定长向量然后投入deep learning model中处理

数据库：Chinese Treebank dataset(widely used)

Word segmentation工具：Jieba(widely used)

**word-based model的缺点：**

1. Word-based data sparsity导致过拟合和OOV（Zipf定律，很多Chinese word的频率极低） 有表格证明sparsity
2. state-of-the-art word segmentation performance is far from perfect 原因：Chinese word Segmentation (CWS)是复杂任务，Chinese word 的边界很模糊
3. CWS labeled dataset很小，且涉及领域很少

**先前研究中指出的CWS对NLP任务提升影响较小（例子）：**

1. 在information retrieval，CWS并没有一直比未分割的模型表现的好，在那些表现好的例子当中word-based model 和char-based model表现相近
2. phrase-based machine translation，CWS仅表现出可有可无的提升
3. text classification，Liu et al. (2007)得出结论CWS并非必要
4. 在cognitive science中，根据eye movement data,中文阅读者的注意力不是频繁地落在单词的中心，表明char应该是阅读理解的基本单位
5. 中文读者读有无空格分割文本的速率相同

**实验详述：**

在四个任务中比对word segmentation在deep learning-based Chinese NLP中的影响

1. **Language Modeling**

任务描述：predicting the upcoming word given previous context representation

语料库：Chinese Tree-Bank 6.0 (CTB6)

数据集：0.8training，0.1validation，0.1test

Segment工具：jieba

Hyperparameter：vector dimensionality, learning rate, batch size

区分设定：context representation分别是char-based和 word-based

补充对照组：hybrid model

数据结果：（table 3），char-based model ≈hybrid model （char only）＞hybrid model （word+char）＞the word-based model (ppl评分)

分析：hybrid model ＞the word-based model：说明char已经将所有语义信息编码

char-based model ≈hybrid model （char only）：说明WS没有提供新信息

结论：language modeling task on CTB, WS未提供additional performance boost

1. **Machine Translation**

设定：standard Ch-En setting,结果比对Ch-En and the En-Ch settings.

数据集：training：1.25M sentence pairs extracted from the LDC corpora .Validation：NIST 2002 and the models are evaluated on NIST 2003,2004, 2005, 2006 and 2008

比对架构：standard framework of SEQ2SEQ +attention

结果:（table4,5）

中英翻译:在vanilla SEQ2SEQ +attention model下：char-based model表现更好

英中翻译：char-based model表现更好，并且比中英翻译表现更佳（原因ch-en中word-based和char-based仅出现在encode阶段，而en-ch中encode和decode均有出现）

1. **Sentence Matching/Paraphrase**

任务描述：根据是否为相同含义来设定标签

数据集BQ：120,000 Chinese sentence pairs, associated with a label表示是否有相同语义

数据集LCQMC：区分两句子是否有相同的intension

任务可描述为预测正确标签（对比char-based和word-based）

对word-based model处理：Jieba来segment

模型：bilateral multi-perspective matching model (BiMPM)

设定：standard settings proposed by BiMPM

结果：（table 6）在两个数据剧集中，char-based的表现均优于wod-based

1. **Text Classification**

数据集：ChinaNews, Ifeng, JD Full, JD binary, Dianping

任务描述： trained the word-based version and the char-based version of bi-directional LSTM models to solve this task

结果：（table 7）：仅有ChinaNews中char-based低于word-based，但是差距很小。其他数据集中char-based都远高于word-based

**分析：**

1. **Data Sparsity**

词频阈值和词库大小相关（低于词频阈值的就删去），对char-based和word-based，阈值为0时均表现不佳（原因：低频词汇很难学习语义）

对char模型，最佳表现的阈值为5。而对word-based模型，最佳表现得阈值是50。这就是说为了获得最好的语义学习，word-based模型由于data sparsity的存在更难达到要求。

1. **OOV words**

其中一个word-based模型表现劣势的可能解释就是有太多OOV

验证方法：对不同的阈值，我们去除掉训练集中所有带OOV的句子。结果在figure 4.

可以发现随着阈值提高，两模型的结果逐渐接近。起初时，char的表现甚至有上升说明char不太受oov影响，但是随着阈值逐渐升高，word-based的表现越来越好，说明word-based受到OOV影响较大

1. **Overfitting**

Data sparsity会导致overfitting

在BQ数据库中进行试验：为了得到最佳结果word-based（0.5）的dropout rate要高于char-based（0.3），说明overfitting在word-based模型当中是个严重问题

1. **visualization**

figure3：char-based model更容易在两句子中映射，而word-based更难做到

**结论**

根据以上试验和分析：word-based model更劣势，原因是sparseness of word distributions，从而导致OOV，overfitting以及lack of domain generalization ability.