基于词间依存的汉语基本块依存关系识别

李 丽 赵文娟 樊孝忠

(北京理工大学计算机学院 北京 100081)

摘 要 基本块的分析是句法分析中的重要技术,根据依存理论,提出了一种分析基本块之间的依存关系的方法。首先使用BIO标记来识别基本块,然后根据词之间的依存关系判别基本块之间的依存关系。实验表明,基本块识别的正确率和召回率分别为82.3%和78%,基本块之间依存关系识别的正确率和召回率分别为89%和90.5%。

关键词 基本块,依存关系,词之间的依存关系,句法分析

中图发分类号 TP391 **文献标识码** A

Chinese Chunk Dependency Relationship Parsing Based on Words Dependency

LI Li ZHAO Wen-juan FAN Xiao-zhong

(School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

Abstract Chinese Chunking is an import technique in the Syntax parsing. This paper adopted the theoretical analysis of dependency to parse the Chinese Chunk Dependency Relationship. First of all, BIO tags are used to identify the Chinese chunk, then according to the dependency of words to distinguish the dependency of Chinese Chunk. Experimental results show that, the Chinese Chunking rates of accuracy and recall are respectively 82.3% and 78%, and the Chinese Chunk Dependency rates of accuracy and recall are respectively 89% and 90.5%.

Keywords Chinese chunk, Dependency relationship, Words dependency, Syntax parsing

1 引言

句法分析是中文问答系统中答案抽取、信息检索等模块的关键技术。目前的句法分析研究方法多采用完全句法分析法,然而由于汉语是一种意和性的语言,有其自身的复杂性和不确定性,该方法并不能达到令人满意的效果。因此,一些研究者提出了浅层句法分析方法。浅层句法分析主要关注句子中的局部成分及其关系,并利用这些局部关系来降低分析的复杂性,从而有效地支持完全句法分析。最初的浅层句法分析方法由于只考虑了句子的基本特征、句法特征或语义特征信息,在问答系统中对答案检索与抽取的效果并不理想。因此在现有的研究中,很多研究人员采用分析句子中词之间依存关系的方法来明确词语之间的支配关系,从而更有效地指导答案的抽取。例如,文献[1]提出了一种可以同时获取基本块的外部句法标记和内部关系描述的汉语基本块分析方法,即通过设计关系标记集来描述基本块内部词之间的句法依存关系,获得了很好的内部关系分析结果。文献[2]利用语义、语法等语言知识,建立了一种基于依存关系的句法分析统计模型,该模型是一个词汇化句法分析模型,能结合分词、词性标注进行句法分析,取得了很好的效果。以上这些方法都是分析句子中词之间的依存关系,虽然可以有效地指导答案的抽取,但分析的粒度较小,并且句子中各句法成分之间的依存关系也不明确。本文提出一种基于词依存的汉语基本块依存关系识别方法,该方法结合依存语法理论,分析基本块之间的依存关系,从整体上把握句子的结构和各个成分之间的支配关系。

本文研究汉语句子中基本块的依存关系,其中涉及基本块的标注及识别、基本块之间依存关系的判定等。本文第1节为引言;第2节介绍汉语基本块以及依存理论;第3节介绍基本块的标注、识别,条件随机场模型及特征的选择;第4节介绍由词之间的依存转化为基本块之间依存的方法;第5节是实验结果及分析;最后为全文的总结。

2 汉语基本块

2.1 基本块的定义

基于语块的分析属于浅层句法分析,它将具有语法关联的词组成一个语块,为完全句法分析提供一个中间结果[3]。Abeny最早提出了一个完整的组块描述体系,并把组块定义为从句内的一个非递归的核心成分[4]。结合汉语的特点,周强定义了汉语基本块描述体系,即基本块=基本拓扑结构+句法形式描述+语义内容描述,并给出了相应的基本块标记集合[5]。

基本块不同于词和短语。词是最小的能够独立运用的语言单位;短语是由两个或两个以上的词组合起来构成的;基本块则是一种结构,是符合一定句法功能的基本短语。基本块主要描述句子中直接相邻的、以名词、动词、形容词等实词为中心聚合组成的具有特定语义内容的词语序列,一般不包括各种功能词,如连词、叹词、语气词、助词、标点符号等[6]。基

李 丽(1990—),女,硕士,主要研究方向为自然语言处理,E-mail:lilichn@126.com;赵文娟(1988—),女,硕士,主要研究方向为自然语言处理;樊孝忠(1948—),男,教授,博士生导师,主要研究方向为自然语言处理。

