基于句法分析的汉语词义消歧*

张春祥[1a,1b], 栾 博[1a], 高雪瑶[1a], 卢志茂[2]
(1. 哈尔滨理工大学 a. 计算机科学与技术学院; b. 软件学院, 哈尔滨 150080; 2. 哈尔滨工程大学 信息与通信工程学院, 哈尔滨 150001)

摘 要: 为了提高词义消歧的质量,对歧义词汇的上下文进行结构分析,提出了一种利用句法知识来指导消歧过程的方法。在歧义词汇上下文的句法树中,提取句法信息和词性信息作为消歧特征;同时,使用朴素贝叶斯模型作为消歧分类器。利用词义标注语料对分类器的参数进行优化,然后对测试数据中的歧义词汇进行消歧。实验结果表明,消歧的准确率有所提升,达到了 66.7%。

关键词: 词义消歧; 句法信息; 词性; 消歧分类器
中图分类号: TP391.2 **文献标志码**: A **文章编号**: 1001-3695(2014)01-0040-03
doi:10.3969/j.issn.1001-3695.2014.01.008

Chinese word sense disambiguation based on parsing analysis

ZHANG Chun-xiang[1a,1b], LUAN Bo[1a], GAO Xue-yao[1a], LU Zhi-mao[2]
(1. *a. School of Computer Science & Technology, b. School of Software, Harbin University of Science & Technology, Harbin* 150080, *China*; 2. *College of Information & Communication Engineering, Harbin Engineering University, Harbin* 150001, *China*)

Abstract: In order to improve the quality of word sense disambiguation, this paper analyzed the structure of the context including an ambiguous word and proposed a new method of word sense disambiguation based on parsing knowledge. It extracted parsing information and part of speech as disambiguation features. This paper used naive Bayesian model as word sense disambiguation classifier and trained its parameters on sense-annotated corpus. Then it applied the classifier to disambiguate ambiguous words in test data. Experimental results show that the accuracy rate of disambiguation is improved and arrives at 66.7%.

Key words: word sense disambiguation(WSD); parsing information; part of speech; disambiguation classifier

0 引言

一词多义是自然语言固有的特性,多义词在独立的语言环境中不能体现出它所表达的含义,只有在特定的上下文中才能确定它的词义。词义消歧(WSD)的任务就是将最确切的词义分配到给定上下文的歧义词中去[1]。词义消歧一直是自然语言处理领域中的研究热点和难点。目前,根据消歧知识的来源不同词义消歧方法大体可以分为基于统计的方法和基于知识的方法两类。李旭等人[2]提出了一种改进的全文无指导词义消歧模型,结合互信息和 Z-测试结果来选取特征,通过统计学习技术来估算 EM 的初始参数,然后利用 EM 算法进行迭代计算。实验结果表明,改进方法有效地提高了汉语词义消歧的准确率,具有良好的扩展性和实用性。Pedersen 通过集成多个分类器来改善词义消歧的效果,分别构造了组合 Bayes 模型[3]和组合决策树分类模型[4]。在组合 Bayes 模型中,假设特征词之间是相互独立的,但是没有规定上下文窗口的大小。何径舟等人[5]分析了特征模板对消歧结果的影响,提出了一套基于最大熵分类模型的自动特征选择方法,主要包括针对所有歧义词的统一特征模板选择算法和针对单个歧义词的独立特征模板优化算法。实验结果表明,使用自动选择的特征不仅简化了特征模板的表示,而且提高了汉语词义消歧的性能。吴云芳等人[6]以支持向量机模型、Bayes 模型和决策树作为单元分类器,将均值、乘法和最大值三种集成方法应用于汉语词义消歧中,同时,系统地比较了不同集成方法在消歧过程中的性能表现。Cruys 等人[7]提出了一个基于潜在语义的词义归纳和消歧联合模型,在词义归纳和消歧过程中,歧义词及其上下文被映射到潜在语义词空间中的有限主题维上。《同义词词林》和 WordNet 都是常用的义类词典。鲁松等人[8]利用《同义词词林》来寻找与某个义项具有相同、相似或相关语义范畴的词汇,将其上下文视为文档,使用向量空间模型来解决词义消歧问题。Chen 等人[9]提出了 AALesk 算法,参考了 WordNet 中的语义关系,并为每种关系分配权重,在统计理论框架下以语义来指导词义消歧过程。Navigli 等人[10]给出了一种多语联合词义消歧方法,该方法利用了多语知识库 BabelNet,在不同语言之间进行了基于图的词义消歧,同时使用词汇的不同语言译文作为补充来实施消歧。实验结果表明,在使用了覆盖面较广的多语词汇知识和基于图的消歧算法之后,单语和多语词义消歧性能有了较大程度的提高。Navigli 等人[11]还提出了一个访问多语词汇知识库 BabelNet 的 API 函数,其目标是为词义消歧领域的研究者提供一个易于执行多语词义分析的工具,以使消歧过程可以融入更多的语言学知识。卢志茂等人[12]将依存分析技术引入无指导学习过程,充分利用依存分析结果来确定能够

