句法分析工具比较

刘荣

2015-04-13

目录

一 目前工具

1. 复旦大学 fnlp
2. 斯坦福大学
3. Hanlp
4. 哈工大 ltp

二 复旦 NLP

2.1 简介

FNLP 主要是为中文自然语言处理而开发的工具包，也包含为实现这些任务的机器学习算法和数据集。目前，FNLP 功能包括信息检索（文本分类 新闻聚类），中文处理（中文分词 词性标注 实体名识别 关键词抽取 依存句法分析 时间短语识别 指代消歧），结构化学习算法（在线学习 层次学习 聚类）。FNLP 源代码可免费从 github(https://github.com/xpqiu/fnlp/)下载使用，也可使用在线 demo（http://jkx.fudan.edu.cn/nlp/)体验工具效果。另外，可以通过论文《FudanNLP: A Toolkit for Chinese Natural Language Processing》查看具体算法的实现和工具的具体性能分析。

2.2 算法

在 FNLP 中，依存句法分析使用的是基于转换的依存句法分析[1].在基于转换的分析方法中，依存分析被看作是对输入句子执行若干动作，由这些动作建立起句子中词与词之间的联系。每一个动作都将当前的分析状态转换到新的状态。基于转换的分析方法并不搜索全局最优的动作序列，而是采用贪婪的策略，根据当前状态选择局部最优的动作，一个动作一旦执行就不会在更改，因而又称确定性分析方法.

2.3 格式

1. 依存分析中的关系

Relations	Chinese	Definitions
SUB	主语	Subject
PRED	谓语	Predicate
OBJ	宾语	Object
ATT	定语	Attribute
ADV	状语	Adverbial Modifier
COMP	补语	Complement
SVP	连动	Serial Verb Phrases
SUB-OBJ	兼语	Pivotal Construction
VOC	语态	Voice
TEN	时态	Tense
PUN	标点	Punctuation

表格 1 依存句法中的关系

2. 实例分析

将 FNLP 加载到 Eclipse 中进行功能验证，输入“如何办理身份证”。程序输出格式(Conll)：

```
0 如何 副词 1 状语
1 办理 动词 -1 核心词
2 身份证 名词 1 宾语
3 ？ 标点 1 标点
```

其中运行时间为 0.034s(不包括资源加载的时间)。另外，可以从可视化工具“DependencyViewer”展现：

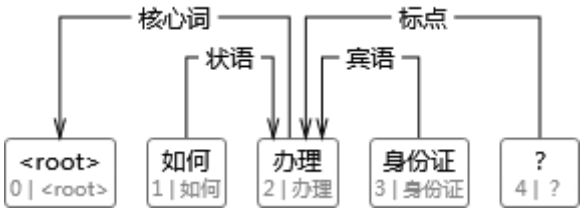


图 1 依存分析树

2.4 性能和效率

性能分析测试结果摘自【2】，如下图所示：

Task	Accuracy	Speed[1]	Memory
CWS	97.5%	98.9K	66M
POS	93.4%	44.5K	110M
NER	98.40%	38K	30M
DePar	85.3%	21.1	80M
TPR	95.16%	22.9k	237K
AR	70.3%	35.7K	52K

[1] characters per second. Test environment: CPU 2.67GHz, JRE 7.

Table 5: System Performances

图 2 依存句法分析性能

其中，依存句法分析的模型大小为 10.6M。

三 斯坦福 NLP

3.1 简介

斯坦福句法分析工具使用 java 语言开发，支持中文的分词，词性标注，句法分析和命名实体识别。其句法分析工具可以在【http://nlp.stanford.edu/software/lex-parser.shtml】下载并使用。也可以使用在线 demo 验证 http://nlp.stanford.edu:8080/parser/。

3.2 算法

斯坦福句法分析集成了三种算法：概率上下文无关文法(PCFG) 基于神经网络的依存句法分析和基于转换的依存句法分析（Shift Reduce）。具体算法参考【4,5】