本块是比词更大的信息单元，在句子中担当的成分可以分为名词块、数量块、时间块、空间块、动词块、形容词块和副词块。

2.2 基本块依存关系

对自然语言进行语义分析基于语义学理论，常用的语义分析方法有概念依存理论、格语法、概念从属理论、语义场理论和知网等[7]。依存是一种将句子描写层级结构化的语言方法。现代依存语法理论的创立者是法国语言学家 Tesniere，他在《结构句法基础》中提出依存关系把握着两个成分，即中心成分和依存成分，中心成分通常是动词，其他成分受其支配，这样便于表示句子中词与词之间的关系[8]，也便于进一步的语义分析。

在判定基本块之间的依存关系时，可以由词之间的依存推出块之间的依存。在识别出基本块的边界后，将属于同一基本块的词组合，在组合后的基本块中，找出该基本块与其他基本块有联系的词，该词与其他块的依存关系作为其所在块与其他块的依存关系，并打断基本块中词之间的依存关系。

3 汉语基本块分析方法

汉语基本块分析的目的是给定一个汉语句子，它能识别出句子中每个基本块的边界，并为每个基本块标注相应的句法标记和位置标记。

3.1 基本块的标注

基本块的分析可以转化为序列标注问题，将句子中每个词语标注一个合适的类别标记，以实现基本块的自动分析。

在标记基本块的边界时，采用 BIO 标记集合。每个标记由两部分组成，如表 1 所列，第一部分是基本块在句子中承担的成分，主要包括名词块、动词块、形容词块、数量块、空间块、时间块和副词块；第二部分是词语在基本块中的位置，基本块的开始位置用 B 标记，内部位置用 I 标记，不属于基本块的词标记为 O。这两部分之间用“-”来连接。

表 1 基本块的句法标记和位置标记集合

句法标记	内容描述	位置标记	内容描述
np	名词块	B	开始位置
vp	动词块	I	内部位置
ap	形容词块	O	不属于块
mp	数量块		
sp	空间块		
tp	时间块		
dp	副词块		

表 2 基本块标记示例

词	词性	基本块边界标记
大学生	n	np-B
通过	P	O
社会	n	np-B
实践	vN	np-I
加深	V	vp-B
了	uA	vp-I
对	P	O
革命	n	np-B
老区	n	np-I
的	uJDE	O
理解	vN	np-B
。	wE	O

该序列标注问题描述为：设输入的序列为 $X=x_1x_2x_3\cdots x_n$，其中 x_i 为带词性标注的词语，相应的输出序列为 $Y=y_1y_2y_3\cdots y_n$，其中 y_i 对应 x_i 并带有边界标记信息。则对一个输入序列 X 进行标注的过程就是为其寻找一个最优的输出序列标记 Y 的过程。根据序列 Y 就可以进行基本块的识别及依存关系的判定工作。如对输入的句子“大学生/n 通过/p 社会/n 实践/vN 加深/v 了/uA 对/p 革命/n 老区/n 的/uJDE 了解/vN。/X”，其对应的边界标记如表 2 所列。

3.2 基本块的识别方法

识别基本块的思路是标记为同一类别 X-B 和 X-I 的词构成一个基本块，该基本块直到遇到下一个标记为 X-B 或者 O 的词为止，X 为 np、vp、ap、mp、sp、tp、dp 中的一个。基本块识别的输入是带有基本块边界标记的句子，其格式为每个词语及其边界标记占一行，句子之间以空行隔开。例如某个词 $word_i$ 的边界标记为 X-B，若其后的词 $word_{i+1}word_{i+2}word_{i+3}word_{i+4}$ 的边界标记为 X-I，$word_{i+5}$ 的边界标记不是 X-I 或者是 O，则将词 $word_i\ word_{i+1}word_{i+2}word_{i+3}word_{i+4}$ 组合成一个基本块，其在句中的成分为 X。

将词转换为基本块的转换算法为：

(1)如果某一行以“B”结尾或者为空行，且其下一行以“B”结尾或下一行为空行，则该行不与其他行合并，自成一个基本块，其在句中的成分由“X-B”改为“X”。

(2)如果第 N 行以“I”结尾，且其下一行以“B”结尾或下一行为空行，则遍历该行前面的行，若第 M 行以“B”结尾，遍历终止。从第 M 行到第 N 行的所有词组成一个基本块，该基本块在句中的成分由“X-B”改为“X”。

