结合语义知识的汉语词义消歧

张春祥[1,2],邓 龙[3],高雪瑶[3],卢志茂[2]
ZHANG Chunxiang[1,2], DENG Long[3], GAO Xueyao[3], LU Zhimao[2]

1.哈尔滨理工大学 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
3.哈尔滨理工大学 计算机科学与技术学院,哈尔滨 150080
1.School of Software, Harbin University of Science and Technology, Harbin 150080, China
2.College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China
3.School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China

ZHANG Chunxiang, DENG Long, GAO Xueyao, et al. Chinese word sense disambiguation with semantic knowledge. Computer Engineering and Applications, 2016, 52(3):119-122.

Abstract: Word sense disambiguation is an important problem in natural language processing. In order to improve the precision of word sense disambiguation, semantic knowledge of left and right word units is mined starting from the target polysemous word. Based on the Bayesian model, a new method of word sense disambiguation is proposed with semantic information of left and right word units. SemEval-2007: Task#5 is used as training corpus and test corpus. The classifier of word sense disambiguation is optimized. Then the optimized classifier is tested. Experimental results show that the precision of word sense disambiguation is improved.

Key words: word sense disambiguation; polysemous word; Bayesian model; semantic information

摘 要:词义消歧一直是自然语言处理领域中的关键性问题。为了提高词义消歧的准确率,从目标歧义词汇出发,挖掘左右词单元的语义知识。以贝叶斯模型为基础,结合左右词单元的语义信息,提出了一种新的词义消歧方法。以SemEval-2007: Task#5作为训练语料和测试语料,对词义消歧分类器进行优化,并对优化后的分类器进行测试。实验结果表明:词义消歧的准确率有所提高。

关键词:词义消歧;歧义词汇;贝叶斯模型;语义信息
文献标志码:A **中图分类号:**TP391.2 doi:10.3778/j.issn.1002-8331.1402-0041

1 引言

词义消歧的目的是确定歧义词汇在特定上下文环境中的意义。词义消歧的准确率在机器翻译、信息检索、文本分析和自动文摘等相关应用中都有着很大的影响。杨陟卓在传统的网络模型中引入了词语距离信息,提出了基于词语距离的网络图词义消歧方法[1]。范冬梅根据贝叶斯假设给出了一种基于信息增益的特征选择方法,通过挖掘上下文词语的位置信息来改善词义分类效果[2]。鲁松提出了一种基于向量空间模型的有监督学习方法,通过计算上下文向量与义项向量之间的距离来进行消歧[3]。Huang结合半监督统计学习技术给出了一种新的词义消歧算法,通过设定多种阈值来扩展训练数据[4]。Niu提出了一种混合数据自动划分方法,通过改善扩展标记传播算法的分类结果来提高词义消歧质量[5]。

基金项目:国家自然科学基金(No.60903082);教育部春晖计划(No.S2009-1-15002);中国博士后科学基金项目(No.2014M560249);黑龙江省自然科学基金(No.F2015041)。

作者简介:张春祥(1974—),男,博士,教授,硕士生导师,研究领域为自然语言处理,E-mail:z6c6x6@aliyun.com;邓龙(1989—),男,硕士研究生,研究领域为自然语言处理;高雪瑶(1979—),女,博士,副教授,硕士生导师,研究领域为自然语言处理和图形学;卢志茂(1972—),男,博士,教授,博士生导师,研究领域为自然语言处理。

收稿日期:2014-02-10 **修回日期:**2014-07-17 **文章编号:**1002-8331(2016)03-0119-04
CNKI**网络优先出版:**2014-08-29, http://www.cnki.net/kcms/doi/10.3778/j.issn.1002-8331.1402-0041.html

Le探讨了半监督词义消岐中的无标注数据的使用问题,结合分类组合策略给出了相应的解决方案,同时给出了自举算法的几个新变种[6]。Le以Dempster-Shafer证据理论和有序加权平均运算符为基础提出了词义消歧分类器的加权组合框架[7]。Huang给出了一种基于位置的消歧算法,以测度上下文的相似性[8]。Yoon使用未标注语料和机读词典来实现自动词义消歧。从未标注语料中学习词汇之间的相似性矩阵,并使用了机读词典中的释义[9]。Tristan提出了一种可模块化和可扩展的词义消歧框架DKPro[10]。Katrin给出了两种基于分级模式的词义标注方法。在词义相似性标注中,对每一个词典语义的应用范围按序进行排列。在Usim标注中,直接对词条的适用性进行排序[11]。李旭提出了一种改进的全文无监督词义消歧模型,结合互信息和Z-测试结果来选取特征[12]。在半监督学习框架中,Yu使用平行双语语料来解决监督化词义消歧方法中语义类定义和训练数据获取的难题[13]。Tacoa使用自动生成的词典来扩充上下文,克服数据稀疏的影响,提高了词义消歧质量[14]。Liu以Google距离为基础提出了一种新的无监督词义消歧算法[15]。