收稿日期: 2013-05-09; **修回日期**: 2013-06-26 **基金项目**: 黑龙江省教育厅科学技术研究资助项目(12531106)

作者简介: 张春祥(1974-),男,黑龙江哈尔滨人,教授,博士,主要研究方向为自然语言处理(z6c6x666@163.com);栾博(1988-),女,硕士研究生,主要研究方向为自然语言处理;高雪瑶(1979-),女,副教授,博士,主要研究方向为自然语言处理;卢志茂(1972-),男,教授,博士,主要研究方向为自然语言处理.

对词语词义构成内在限制的上下文,来进行词义消歧。在利用句法信息进行词义分类时,很多研究者采用了依存分析技术。短语结构文法也是一种句法分析技术,主要是面向语法层次的,描述了多词组块之间的依赖关系。本文对歧义词汇的上下文进行短语句法分析,提出了一种融合句法知识的词义消歧模型。

1　基于句法树的消歧特征获取

在传统的开设直线窗口的方法中,消歧特征是歧义词汇的左右邻接词。然而在很多情况下,决定歧义词真实含义的信息往往来自与歧义词距离较远的词汇。为了获取更多的消歧信息,只能扩大直线窗口的大小,但这将引入更多的噪声,而且还会产生数据稀疏问题。因此,本文将从句法树中抽取消歧特征。首先对包含歧义词汇的汉语句子进行分词处理;然后对句子进行短语结构拆分,标记每个短语的句法类别;最后根据层次关系建立对应的句法分析树。消歧特征提取的主要思路是:先根遍历汉语句法树,确定歧义词所在的节点;然后以歧义词为中心向左、向右和向上开设窗口,形成树状消歧窗口。尽管句法分析的结果还不能令人完全满意,但是对于词义消歧来说仍然能够提供部分指导信息。从歧义词汇的上下文句法树中,抽取其父节点、左兄弟节点和右兄弟节点的句法知识以及词性知识,融入词义消歧过程。

为了便于从句法树中获取歧义词汇的消歧特征,将每个汉语单词存放在节点 Node 中。Node 包含以下属性:Word 表示单词;POS 为词性;Phrase 表示短语句法标记;CPOS 为核心节点词性;Parent 指向该节点的父节点;LBrother 指向该节点的左兄弟节点;RBrother 指向该节点的右兄弟节点;LChild 指向该节点的最左孩子节点;RChild 指向该节点的最右孩子节点。对于汉语句子 CS 的句法树 TS 而言,歧义词汇 w 的消歧特征提取算法如下所示:

```
Node * Find_WNode( Node * pNode, CString w) {
    if( pNode == NULL)
        return NULL;
    if( ! pNode→IsLeafNode( ) ) {
        if( pNode→Word == w)
            return pNode;
    }
    else{
        if( pNode→RBrother! = NULL)
            Find_WNode( pNode→RBrother, w) ;
    }
}
CString Feature_Extraction( Node * TS, CString w) {
    Node * p = Find_WNode( TS, w) ;
    CString f = "";
    if( p! = NULL) {
    if( ! p→LBrother→IsLeafNode( ) )
        f = "LS = " + p→LBrother→POS;
    else
        f = "LP = " + p→LBrother→Phrase + "LCS = " + p→LBro-ther→
CPOS;
    if( ! p→RBrother→IsLeafNode( ) )
        f = f + "RS = "p→RBrother→POS;
    else
        f = f + "RP = " + p→RBrother→Phrase + "RCS = " + p→RBrot-
her→CPOS;
        f = f + "PP = " + p→Parent→Phrase + "PCS = " + p→Parent→
CPOS;
    }
    return f;
}
```