(3)对于其他行则不做处理。

这样就从词转换为了基本块。

3.3 条件随机场模型

条件随机场模型(CRF)是 Lafferty 于 2001 年在最大熵模型和隐马尔科夫模型的基础上提出的一种判别式概率无向图学习模型，是一种用于标注和切分有序数据的条件概率模型[9]。它消除了隐马尔科夫模型的强度独立性假设，并解决了最大熵模型的标注偏执问题，提高了序列标注性能[10]。在训练阶段，条件随机场模型使用最大化后验估计来估计模型的参数。在测试阶段，使用 Viterbi 算法来自动预测出每条序列的标注结果。其公式表示如下：

$$x=\{x_1,x_2,\cdots,x_n\} \tag{1}$$

$$y=\{y_1,y_2,\cdots,y_n\} \tag{2}$$

$$p(y|x,\lambda)\propto\exp\left(\sum_j\lambda_j t_j(y_{i-1},y_i,x,i)+\sum_k\mu_k s_k(y_i,x,i)\right) \tag{3}$$

式(1)表示观察序列，式(2)是有限状态的集合，则根据随机场的基本理论，有式(3)。

由于其在各种序列化标注问题(包括词类标注、命名实体识别、语义角色标注)中显示出了很好的效果，本文也将其用于基本块的边界标注中。

3.3.1 条件随机场模型特征的选择

在利用条件随机场模型时，把重点放在了特征模板的选择上，除了要包含词、词性特征、词与词性的组合特征外，还要扩展处理窗口的大小以便融入更多的上下文信息，然后用统计的方法构建分析模型来达到更好的识别效果。本文使用了两种特征模板来进行对比，一是选用 CRF 自带的特征模板作为特征模板一，另一个是参考文献[11]制定的特征模板二。

特征模板二包含的信息：

(1)前后各两个词的词语和词性特征。

(2)相邻两个词的词语组合特征。

(3)当前词与前后词的词语组合特征,当前词与前(后)相邻两个词的词语组合特征。

(4)窗口大小为5时各相邻词的词性组合特征。

(5)当前词与前后词的词性组合特征,当前词与前(后)相邻两个词的词性组合特征。

在用这两个模板进行训练时,发现模板一比模板二迭代的次数多,而且terr和serr的值比模板二的要大,综合以上因素,在实验阶段,本文使用特征模板二。

4 基本块依存关系的识别

本文用到的语料库是清华汉语树库(TCT)[12],它有完整的层次结构树,句法树上的每个非终结符节点都有成分标记和关系标记,具有很好的信息覆盖率和语料适应性。在进行基本块的边界标注和识别之前,首先将树库的格式转化为表3的格式。对转化后的语料库使用CRF训练得到模板,此模板用于基本块的边界识别。本文在识别基本块的依存关系时,利用哈工大的依存句法分析器[13],得到词之间的依存,然后由词之间的依存转化为基本块之间的依存。

表3 清华汉语树库转化后的格式

词的位置	词	词性	基本块标记
0	思科	nR	np-B
1	公司	n	np-I
2	是	vC	O
3	全球	n	np-B
4	最	aD	ap-B
5	大	a	ap-I
7	互联网	n	np-B
8	设备	n	np-I
9	供应商	n	np-I

经过基本块边界标注和词之间的依存分析后,将属于同一类别的词进行组合,在基本块识别的同时,进行依存关系的识别。其识别算法为:

(1)若某一行中包含"X-B",且其下一行包含"B"或是空行,则该词是一个基本块,根据其依存词的位置找出依存的词,其与其他基本块的依存关系不变。

(2)若某一行包含"X-I",且其下一行包含"B"或是空行,在识别基本块的同时,根据该块中所有依存词的位置,找出与其他基本块有依存关系的词,该词与其他词的依存关系作为该词所在块与其他块的依存关系。再根据该词对应的依存词的位置,找出依存的词。

