链滴

11 款开放中文分词引擎大比拼

作者：BosonNLP

原文链接：https://ld246.com/article/1447062320195

来源网站：链滴

在逐渐步入DT（Data Technology）时代的今天，自然语义分析技术越发不可或缺。对于我们每天交道的中文来说，并没有类似英文空格的边界标志。而理解句子所包含的词语，则是理解汉语语句的一步。汉语自动分词的任务，通俗地说，就是要由机器在文本中的词与词之间自动加上空格。

一提到自动分词，通常会遇到两种比较典型的质疑。一种质疑是来自外行人的：这件事看上去平凡之，好像一点儿也不“fancy”，会有什么用呢？另一种质疑则是来自业内：自动分词研究已经进行了年，而网上也存在各种不同的开放分词系统，但对于实际商用似乎也未见一个“即插即用”的系统。

那么，目前常见的开放分词引擎，到底性能如何呢？为了进行测试，我们调研了11款网上常见的并且开提供服务的分词系统，包括：

	分词系统	标识
1	BosonNLP	
2	IKAnalyzer	
3	NLPIR	
4	SCWS	
5	结巴分词	
6	盘古分词	
7	庖丁解牛	
8	搜狗分词	
9	腾讯文智	
10	新浪云	
11	语言云	

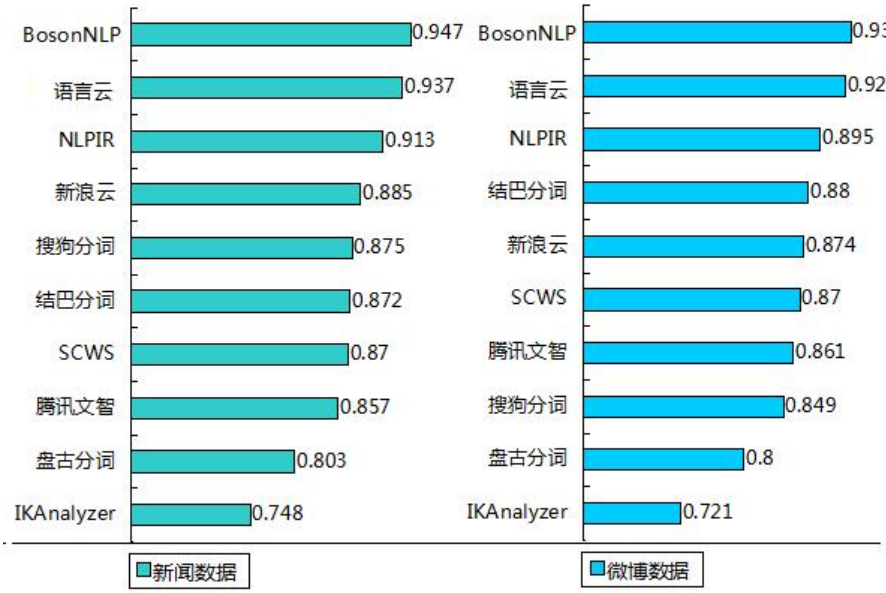
分词的客观量化测试离不开标注数据，即人工所准备的分词“标准答案”。在数据源方面，我们将测分为:

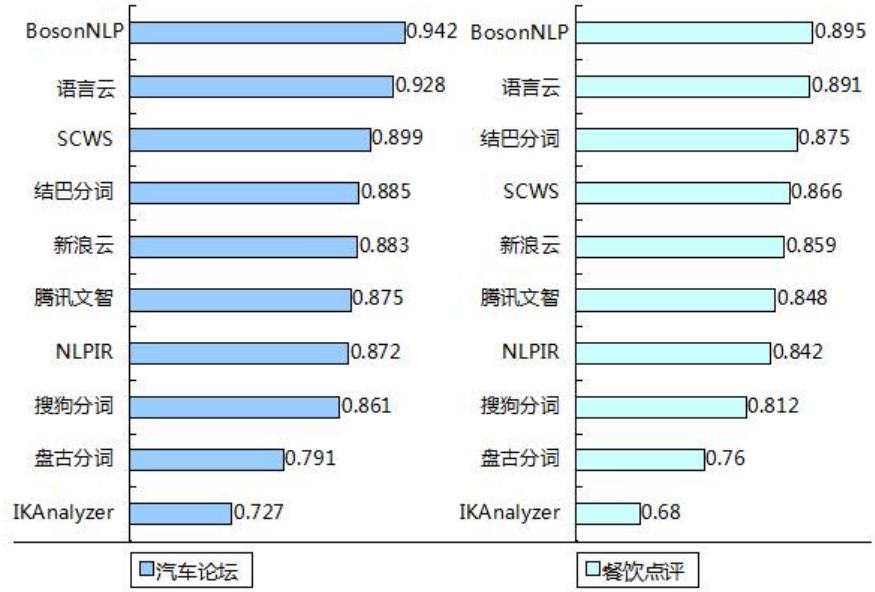
1. 新闻数据：140篇，共30517词语；
2. 微博数据：200篇，共12962词语；
3. 汽车论坛数据（汽车之家）100篇：共27452词语；
4. 餐饮点评数据（大众点评）：100条，共8295词语。