本文提出了一种基于语义类别的汉语词义消歧方法。在目标歧义词汇所在的句子中,开设消歧窗口,提取左右词单元。将左右词的语义代码作为消歧特征,采用贝叶斯模型进行词义消歧。实验结果表明:消歧的性能有所提高。

2　词义消歧特征的抽取

文本消歧特征可以使用一定语言环境中的词单元来表示,主要体现为词单元的语言学信息之间的共现。词汇的语义类别是由其所处的上下文环境和本身所具有的词义决定的。其上下文的语义为确定歧义词汇的语义提供了一定的指导信息。以歧义词汇为中心,通过开设窗口来提取这些上下文信息。本文通过选择歧义词汇的左右词单元的语义类别作为消歧特征,来判别它的真实含义。

对于含有歧义词“*说明*”的汉语句子,消歧特征的提取过程如下所示:

汉语句子:一些学生和家长的看法似乎更能*说明*问题。

分词结果:一些 学生 和 家长 的 看法 似乎 更能 *说明* 问题 。

词性标注结果:一些/m 学生/ng 和/c 家长/ng 的/ur 看法/ng 似乎/d 更/d 能/vg *说明*/vg 问题/ng 。/wj

《同义词词林》给出了汉语词汇的语义代码,可以为词义消歧过程提供丰富的语义知识。在词条中,用语义代码来表示语义分类体系。例如,“说明”的语义代码为Dk15,表示其处于D大类,k中类和15小类。可以把整个语义分类体系想象成一棵语义树,根结点的儿子是所有大类,某个大类的儿子是它下属的中类,叶子结点为各个小类。本文将采用《同义词词林》作为语义词典来确定左、右词汇的语义代码,同时获取目标词汇的语义类别。在《同义词词林》中,“说明”共有5种不同的语义。第一个语义类别为Dk15,其汉语同义词为“便条”、“处方”和“说明书”;第二个语义类别为Dk23,其汉语同义词为“按语”、“批语”、“引文”和“注解”;第三个语义类别为Hg12,其汉语同义词为“解释”、“注释”、“引证”和“参阅”;第四个语义类别为Hi14,其汉语同义词为“表示”、“表达”、“阐明”和“表露”;第五个语义类别为Ja04,其汉语同义词为“显示”、“预示”和“证明”。

汉语词汇“说明”在《同义词词林》中的树状结构如图1所示。

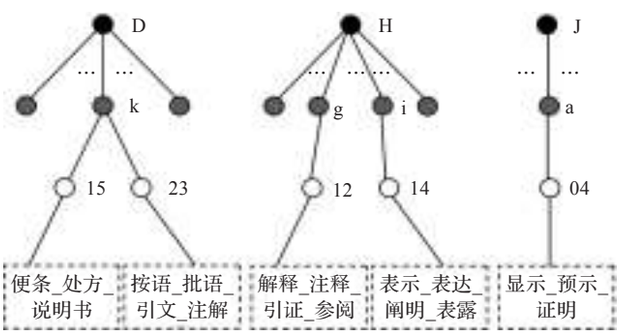

图1　“说明”的语义树结构

在此句中“说明”为歧义词。根据其上下文语境,可以推断它的真实语义类别为Ja04。

在《同义词词林》中,“能”也是一个歧义词汇。其语义类代码和汉语同义词为:(1)Ee17,“能干”和“无能”;(2)Dd14,“力气”、“力量”和“能量”;(3)De04,“智慧”、“才能”、“能力”和“功夫”;(4)Gc02,“能”、“能够”和“不能”。这意味着决定词义的上下文是歧义的,不能为消歧过程提供任何有意义的指导信息。本文将采用Dice系数来确定词单元的语义类别。