在句法树上,当歧义词的左(右)兄弟节点为空时,则将其左(右)邻接单词的词性信息补充进消歧特征中。对于含有歧义词汇“说”的汉语句子 CS,消歧特征的获取过程如下所示:

汉语句子 CS:他 说 到 这 一 问题,开始 沉思 了 。

句法分析结果 TS:S[SS[他/r VO[说/vg VO[到/vg BNP[BMP[这/r 一/m] 问题/ng]]]]],/wo VP[VP[开始/vg 沉思/vg] 了/usde] 。/wj]

其中,“[”和“]”成对出现,描述了一个短语结构。“[”之前的标注为短语句法类型,如 S、SS、VO、BNP、BMP 和 VP。S 为句子结构;SS 表示子句结构;VO 为动宾结构;BNP 表示基本形容词短语;BMP 为基本量词短语;VP 表示动词短语。“/”之后的标注为单词的词性类别。其中,r 是代词;vg 为动词;m 是数词;ng 为名词;wo 是逗号;usde 为助词;wj 是句号。每一对“[”和“]”都刻画了一个短语结构,也描述了句子的一个层次结构。为了突显内层结构对外层结构的支配作用和外层结构对内层结构的解释说明作用,由外至内,每一对“[”和“]”作为一个独立的节点向下扩展,形成了一棵多叉树,即句法分析树。汉语句子 CS 的句法树如图 1 所示。在短语“VO[到/vg BNP[BMP[这/r 一/m] 问题/ng]] ”中,VO 表明其组成形式只能是动词成分加名词成分。该短语包括节点“到/vg”和节点 BNP,表明其句法类型只能是 VO。

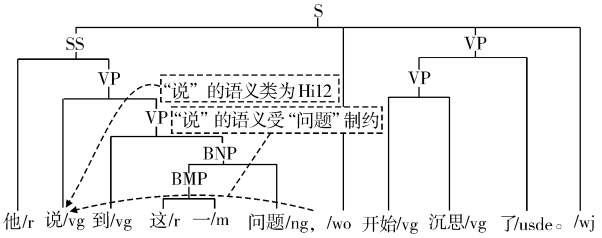

图 1　汉语句子 CS 的句法分析树

在汉语中,单词“说”是一个多义词。在《同义词词林》中共包含四种词义。第一种词义类别是 Hi12,汉语同义词为“讲话”和“谈天”;第二种词义类别是 Hg12,汉语同义词为“解释”“注释”“引证”和“参阅”;第三种词义类别是 Hi16,汉语同义词为“介绍”和“推荐”;第四种词义类别是 Hi34,汉语同义词为“鼓励”和“劝告”。从图 1 中可以发现,“说”的词义类别应该为 Hi12。如果开设直线窗口,那么“说”的左、右邻接词单元分别为“他/r”和“到/vg”。若以词性作为判别依据,则消歧特征为 r + vg。但是从语法角度来看,“说”的词义直接受词单元“问题/ng”制约。词单元“问题/ng”离“说/vg”的距离比较远,无法从直线窗口中提取,然而却能够从树状窗口中获取。使用消歧特征提取方法,可以获得“说”的右兄弟节点的句法类别 VO。右兄弟的核心节点为词单元“问题/ng”,因此右兄弟节点的核心词性为 ng。此处核心节点的计算过程为:core(VO[到/vg BNP[BMP[这/r 一/m] 问题/ng]]) = core(BNP[BMP[这/r 一/m] 问题/ng]) = “问题/ng”。父亲节点的句法信息为 VP,核心节点词性为 ng。在句法树中,“说”无左兄弟节点,所以取它的左侧邻接词单元“他/r”的词性 r。从树状窗口中获取的消歧特征为 r + VO + ng + VP + ng。对比两种消歧特征,在直线窗口中,“说”的右兄弟词性为 vg;在树状窗口中,“说”的右兄弟词性为 ng。由此可以看出树状窗口所提取的消歧特征更为合理。