准确度计算规则：

1. 将所有标点符号去除，不做比较。
2. 参与测试的部分系统进行了实体识别，可能造成词语认定的不统一。我们将对应位置替换成了人工注的结果，得到准确率估算的上界。
3. 经过以上处理，用SIGHAN 分词评分脚本比较得到最终的准确率，召回率和F1值。

以上所有数据采用北大现代汉语基本加工规范对所有数据进行分词作为标准。具体数据下载地址请参附录。通过这四类数据综合对比不同分词系统的分词准确度。





上图为参与比较的10款分词引擎在不同数据的分词准确度结果。可以看出，在所测试的四个数据集上BosonNLP和哈工大语言云都取得了较高的分词准确率，尤其在新闻数据上。因为庖丁解牛是将所有能成词的词语全部扫描出来（例如：“最不满意”分为：“最不 不满 满意”），与其他系统输出规不同，因而不参与准确率统计。

为了更直接的比较不同数据源的差别，我们从每个数据源的测试数据中抽取比较典型的示例进行更直的对比。

【新闻数据】

新闻数据的特点是用词规整，符合语法规则，也是普遍做得比较不错的一个领域。对比其他数据源，7家系统都在新闻领域达到最高。包括IKAnalyzer、盘古分词、搜狗分词、新浪云、NLPIR、语言云、osonNLP。并且有三家系统准确率超过90%。

样例：香港 中文 大学 将 来 合肥 一中 进行 招生 宣传 今年 在 皖 招 8 人 万家 热线 安徽 第一 门户

【微博数据】

微博数据用词多样、话题广泛，并常包含错别字及网络流行词。能够比较全面的体现每家分词系统的

确度。

样例：补 了 battle 赛 峰暴班 的 两 个 弟弟 妹妹 @杨宝心 @修儿 一个 是 我 很 挺 的 好 弟弟 一个 是 我 推荐 进 好声音 的 妹子 虽然 都 在 battle 阶段 都 下来 了 但是 我 依然 像 之前 那样 觉得 你们 非常 棒

【汽车论坛】

汽车数据是针对汽车领域的专业评价数据，会出现很多的专业术语。例如示例中的“胎噪”、“风燥等，如果系统没有足够强大的训练词库或领域优化，会使准确率有较大程度降低。比较有意思的是，比其他数据源，有3家系统都在汽车论坛领域达到最高：腾讯文智、SCWS中文分词、结巴分词。

样例：舒适性 胎噪 风噪 偏 大 避震 偏 硬 过 坎 弹跳 明显

【餐饮点评】

餐饮点评数据为顾客评论数据，更偏重口语化。会出现很多类似“闺蜜”、“萌萌哒”口语化词语和多不规范的表达，使分词更加困难。

样例：跟 闺蜜 在 西单 逛街 想 吃 寿司 了 在 西单 没 搜 到 其他 的 日料店 就 来 禾绿 了 我们 俩 都 觉得 没 以前 好 了

各家系统对于多数简单规范的文本的分词已经达到很高的水平。但在仔细对比每一家中文分词后依旧现切分歧义词和未登陆词（即未在训练数据中出现的词）仍然是影响分词准确度的两大“拦路虎”。

1.切分歧义：根据测试数据的切分结果，一类属于机器形式的歧义，在真实语言环境下，只有唯一可的正确切分结果，称其为伪歧义。另一类有两种以上可实现的切分结果，称为真歧义。由于真歧义数无法比较正确或者错误。所有我们着重举例来比较各家系统对伪歧义的处理效果。

正确： 在 伦敦 奥运会 上 将 可能 有 一 位 沙特阿拉伯 的 女子
（BosonNLP、新浪云、语言云、NLPIR、腾讯文智）

错误： 在 伦敦 奥运会 上将 可能 有 一 位 沙特阿拉伯 的 女子
（PHP 结巴分词、SCWS 中文分词、搜狗分词、庖丁解牛）

示例中原意指伦敦奥运会可能有一位沙特阿拉伯的女子，错误分词的意思是指上将（军衔）中有一位沙特阿拉伯的女子，句意截然不同。当然，分析的层次越深，机器对知识库质量、规模等的依赖性就强，所需要的时间、空间代价也就越大。

2.未登录词：未登录词大致包含三大类：

a)新涌现的通用词：类似“神马”、“纳尼”、“甩卖”、“玫瑰金”等新思想、新事物所带来的新汇，不管是文化的、政治的、还是经济的，在人们的生活中不断涌现。同时很多词语也具有一定的时性。