Dice系数是一种集合相似性度量函数,通常用于计算两个样本的相似度,也可以计算两个字符串之间的相似程度。在《同义词词林》中,利用Dice系数来确定词汇 w 的语义类 S_w,其计算如公式(1)所示。

$$S_w = \arg\max_{s \in SenSet(w)} SenSim(w, w_s) \tag{1}$$

$$SenSim(w, w_s) = \frac{sim(w, w_s)}{length(w) + length(w_s)} \tag{2}$$

此处,$SenSet(w)$ 为汉语单词 w 的语义类集合,例如:$SenSet$(能)={Ee17,Dd14,De04,Gc02};w_s 为汉语单词 w 在语义类别 S 下的同义词词汇,例如:w_{Ee17}=“能干_无能”;$sim(w, w_s)$ 为汉语单词 w 和同义词词汇 w_s 之间

共有的字符个数，例如：*sim*(能，能干_无能)=3，“能”与“能干_无能”之间共有3个相同的汉字字符“能”；*length*(X)为汉语字符串 X 的长度，例如：*length*(能)=1，*length*(能干_无能)=4。使用公式(2)计算语义类代码Ee17的相似度为0.6。

按如上方法可以计算出“能”的所有语义类代码的相似度，其结果如表1所示。

表1　“能”的语义类代码的相似度

序号	词义类	汉语同义词汇	相似度
1	Ee17	能干_无能	0.60
2	Dd14	力气_力量_能量	0.29
3	De04	智慧_才能_能力_功夫	0.33
4	Gc02	能_能够_不能	0.67

按照公式(1)，可以确定汉语单词“能”的语义类代码为Gc02。

在《同义词词林》中，“问题”也是一个歧义词汇。其语义类代码和汉语同义词为：(1)Dk03，“科学”、“学科”和“学业”；(2)Dk07，“题材”、“体裁”和“布局”；(3)Da01，“事情”、“事件”和“事故”；(4)Da04，“疑问”和“嫌疑”；(5)Dd09，“关键”和“核心”。按照公式(1)，可以确定汉语单词“问题”的语义类代码为Da04。

从以上句子中所抽取的消歧特征为Gc02和Da04。根据消歧特征Gc02和Da04来判断汉语歧义词汇“说明”的真实含义。

3　基于语义类的贝叶斯分类器

贝叶斯分类器是由托马斯·贝叶斯提出来的，使用概率计算方法来实现。它根据某一事件过去的发生概率来推断这一事件当前的发生概率。在词义消歧问题中，贝叶斯决策规则的具体表现形式为：若 $P(S_j|Context)\geqslant P(S_i|Context)(i=1,2,\cdots,m)$，则歧义词汇 w 的语义为 S_j。其中，*Context* 为歧义词汇 w 所在的上下文环境。通常，*Context* 由 w 两侧的词单元组成，为消歧过程提供必要的指导信息。歧义词汇 w 有 m 种语义类别 $S_1,S_2,\cdots,S_m$。在 *Context* 中，它的真实语义类别为 S_j。$P(X)$ 为 X 出现的概率。在上下文环境 *Context* 中，如果歧义词汇 w 取语义类别 S_j 的概率要大于其余任何一种语义类别 $S_i(j\neq i)$，则词汇 w 的语义类别应该判定为 S_j。贝叶斯决策具有最小的误差概率。

对于歧义词汇 w 而言，它有 m 种语义 $S_1,S_2,\cdots,S_m$，上下文消歧特征分别为 F_L 和 F_R。其中，F_L 和 F_R 为语义类代码。《同义词词林》的语义代码共分三层，$F_L=f_{l1}f_{l2}f_{l3}$，$F_R=f_{r1}f_{r2}f_{r3}$。根据 $f_{l1}f_{l2}f_{l3}$ 和 $f_{r1}f_{r2}f_{r3}$ 来判断歧义词汇 w 语义的过程如公式(3)所示。

$$\begin{aligned}S&=\arg\max_{i=1,2,\cdots,m}P(S_i|F_L,F_R)=\\&\arg\max_{i=1,2,\cdots,m}P(S_i|f_{l1}f_{l2}f_{l3},f_{r1}f_{r2}f_{r3})=\\&\arg\max_{i=1,2,\cdots,m}\frac{P(S_i,f_{l1}f_{l2}f_{l3},f_{r1}f_{r2}f_{r3})}{P(f_{l1}f_{l2}f_{l3},f_{r1}f_{r2}f_{r3})}\approx\\&\arg\max_{i=1,2,\cdots,m}P(S_i,f_{l1}f_{l2}f_{l3},f_{r1}f_{r2}f_{r3})=\\&\arg\max_{i=1,2,\cdots,m}P(f_{l1}f_{l2}f_{l3},f_{r1}f_{r2}f_{r3}|S_i)\cdot P(S_i)\approx\\&\arg\max_{i=1,2,\cdots,m}P(f_{l1}f_{l2}f_{l3}|S_i)\cdot P(f_{r1}f_{r2}f_{r3}|S_i)\cdot P(S_i)\end{aligned}\tag{3}$$