2　词义消歧分类器

贝叶斯分类器适合处理单点分类问题。单点分类是指序

列中的符号所对应的状态与序列中的其他符号所对应的状态无关。在给定的上下文环境中,正确的词义出现的概率应该最大,歧义词 w 的词义判别规则如式(1)所示。

$$S = \arg\max_{i=1,2,\cdots,n} P(s_i \mid F_L, F_R, F_P) \tag{1}$$

$$P(s_i \mid F_L, F_R, F_P) \approx P(F_L, F_R, F_P \mid s_i) \cdot P(s_i) \approx$$

$$P(F_L \mid s_i) \cdot P(F_R \mid s_i) \cdot P(F_P \mid s_i) \cdot P(s_i)$$

为了计算条件概率 $P(F_L, F_R, F_P \mid s_i)$,此处使用了贝叶斯假设,即各个上下文词单元都是相互独立的,所抽取的句法和词性信息也是相互独立的。其中,F_L 为左兄弟节点的消歧特征。若左兄弟为叶节点,则 F_L 为词性信息;若左兄弟为非叶节点,则 F_L 为句法类别和核心节点词性信息。F_R 为右兄弟节点的消歧特征。若右兄弟为叶节点,则 F_R 为词性信息;若右兄弟为非叶节点,则 F_R 为句法类别和核心节点词性信息。F_P 为父节点消歧特征,包括句法类别和核心节点词性信息。$s_i (i = 1, 2, \cdots, n)$ 为 w 的 n 种词义类别。$P(X)$ 为 X 的出现概率。$P(F_t \mid s_i)$ 的计算过程如下:若 F_t 是来自叶节点的特征,则 $P(F_t \mid s_i)$ 为词性出现的概率;若 F_t 是来自非叶节点的特征,则 $P(F_t \mid s_i)$ 为词性出现的概率与句法类别出现的概率之积。其中,$t = L, R, P$。整个贝叶斯模型的求解过程遵循了在给定条件下决策事件错误率最小的原则。

3 实验分析

为了衡量本文所提出方法的性能,收集整理了包含歧义词汇的110个汉语句子。首先使用分词工具对这110个汉语句子进行单词切分,两名人工标注者对自动分词结果进行人工校正;然后采用词性标注工具来标记每一个单词的词性,两名人工标注者对自动词性标注结果进行人工校正;最后利用句法分析工具对每一个汉语句子进行结构标注。所使用的汉语分词工具、词性标注工具和句法分析工具是由哈尔滨工业大学计算机科学与技术学院语言语音教育部—微软重点实验室开发的。按照《同义词词林》手工标注歧义词汇的语义类。将这110个汉语句子分成两部分,一部分为训练数据集,另一部分为测试数据集。在训练数据集中共包含80个汉语句子;在测试数据集中共包含30个汉语句子。为了衡量句法信息和词性信息对词义消歧的影响,在训练数据集上进行了统计。在给定的词义类别下,统计每种句法类别出现的次数,其结果如表1所示。

表1 句法信息对词义消歧的影响

	VP	VO	SS	BVP	VJ	NP	VC	VOO	BNP	S
Hi12	14	16	6	2	1	1	1	2		
Hg12	8	10	2	1		3			1	3
Hi16	1	3	1				1			

从表1可以看出,对于“说”而言,句法类别VP、VO和SS出现的次数较多,其中,SS的消歧能力比较高。在给定的词义类别下,统计每种词性出现的次数,其结果如表2所示。从表2可以看出,对于“说”而言,词性信息nx、ng、r、used、vg、wo和wj出现的次数较多,其中,r、used和wj的消歧能力比较高。