3.3 格式

3.3.1 句法分析工具输出格式参数

stanfordNLP 句法分析的工具指南，包含了命令的参数和输出格式，具体见【3】。以下是摘取的句法分析的输出格式：

Oneline：成分句法分析输出文件的格式为每行一句的广义表形式的树结构。

Penn：成分句法分析输出文件的格式为层次化树的形式。默认选项为 penn。

latexTree：格式类似于 penn Words：只给出分词格式。如：

继续 播报 详细 的 新闻 内容 。

wordsAndTags：给出分词文本和标记。如

继 续 /VV 播 报 /VV 详 细 /VA 的 /DEC 新 闻 /NN 内 容 /NN 。 /PU

rootSymbolOnly：只给出 ROOT 结点

typedDependencies：给出依存句法分析结果。

mmod(播报-2, 继续-1) rcmod(内容-6, 详细-3) cpm(详细-3, 的-4) nn(内容-6, 新闻-5) dobj(播报-2, 内容-6)

conllStyleDependencies、conll2008：conll 格式(每行一词，每词十项)如下：

1 继续 _ VV _ _ 2 _ _ _
2 播报 _ VV _ _ 0 _ _ _
3 详细 _ VA _ _ 4 _ _ _
4 的 _ DEC _ _ 6 _ _ _
5 新闻 _ NN _ _ 6 _ _ _
6 内容 _ NN _ _ 2 _ _ _
7 。 _ PU _ _ 2 _ _ _

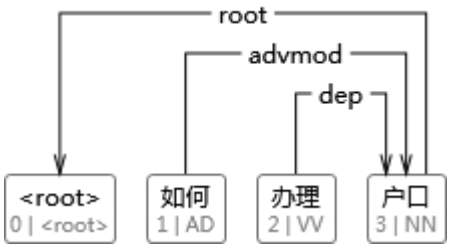
3.3.2 中文输出格式验证

以上为句法分析的格式输出，具体以例子验证，输入“如何办理身份证”。输

出文本格式：

```
1   如何    _   AD  AD  _   3   advmod  _   _
2   办理    _   VV  VV  _   3   dep_    _
3   户口    _   NN  NN  _   0   root    _   _
```

可视化效果：



依存关系：

[advmod(办理-2, 如何-1), root(ROOT-0, 办理-2), dobj(办理-2, 户口-3)]

对于 Stanford 的标记格式(见附录一)：

3.4 性能和效率

没有详细查看其准确率，如果需要请查阅【4,5】。没有找到中文效率的对比。

四 HanLP

4.1 简介

HanLP 是由一系列模型与算法组成的 Java 工具包，目标是普及自然语言处理在生产环境中的应用。不仅仅是分词，而是提供词法分析、句法分析、语义理解等完备的功能。HanLP 具备功能完善、性能高效、架构清晰、语料时新、可自定义的特点。HanLP 完全开源，包括词典。不依赖其他 jar，底层采用了一系列高速的数据结构，如双数组 Trie 树、DAWG、AhoCorasickDoubleArrayTrie 等，这些基础件都是开源的。官方模型训练自 2014 人民日报语料库，您也可以使用内置的工具训练自己的模型。HanLP 的官方地址为【http://hanlp.linrunsoft.com/】，github[https://github.com/hankcs/HanLP].

4.2 算法

HanLP 的依存句法分析有两个模型的实现：最大熵依存句法分析算法和基于 CRF 序列标注的中文依存句法分析算法。具体算法实现参考最大熵【8】CRF【9】

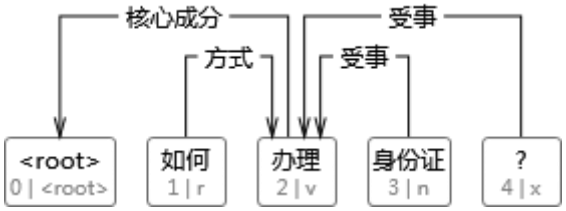
4.3 格式

4.3.1 输出格式

HanLP 的输出格式为 conll 格式，如输出文本格式：

```
1	如何	如何	r	r	_	2	方式	_	_
2	办理	办理	v	v	_	0	核心成分	_	_
3	身份证	身份证	n	n	_	2	受事	_	_
4	?	?	x	x	_	2	受事	_	_
```

