title	s	date	published	tags
日语分词器的介绍与比较	japanese-morphological-analysis-compare	2019-05-17 09:19:53 -0700	true	日语 分词 Elasticsearch

对于搜索引擎来说，分词器的质量是对搜索结果影响最大的一个环节，日语分词（形態素解析）经过多年的发展也有了一些比较成熟的分词系统，下面介绍并比较目前主流的开源日语分词系统。

日语分词词典

在介绍分词器之前，有必要先介绍词典，因为词典是分词的基础，词典的质量直接决定了分词器的质量，而分词系统内部采用哪个词典，也是对分词器效果评估的重要参考。

说到日文词典，可能首先想到是知名的商业词典，如「広辞苑」、「大辞林」等，然而由于商业词典等存在版权和授权问题，一般都无法用于开源项目。其次这类商业词典一般以词义解释为主，而对于分词来说，词汇的释义并不是最重要的，而是词性、词根等，所以也并不是所有的商业词典都适用于分词。

因此开源项目一般都采用具备开源许可证的免费词典，目前使用比较多的有：

- UniDic: 由国立国语研究所推出，主要用于分词问题研究的词典，质量非常优秀，并且按照现代、近代、古代，书面语、口语等划分了多个专门词典，License 也有 GPL/LGPL/BSD 多个选择，可以说是日语分词研究的首选词典了。
- ipadic: 这个词典首先是由「奈良先端科学技術大学院大学松本研究室」所开发的分词软件 ChaSen 整理并使用的，词汇的来源是「情報処理振興事業協会(IPA)」所编著的「IPA 品詞体系(THiMCO97)」，由于 ChaSen 这个软件已经停止更新，因此这个词典的最后一个版本也停止在 2.7.0，更新时间是 2007 年。从应用角度来看 ipadic 并不适合使用在实际产品中，但是胜在体积小巧，适合用于个人学习或 demo，总词条数约 14W 条。
- NAIST-jdic : 由于 ipadic 停止更新，NAIST-jdic 就是在 ipadic 继续整理并将词条数增补至 30W，许可证 BSD。最后更新时间为 2008 年。
- mecab-ipadic / mecab-jumandic 由知名项目 MeCab (下文详细介绍) 按照自己项目的格式整理而成的词典
- mecab-ipadic-NEologd 由工程师 @overlast 个人维护的项目，并得到的 LINE 公司的赞助，主要在 mecab-ipadic 的基础上增补了很多互联网的新词

可以通过这个网页比较几个日语分词词典的差异。

在对日语分词词典有了一定了解后，下面逐一介绍日语分词器

MeCab

MeCab 是京都大学信息专业和日本电信电话株式会社通讯研究所共同研究的项目，模型基于 CRF(条件随机场)，基于 C++实现，主要作者是「工藤 拓」，是日本自然语言研究领域大拿，就职于 Google 负责日语输入法相关项目。

MeCab 主要特点有：

- 不依赖特定词典，因此其他语言如中文、韩语也可以使用
- 性能较好
- 多语言接口
- 整理了若干好用的词典

MeCab 无论在学术方面，还是工程方面，都是非常优秀的，可以看到其他很多日语分词相关的项目，或多或少都受到了 MeCab 的影响或使用了 MeCab 的算法、词典等。

Kuromoji

Kuromoji 由位于东京的 Atilika 公司开发，基于 Java 实现。目前已经捐赠给了 Apache 软件基金会，并内置在 Lucene 和 Solr 中作为默认的日文分词器。

Kuromoji 基本支持前文提到的所有词典，如果未指定的话，默认使用 ipadic。

Kuromoji 的分词算法基于 Viterbi Algorithm，因此可以看做是基于 HMM (隐马尔科夫模型) 的分词。

由于 Atilika 是一个纯商业公司，因此 Kuromoji 也更偏向作为日语分词的工程实现，作为 Java 开源项目，与主流的 Java 搜索项目如 Lucene，Elastic 有很好的匹配，工程方面比较规范，容易上手。而在学术方面的贡献就比较少了。

Juman++

Juman 和 Juman++ 都是京都大学信息专业在 NLP 方面的研究成果，分词模型使用了 RNNLM（递归神经网络语言模型），即采用了深度学习技术。开发采用 C++ 实现，Github 页面也给出了已经训练好的模型。

Juman 除了采用自己整理的词典外，还引入了来自 Wikipedia 的词汇。 Juman 的输出除了标准的分词结果外，还可以输出词汇的分类，从而可以对句子做更好的标签和归类。

不过由于文档较少，想要使用 Juman 训练自己的模型或者更换自己的词典还是比较困难的，整体是一个比较偏学术的项目。

KyTea

KyTea 是由卡内基·梅隆大学的 Graham Neubig 主导研究的项目，实现语言是 C++， 算法方面融合了 SVM 和逻辑回归等多个模型，默认使用的词典是 UniDic。

作者的方向偏学术，因此 KyTea 更多也只是作为研究成果的展示，版本早已停止更新，实际应用的项目也较少。

Sudachi

Sudachi 由 Works Applications 公司开发，和 Kuromoji 非常类似，都是 Java 实现的商业开源项目，对比 Kuromoji， Sudachi 可以调整的参数更细致一些，默认词典同时包括了 UniDic 和 NEologd，算法使用的应该是 Lattice LSTM。

项目开源仅 2 年，更新维护比较勤快，官方提供了 Elasticsearch 插件，对开发者比较友好。

nagisa

nagisa 是 NTT DOCOMO 的「池田 大志」 个人开发的基于 RNN 的项目，训练好的模型可以直接使用 pip 安装后使用，不过由于是 Python 开发，运行效率上就远远比不了上述的 C++项目了。

nagisa 整体代码较少，并给出了完整的训练代码和语料库，如果是以学习日语 NLP 为目的的话，是一个不错的选择。

其他

- janome 纯 Python 实现的 Lattice LSTM

日语分词器的横向比较

将以上所有介绍的日语分词器做一个横向比较，可以根据实际需要自行选择。

	算法/模型	实现语言	词典	处理速度	ES 插件	Lisence
MeCab	CRF	C++	可选	最高	有	GPL/LGPL/BSD
Kuromoji	Viterbi	Java	可选, 默认 ipadic	中	内置	Apache License v2.0		Juman++	RNNLM	C++	自制	高	无	Apache License v2.0		KyTea	SVM 等	C++	UniDic	中	有	Apache License v2.0		Sudachi	Lattice LSTM	Java	UniDic + NEologd	中	有	Apache License v2.0		nagisa	Bi-LSTM	Python	ipadic	低	无	MIT

References

- 形態素解析ツールの比較 (NLP2018)
- 形態素解析の過去・現在・未来