Jieba 分词、海量分词、深度学习算法分词准确度比较

一、测试语料库

1. 宾州中文树库（CTB 7.0）
   包含人民日报、香港新闻电讯和台湾期刊等语料。
2. 台北中央研究院语料库（AS）（繁体）
   包含部分台湾报刊、书籍等语料。
3. 香港城市大学中文语料库（CityU）（繁体）
   包含香港特区新闻报纸、出版刊物等语料。
4. 微软亚研院中文语料库（MSR）
5. 北京大学中文语料库（PKU）

二、测试说明

1. 测试采用国际中文分词测评活动（International Chinese Word Segmentation Bakeoff）提供的测评方法。
2. 海量分词通用版使用的字典只包含简体字，所以在 AS、CityU 语料库上测试时，先转化为简体字然后进行测试。
3. *深度学习算法在各语料库上进行测试所使用的模型是在宾州中文树库（CTB 7.0）上训练所得。*

三、测试结果

1. CTB 7.0

	Recall	Precision	F1-score
Jieba	0.797	0.842	0.819
海量分词	0.912	0.912	0.912
深度学习	**0.951**	**0.948**	**0.949**

2. AS

	Recall	Precision	F1-score
Jieba	0.737	0.740	0.738
海量分词	0.891	0.869	0.880
深度学习	**0.899**	**0.887**	**0.893**

3. CityU

	Recall	Precision	F1-score
Jieba	0.735	0.748	0.742
海量分词	**0.894**	0.863	**0.878**
深度学习	0.880	**0.873**	0.876

4. MSR

	Recall	Precision	F1-score
Jieba	0.812	0.817	0.815
海量分词	**0.905**	**0.875**	**0.890**
深度学习	0.855	0.805	0.829

5. PKU

	Recall	Precision	F1-score
Jieba	0.787	0.853	0.818
海量分词	**0.952**	**0.963**	**0.958**
深度学习	0.889	0.885	0.887

三、小结

深度学习算法在模型训练所使用的语料库（CTB7.0）上进行测试时，测试表现最优；因为 AS 和 CityU 语料库同 CTB7.0 均为新闻类语料，所以在 AS 和 CityU 语料库上的测试结果也同样较好，AS 语料库上三项测试分数均为最高，CityU 语料库上 Precision 分数最高、F1-score 与海量分词相近；在 MSR 和 PKU 语料库上测试表现仅次于海量分词，优于 Jieba 分词。