通过可视化软件格式：



4.3.2 句法树集

HanLP 依存句法分析由清华大学语义依存网络语料进行训练，使用的句法树集为清华大学的依存句法集合，详见附录二。

4.4 性能和效率

4.4.1 准确度

HanLP 的测试结果摘自【9】，为开发者个人的测试结果：

封闭测试集——

UA: 91.61% LA: 85.68% DA: 84.94% sentences: 20000 speed: 582.42816 sent/s

开发集——

UA: 71.22% LA: 55.02% DA: 52.69% sentences: 2000 speed: 513.34705 sent/s

值得一提的是，由于条件有限，模型并没有训练到足够收敛。如果有足够的时间，应该可以得到更加精确的模型。同时模型的体积也达到了 466MB，代价很大。

4.4.2 验证

如上例

五 LTP

5.1 简介

语言技术平台(Language Technology Platform，LTP)是哈工大社会计算与信息检索研究中心

历时十年开发的一整套中文语言处理系统。LTP 制定了基于 XML 的语言处理结果表示，并在此基础上提供了一整套自底向上的丰富而且高效的中文语言处理模块(包括词法、句法、语义等 6 项中文处理核心技术)，以及基于动态链接库(Dynamic Link Library, DLL)的应用程序接口，可视化工具，并且能够以网络服务(Web Service)的形式进行使用。

5.2 算法

基于图的依存分析方法由 McDonald 首先提出，他将依存分析问题归结为在一个有向图中寻找最大生成树(Maximum Spanning Tree)的问题。 在依存句法分析模块中，LTP 分别实现一阶解码(1o),二阶利用子孙信息解码(2o-sib)和二阶利用子孙和父子信息(2o-carreras)等三种算法。

5.3 格式

输出格式为弧和依赖的父亲节点编号，需要转换为可以相应的 coll 格式和树的格式。

5.4 性能和效率

5.4.1 性能

性能数据参考[11]，如下图：

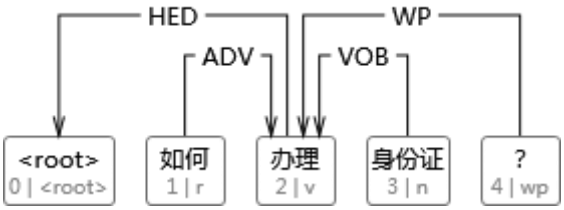
model	1o		2o-sib		2o-carreras	
	UAS	LAS	UAS	LAS	UAS	LAS
开发集	0.8192	0.7904	0.8501	0.8213	0.8582	0.8294
测试集	0.8118	0.7813	0.8421	0.8106	0.8447	0.8138
速度	81.71 sent./s		15.21 sent./s			
运行时内存	338.06M		974.64M			

5.4.2 验证

本次验证使用的 java 调用 LTP 动态链接库，测试用例“如何办理户口“，文本格式：

```
0   如何   如何   r   r   2   ADV
1   办理   办理   v   v   0   HED
2   身份证 身份证 n   n   2   VOB
3   ？   ？   wp  wp  2   WP
```

可视化效果：



六 总结

以上对四种支持中文句法分析的 NLP 工具包进行了简单的介绍，简单总结：

1. 复旦 NLP 是一个中文自然语言处理工具包，集成了基于转换的依存句法分析算法，并且 java 调用简单方便。另外，句法关系用中文标示，简单易用。

2. 哈工大 NLP 是一个用 c 语言开发的中文自然语言处理工具包，集成了基于图的句法分析算法。但是需要利用 juni 编译才能利用 java 调用，不利于跨平台的应用。

3. HanLP 作为个人开发的中文自然语言处理工具包，虽然能够进行句法分析。但是，在测试集上的效果不是很理想，可能不是很稳定。

4. 斯坦福 NLP 作为一个强大的自然语言处理工具包，集成了多种句法分析的算法(包括最新的神经网络模型)并有相应训练的中文模型。从测试的效果来看，中文的效果不是很好，可能用的中文模型不是很正确。

另外，在运行时间上大约都为 2ms 级别左右，在使用时时间上应该不是问题。同时，所有的工具都已经安装进行了实例测试。

参考文献：

[1] Yamada H, Matsumoto Y. Statistical dependency analysis with support vector machines[C]//Proceedings of IWPT. 2003, 3: 195-206.

[2] FudanNLP: A Toolkit for Chinese Natural Language Processing

[3] stanfordParser 指南 http://wenku.baidu.com/view/8d672929ed630b1c59eeb595.html

[4] Pi-Chuan Chang, Huihsin Tseng, Dan Jurafsky, and Christopher D. Manning. 2009. Discriminative Reordering with Chinese Grammatical Relations Features. In *Proceedings of the Third Workshop on Syntax and Structure in Statistical Translation*.