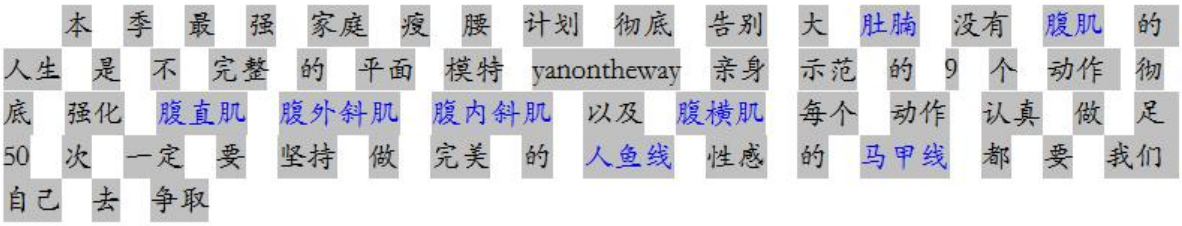
b)专业术语：是相对日常用语而言的，一般指的某一行业各种名称用语，大多数情况为该领域的专业士所熟知。这种未登录词理论上是可预期的。能够人工预先添加到词表中（但这也只是理想状态，在实环境下并不易做到）。

c)专有名词：如中国人名、外国译名、地名、公司名等。这种词语很多基本上不可通过词典覆盖，考分词系统的新词识别能力。

【新涌现的通用词或专业术语】

示例中的蓝色字包括专业术语：“肚腩”、“腹肌”、“腹直肌”、“腹外斜肌”、“腹横肌”；新

现的通用词：“人鱼线”、“马甲线”。大多数的系统对于示例文本的分词结果都不够理想，例如：大肚腩”（SCWS中文分词）“腹直肌腹外斜肌”（搜狗分词、IKAnalyer、NLPIR、SCW中文分词）、“人鱼线”（PHP结巴分词）。总的来说这两种类型的数据每家系统都存在一定的缺，相对而言哈工大的语言云在这方面表现的较好。



【专有名词】

示例出现的专有名词包括“蒂莫西伊斯顿”（姓名）、“英国”“意大利”“北欧”（地点）、“金敦”（机构名）、“伊丽莎白 格林希尔兹”（机构名）。而这种用词典无法穷尽的专有名词也成为家分词准确率降低的重要原因。其中搜狗分词、IKAnalyer、PHP结巴分词、腾讯文智、SCWS中文词在新词识别时较为谨慎，常将这类专有名词切分成多个词语。

当然在分词准确度可以接受的情况下，很多细节问题，包括是否有出错情况、是否支持各种字符、是标注词性等都可能让我们望而却步。在分词颗粒度选择当中，BosonNLP、SCWS、盘古分词、结巴词、庖丁解牛都提供了多种选择，可以根据需求来采用不同的分词粒度。与北大的分词标准对比来说新浪云默认的分词粒度较大，而搜狗分词、腾讯文智分词粒度相对较小。除此之外，BosonNLP、新云、NLPIR、腾讯文智同时提供了实体识别、情感分析、新闻分类等其他扩展服务。下表给出了各家统在应用方面的详细对比。

分词服务	分词粒度	出错情况	支持处理字符	新词识别	词性标注	认证方法	接口
BosonNLP	多选择	无	识别繁体字	有	有	Token	REST API
IKAnalyzer	多选择	无	兼容韩文、日文字符	有	无	无	jar包
NLPIR	多选择	中文间隔符，返回局部乱码	未知	有	有	无	多语言接口
SCWS	多选择	无	未知	有	有	无	PHP库|命令行工具
结巴分词	多选择	无	识别繁体字	有	有	无	python库
盘古分词	多选择	无	识别繁体字并自动转换	有	无	无	无
庖丁解牛	多选择	无	识别繁体字	有	无	无	jar包
搜狗分词	小	放弃超过一定字符长度的句子	识别繁体字并自动转换	未知	有	无	支持上传文档，但是一直失败
腾讯文智	小	空白字符、中文间隔符，整段返回错误码	未知	有	返回中文词性	Signature	REST API
新浪云	大	无	未知	有	有	需要在新浪有一个仓库	REST API
语言云	适中	无	识别繁体字	有	有	Token	REST API

中文分词是其他中文信息处理的基础，并且在很多领域都有广泛的应用，包括搜索引擎、机器翻译（T）、语音合成、自动分类、自动摘要、自动校对等等。随着非结构化文本的广泛应用，中文分词等本处理技术也变得越来越重要。通过评测可以看出，部分开放分词系统在不同领域已经达到较高准确

。对于数据分析处理的从业者，相信在此之上构建数据分析系统、人机交互平台，更能够起到事半功的效果。

注意：分词数据准备及评测由BosonNLP完成。

附录

评测数据地址

http://bosonnlp.com/dev/resource

各家分词系统链接地址

BosonNLP：http://bosonnlp.com/dev/center

IKAnalyzer：http://www.oschina.net/p/ikanalyzer

NLPIR：http://ictclas.nlpir.org/docs

SCWS中文分词：http://www.xunsearch.com/scws/docs.php

结巴分词：https://github.com/fxsjy/jieba

盘古分词：http://pangusegment.codeplex.com/

庖丁解牛：https://code.google.com/p/paoding/

搜狗分词：http://www.sogou.com/labs/webservice/

腾讯文智： http://www.qcloud.com/wiki/API%E8%AF%B4%E6%98%8E%E6%96%87%E6%A1%3

新浪云：http://www.sinacloud.com/doc/sae/python/segment.html

语言云：http://www.ltp-cloud.com/document