表2 词性信息对词义消歧的影响

类别	nx	ng	r	i	d	t	f	used	wyr	vg	wo	wj	m	a	q	c
Hi12	6	17	13		2	2	4	8		7	15	10	1	1		
Hg12	3	11	2	1			3	2	1	7	19	3	1		1	2
Hi16	1	3	1		2					1	2			1		

为了比较本文所提出方法的性能,共进行了两组实验。实验1使用了开设词窗的方法[13],利用歧义词汇的左右邻接单词的词性作为消歧特征,使用朴素贝叶斯模型作为词义消歧分类器。实验2采用本文所提出的方法来获取消歧特征,利用式(1)作为词义消歧分类器。使用训练数据分别对两组实验中的分类器进行训练,然后采用优化后的分类器对测试数据进行词义分类。实验结果如表3所示。从表3中可以看出,实验2的分类准确率要高于实验1,达到了66.7%。其原因是在实验2中,利用句法信息来指导消歧过程,所获取的消歧特征对词义分类的效果比较好。

表3 两组实验词义消歧的准确率

实验	正确分类数	错误分类数	准确率/%
1	18	12	60
2	20	10	66.7

4 结束语

本文将句法信息引入词义消歧模型之中,对包含歧义词汇的汉语句子进行分词、词性标注和句法分析处理。从句法树中抽取消歧特征,同时使用朴素贝叶斯模型进行词义分类。使用带有语义标注的语料来训练分类器,然后对测试数据进行分类。对比实验表明词义消歧的性能有所提升。

参考文献:

[1] BROWN P, FPIETRA S A D, PIETRA V J D, *et al*. Word sense disambiguation using statistical methods[C]//Proc of the 29th Annual Meeting on Association for Computational Linguistics. Stroudsbury: ACL, 1991: 264-270.

[2] 李旭, 刘国华, 张东明. 一种改进的汉语全文无指导词义消歧方法[J]. 自动化学报, 2010, 36(1): 184-187.

[3] PEDERSEN T. A simple approach to building ensembles of naive Bayesian classifiers for word sense disambiguation[C]//Proc of the 1st Meeting of the North American Chapter of the Association for Computational Linguistics. Stroudsburg: ACL, 2000: 63-69.

[4] PEDERSEN T. Evaluating the effectiveness of ensembles of decision trees in disambiguating senseval lexical samples[C]//Proc of the 40th Annual Meeting of the Association for Computational Linguistics. Stroudsburg: ACL, 2002: 81-87.

[5] 何径舟, 王厚峰. 基于特征选择和最大熵模型的汉语词义消歧[J]. 软件学报, 2010, 21(6): 1287-1295.

[6] 吴云芳, 王森, 金澎. 多分类器集成的汉语词义消歧研究[J]. 计算机研究与发展, 2008, 45(8): 1354-1361.

[7] Van De CRUYS T, APIDIANAKI M. Latent semantic word sense induction and disambiguation[C]//Proc of the 49th Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2011: 1476-1485.

[8] 鲁松, 白硕, 黄雄. 基于向量空间模型中义项词语的无导词义消歧[J]. 软件学报, 2002, 13(6): 1082-1089.

[9] CHEN Yi-qun, YIN Jian. Sense rank AALesk: a semantic solution for word sense disambiguation[C]//Proc of the 2nd International Conference on Fuzzy Systems and Knowledge Discovery. Berlin: Springer-Verlag, 2005: 710-717.

[10] NAVIGLI R, PONZETTO S P. Joining forces pays off: multilingual joint word sense disambiguation[C]//Proc of Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning. Stroudsburg: ACL, 2012: 1399-1410.

[11] NAVIGLI R, PONZETTO S P. Multilingual WSD with just a few lines of code: the BabelNet API[C]//Proc of the 50th Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 2012: 67-72.

[12] 卢志茂, 刘挺. 基于依存分析和贝叶斯网络的无指导汉语词义消歧[J]. 高技术通讯, 2004, 14(2): 7-11.

[13] YAROWSKY D. Unsupervised word sense disambiguation rivaling supervised methods[C]//Proc of the 33rd Annual Meeting on Association for Computational Linguistics. Stroudsburg: ACL, 1995: 189-196.