(3)重新对基本块进行排序,并确定基本块所依存的块的位置。

如表4的例句,"思科公司"组合成一个np基本块,"全球"自身形成一个np基本块,"最大"组合成一个ap基本块,"互联网设备供应商"形成一个np基本块,"是""的"不属于任何基本块。然后找出基本块中与其他基本块有依存关系的词,该词与其他基本块的依存关系作为其所在基本块与其他基本块之间的依存关系,并打断基本块内词之间的依存关系。如"互联网设备供应商"与"是"有依存关系的词是"供应商","供应商"与"是"的依存关系为VOB,则"互联网设备供应商"与"是"的依存关系为VOB。"互联网设备供应商"这个基本块内,"互联网"依存于"设备","设备"依存于"供应商",打断它们之间的依存,保证一个基本块与另外的某个基本块有一个依存关系。如表5所列,其中,所依存的块的位置这列中,"-1"表示的是这句话的中心词。依存关系的标记采用哈工大的依存关系标记集,共24个依存关系的标记。

表4 经过依存分析后的文本

词的位置	词	词性	基本块标记	依存词的位置	依存关系
0	思科	nR	np-B	1	ATT
1	公司	n	np-I	2	SBV
2	是	vC	O	-1	HED
3	全球	n	np-B	5	SBV
4	最	aD	ap-B	5	ADV
5	大	a	ap-I	6	DE
6	的	uJDE	O	9	ATT
7	互联网	n	np-B	8	ATT
8	设备	n	np-I	9	ATT
9	供应商	n	np-I	2	VOB

表5 基本块之间依存关系示例

基本块的位置	基本块	所依存的块的位置	依存关系
0	思科公司	1	SBV
1	是	-1	HED
2	全球	3	SBV
3	最大	4	DE
4	的	5	ATT
5	互联网设备供应商	1	VOB

5 实验结果及分析

5.1 语料库的处理

本文在进行基本块的边界识别时,输入数据格式为"词/词性",格式为表2中的前两列,即句子已经分词且每个词及词性占一行,句子与句子之间以空行分割。清华汉语库[12]的格式如表6所列,与所需的输入格式不符,所以要进行格式的转化,首先要去掉无用的信息,去掉全部由数字组成的行,提取出第一、二、四列,且句子之间用空行隔开,并为每个词标上在句子中的位置,处理完后的形式如表3所列。

表6 清华汉语树库的格式

词	词性	关系标记	成分标记	组合标记
执法	vN	M	np-B	np-ZX
部门	n	R	np-I	np-ZX

5.2 实验结果

为了得到CRF的训练模板,使用清华汉语树库,按照8:2的比例分成两份,较大的一份用特征模板二进行训练得到训练模板,另一份用于测试。

在测试生语料时,首先利用中科院的分词工具进行分词,然后再利用哈工大的依存分析器分析词之间的依存关系。在进行分词时获取词信息及词性信息。由于中科院的词性标注集与清华汉语库的词性标注集不一样,还要进行词性标注的映射,将中科院的词性标注信息转化为清华汉语库的词性标注信息。

分析完词之间的依存关系后,再利用由CRF得到的训练模板处理数据,得到带有基本块边界标记的且有词之间依存关系的数据,然后利用第3节的思路,将词组合成相应的基本块,再判定基本块之间的依存关系。

实验数据是从清华汉语树库中随机选择100个句子组成测试语料,以验证所提方法的有效性。首先用中科院的分词工具进行分词,然后用哈工大的依存分析器分析词之间的依存关系,再用由CRF得到的训练模板标注基本块的句法成分和在基本块中的位置。实验采用正确率和召回率进行评价,

计算方法如式(4)和式(5),结果如表7所列。

$$正确率=\frac{实验得到的正确数据数量}{实验得到的所有数据数量}*100\% \quad (4)$$

$$召回率=\frac{实验得到的正确数据数量}{人工标注的数据数量}*100\% \quad (5)$$

表7 正确率与召回率

识别的数据	正确率	召回率
基本块	82.3%	78%
依存关系	89%	90.5%

分析识别错误的基本块,主要有以下几个原因:

(1)分词时,把某些词分成了两个词。

(2)这两个词的词性与原来词的词性不一致。

(3)在把中科院的词性标注集转化为清华树库中的词性标注集时,两个词性标注集合的大小不一样,某些词的词性转化不太标准。

(4)由CRF得到的训练模板还有待改善。

分析识别错误的基本块间的依存关系,原因如下:

若某个基本块内有多个词与其他基本块有依存关系,则无法判断哪个词是该基本块的核心词,也就不能准确地识别该基本块与其他基本块的关系。默认的是选择第一与其他基本块有依存关系的词,该词所对应的依存关系为整个基本块所对应的依存关系。