[5] Roger Levy and Christopher D. Manning. 2003. Is it harder to parse Chinese, or the Chinese Treebank? *ACL 2003*, pp. 439-446.

[8] 最大熵依存句法分析器的实现 [http://www.hankcs.com/nlp/parsing/to-achieve-the-maximum-entropy-of-the-dependency-parser.html

[9] 基于 CRF 序列标注的中文依存句法分析器的 Java 实现 [http://www.hankcs.com/nlp/parsing/crf-sequence-annotation-chinese-dependency-parser-implementation-based-on-java.html]

[10] Wanxiang Che, Zhenghua Li, and Ting Liu. LTP: A Chinese Language Technology Platform.

In Proceedings of the Coling 2010:Demonstrations. 2010.08, pp13-16, Beijing, China.
[11] LTP 文档 https://github.com/HIT-SCIR/ltp/blob/master/doc/ltp-document-3.0.md

附录：

一 斯坦福句法树的相关标记

标注指代： http://blog.csdn.net/cuixianpeng/article/details/16864785

1.1 句法分析树标注集

ROOT：要处理文本的语句

IP：简单从句

NP：名词短语

VP：动词短语

PU：断句符，通常是句号、问号、感叹号等标点符号

LCP：方位词短语

PP：介词短语

CP：由‘的’构成的表示修饰性关系的短语

DNP：由‘的’构成的表示所属关系的短语

ADVP：副词短语

ADJP：形容词短语

DP：限定词短语

QP：量词短语

NN：常用名词

NR：固有名词

NT：时间名词

PN：代词

VV：动词

VC：是

CC：表示连词

VE：有

VA：表语形容词

AS：内容标记（如：了）

VRD：动补复合词

CD: 表示基数词

DT: determiner 表示限定词

EX: existential there 存在句

FW: foreign word 外来词

IN: preposition or conjunction, subordinating 介词或从属连词

JJ: adjective or numeral, ordinal 形容词或序数词

JJR: adjective, comparative 形容词比较级

JJS: adjective, superlative 形容词最高级

LS: list item marker 列表标识

MD: modal auxiliary 情态助动词

PDT: pre-determiner 前位限定词

POS: genitive marker 所有格标记

PRP: pronoun, personal 人称代词

RB: adverb 副词

RBR: adverb, comparative 副词比较级

RBS: adverb, superlative 副词最高级

RP: particle 小品词

SYM: symbol 符号

TO:”to” as preposition or infinitive marker 作为介词或不定式标记

WDT: WH-determiner WH 限定词

WP: WH-pronoun WH 代词

WP$: WH-pronoun, possessive WH 所有格代词

WRB:Wh-adverb WH 副词

1.2 斯坦福依存关系

abbrev: abbreviation modifier，缩写

acomp: adjectival complement，形容词的补充；

advcl : adverbial clause modifier，状语从句修饰词

advmod: adverbial modifier 状语

agent: agent，代理，一般有 by 的时候会出现这个

amod: adjectival modifier 形容词

appos: appositional modifier,同位词

attr: attributive，属性

aux: auxiliary，非主要动词和助词，如 BE,HAVE SHOULD/COULD 等到

auxpass: passive auxiliary 被动词

cc: coordination，并列关系，一般取第一个词

ccomp: clausal complement 从句补充

complm: complementizer，引导从句的词好重聚中的主要动词

conj : conjunct，连接两个并列的词。

cop: copula。系动词（如 be,seem,appear 等），（命题主词与谓词间的）连系

csubj : clausal subject，从主关系

csubjpass: clausal passive subject 主从被动关系

dep: dependent 依赖关系

det: determiner 决定词，如冠词等

dobj : direct object 直接宾语

expl: expletive，主要是抓取 there

infmod: infinitival modifier，动词不定式

iobj : indirect object，非直接宾语，也就是所以的间接宾语；

mark: marker，主要出现在有“that” or “whether”“because”, “when”,

mwe: multi-word expression，多个词的表示

neg: negation modifier 否定词