结束语 本文使用BIO标记集来标注基本块的边界,以识别基本块,由词之间的依存关系来识别基本块间的依存关系,从实验结果看,取得了一定的效果。本文以后的研究工作是改善CRF的特征模板;识别基本块的核心词,把核心词与其他基本块的依存关系作为其所在基本块与其他基本块的依存关系。

参考文献

[1] 宇航,周强.汉语基本块标注系统的内部关系分析[J].清华大学学报,2009,49(10):1708-1711

[2] 袁里驰.基于依存关系的句法分析统计模型[J].中南大学学报,2009,40(6):1630-1635

[3] 陈亿,周强,宇航.分层次的汉语功能块描述库构建分析[J].中文信息学报,2008,22(3):24-31

[4] Steven A. Parsing by Chunks[M]. Robert Berwick, Steven Abney and Carol Tenny, eds. Principle-Based Parsing, Kluwer Academic Publishers, 1991:257-278

[5] 周强.汉语基本块描述体系[J]:中文信息学报,2007,21(3):21-27

[6] 李素建,刘群.汉语组块的定义和获取[C]//语言计算与基于内容的文本处理——全国第七届计算语言学联合学术会议论文集.2003

[7] 唐怡.用于常识推理的中文句子语义知识抽取[D].厦门:厦门大学,2010

[8] 宗成庆.统计自然语言[M].北京:清华大学出版社,2007

[9] Lafferty J, McCallum A, Pereira F. Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data [C]// Proceedings of the 18th International Conf. on Machine Learning. 2001:282-289

[10] 王昕,王金勇,刘春阳,等.基于CRF的汉语语块分析和事件描述小句识别[R].北京:中文信息学会,2009

[11] 程勇,孙承杰,刘远超,等.基于CRFs的级联中文组块识别[R].北京:中文信息学会,2009

[12] 周强.汉语句法树库标记体系[J].中文信息学报,2004,18(4):1-8

[13] http://ir.hit.edu.cn/demo/ltp

(上接第254页)

SARS病毒却已经逐渐淡出人们的视线。以每天都有HIV病毒研究文献发表的频率预测HIV病毒仍然是未来几年医学领域的研究热点。

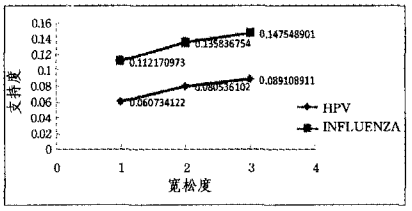

图2 宽松度与支持度的关系

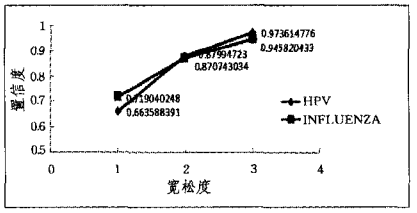

图3 宽松度与置信度的关系

结束语 本文针对文本数据的时间属性,论述了多粒度时间下文本数据的周期模式概念,并在此基础上建立了时态文本数据模型与其周期模型。分别提出了粒度转换和时间间隔定义,给出了宽松周期模式的概念,提出了一个多粒度时间下的文本周期模式挖掘算法,通过实例分析得到了宽松周期的支持度和置信度。本文研究对多粒度时间下的大文本数据的周期模式挖掘具有重要的意义。

参考文献

[1] Bettini C. Testing complex temporal relationships involving multiple granularities and its application to data mining[J]. ACM, 1996,12(4):86-88

[2] Bettini C, Wang S X, Sushil J, et al. Discovering frequent event patternswithmultiple granularities in time sequences [J]. IEEETransactions on Knowledge and Data Engineering, 1998, 10(2):222-237

[3] 孟志青.时态数据采掘中的时态型与时间粒度研究[J].湘潭大学自然科学学报,2000,22(3):1-4

[4] 孟志青.时态关联规则采掘的若干性质[J].计算机工程与应用,2001,37(10):42-44

[5] 姜华,孟志青,肖建华,等.一种时态近似周期的数据挖掘研究[J].软件技术与数据库,2006,32(22):61-63

[6] 程昱.时态数据周期挖掘理论与算法的研究[D].湘潭大学,2005

[7] Li Ying-jiu, Wang X, et al. Discovering Temporal Patterns in Multiple Granularitiesp [C]// TSDM' 00 Proceeding of the First International Workshop on Temporal, Spatial, and Spatio-Temporal Data Mining-Revised Papers. 2007:5-19