nn: noun compound modifier 名词组合形式

npadvmod: noun phrase as adverbial modifier 名词作状语

nsubj : nominal subject，名词主语

nsubjpass: passive nominal subject，被动的名词主语

num: numeric modifier，数值修饰

number: element of compound number，组合数字

parataxis: parataxis: parataxis，并列关系

partmod: participial modifier 动词形式的修饰

pcomp: prepositional complement，介词补充

pobj : object of a preposition，介词的宾语

poss: possession modifier，所有形式，所有格，所属

possessive: possessive modifier，这个表示所有者和那个’S 的关系

preconj : preconjunct，常常是出现在 “either”, “both”, “neither”的情况下

predet: predeterminer，前缀决定，常常是表示所有

prep: prepositional modifier

prepc: prepositional clausal modifier

prt: phrasal verb particle，动词短语

punct: punctuation，这个很少见，但是保留下来了，结果当中不会出现这个

purpcl : purpose clause modifier，目的从句

quantmod: quantifier phrase modifier，数量短语

rcmod: relative clause modifier 相关关系

ref : referent，指示物，指代

rel : relative

root: root，最重要的词，从它开始，根节点

tmod: temporal modifier

xcomp: open clausal complement

xsubj : controlling subject 掌控者

二 清华大学依存关系

参考【http://tcci.ccf.org.cn/conference/2013/dldoc/ev05.pdf】

附表 1：清华大学语料依存关系集合

主语义关系（16）	施事 领有者	关系主体	经验者	描写体	存现体
	受事 结果	内容 整体	类指 部分	占有物 代价	目标 触及部件
辅助语义关系（24）	时间 后延时段	进程时间	起始时间	终止时间	时距
	处所 终状态	原处所	通过处所	终处所	原状态
	相伴体 根据 比较内容	参照体 来源 比较量	目的 工具	原因 手段	方向 材料
定、状语语义关系（10）	限定 评论 频率	数量 方式	描述 程度	同位语 范围	动量
连动词及从句语义关系（9）	结果事件 条件	接续 并列	伴随 递进	事件过程 让步	除了
特殊句法结构(9)	“的”字依存 语气依存	…是…的依存 关联词依存	连接依存 趋向动词依存	方位词依存 介词依存	时态语态依，
特殊关系(2)	核心成分	依存失败			

三 哈工大相关标注

1 词性标注集（863 词性标注集）

Tag	Description	Example	Tag	Description	Example
a	adjective	美丽	ni	organization name	保险公司
b	other noun-modifier	大型, 西式	nl	location noun	城郊
c	conjunction	和, 虽然	ns	geographical name	北京
d	adverb	很	nt	temporal noun	近日, 明代
e	exclamation	哎	nz	other proper noun	诺贝尔奖
g	morpheme	茨, 甥	o	onomatopoeia	哗啦
h	prefix	阿, 伪	p	preposition	在, 把
i	idiom	百花齐放	q	quantity	个
j	abbreviation	公检法	r	pronoun	我们
k	suffix	界, 率	u	auxiliary	的, 地
m	number	一, 第一	v	verb	跑, 学习
n	general noun	苹果	wp	punctuation	，。！
nd	direction noun	右侧	ws	foreign words	CPU
nh	person name	杜甫, 汤姆	x	non-lexeme	萄, 翱

2 依存句法关系

关系类型	Tag	Description	Example
主谓关系	SBV	subject-verb	我送她一束花 (我 <-- 送)
动宾关系	VOB	直接宾语，verb-object	我送她一束花 (送 --> 花)
间宾关系	IOB	间接宾语，indirect-object	我送她一束花 (送 --> 她)
前置宾语	FOB	前置宾语，fronting-object	他什么书都读 (书 <-- 读)
兼语	DBL	double	他请我吃饭 (请 --> 我)
定中关系	ATT	attribute	红苹果 (红 <-- 苹果)
状中结构	ADV	adverbial	非常美丽 (非常 <-- 美丽)
动补结构	CMP	complement	做完了作业 (做 --> 完)
并列关系	COO	coordinate	大山和大海 (大山 --> 大海)
介宾关系	POB	preposition-object	在贸易区内 (在 --> 内)
左附加关系	LAD	left adjunct	大山和大海 (和 <-- 大海)
右附加关系	RAD	right adjunct	孩子们 (孩子 --> 们)
独立结构	IS	independent structure	两个单句在结构上彼此独立
核心关系	HED	head	指整个句子的核心

.