第一层编码 f_{l1} 和 f_{r1} 用大写英文字母来表示，取值范围为A，B，…，L。第二层编码 f_{l2} 和 f_{r2} 用小写英文字母来表示，取值随着 f_{l1} 和 f_{r1} 的变化而变化。当 f_{i1}=A，D，H时，f_{i2}=a，b，…，n；当 f_{i1}=B时，f_{i2}=a，b，…，r；当 f_{i1}=C时，f_{i2}=a，b；当 f_{i1}=E，K时，f_{i2}=a，b，…，f；当 f_{i1}=F时，f_{i2}=a，b，…，d；当 f_{i1}=G时，f_{i2}=a，b，c；当 f_{i1}=I时，f_{i2}=a，b，…，g；当 f_{i1}=J时，f_{i2}=a，b，…，e；当 f_{i1}=L时，f_{i2}=NULL；其中，$i=l,r$。第三层编码 f_{l3} 和 f_{r3} 用两位数字来表示。

4　实验

为了衡量本文所提出方法的性能，以SemEval-2007 #Task5作为实验语料。选取其中常用的13个歧义词汇“儿女”、“气息”、“望”、“开通”、“机组”、“气象”、“表面”、“单位”、“菜”、“中医”、“本”、“赶”和“旗帜”。将包含这些歧义词汇的句子抽出来，分为训练语料和测试语料两部分，具体分布如表2所示。

表2　训练语料和测试语料的分布

歧义词汇	训练语料句子数	测试语料句子数	总句子数
儿女	60	20	80
气息	39	14	53
望	37	13	50
开通	56	20	76
机组	38	14	52
气象	47	16	63
表面	53	18	71
单位	50	17	67
菜	52	18	70
中医	43	16	59
本	68	25	93
赶	56	18	74
旗帜	50	18	68

为了比较本文所提出方法的性能，共进行了两组实验。在第一组实验中，使用了传统的开设词窗的方法，利用歧义词汇的左右邻接单词的词形作为消歧特征，采用贝叶斯模型作为词义消歧分类器。在第二组实验中，针对歧义词汇的训练语料，利用本文所提出的方法来抽取其上下文消歧特征，对公式(3)所示的消歧分类器中

的参数进行估计。同时,针对它的测试语料,使用本文所提出的方法来抽取其上下文消歧特征,利用优化后的消歧分类器来自动地选择歧义词汇的语义类别。将自动选择的语义类别与人工标注的语义类别进行比照,统计自动消歧的精确率。歧义词汇“儿女”、“气息”、“望”、“开通”、“机组”、“气象”、“表面”、“单位”、“菜”、“中医”、“本”、“赶”和“旗帜”的测试语料的消歧精确率如表3所示。

表3 测试语料的消歧精确率 %

	基于词形的消歧精确率	基于语义代码的消歧精确率
儿女	50.0	85.0
气息	64.3	71.4
望	69.2	69.2
开通	70.0	70.0
机组	71.4	71.4
气象	37.5	43.8
表面	50.0	61.1
单位	47.1	47.1
菜	33.3	33.3
中医	37.5	50.0
本	68.0	72.0
赶	27.8	27.8
旗帜	55.6	55.6

从表3中可以看出:基于语义代码的消歧精确率要大于等于基于词形的消歧精确率。对词汇“儿女”而言,其精确率的增长达到了35%;对词汇“气息”而言,其精确率的增长达到了7.1%;对词汇“气象”而言,其精确率的增长达到了6.3%;对词汇“表面”而言,其精确率的增长达到了11.1%;对词汇“中医”而言,其精确率的增长达到了12.5%;对词汇“本”而言,其精确率的增长达到了4.0%。对词汇“望”、“开通”、“机组”、“单位”、“菜”、“赶”和“旗帜”而言,其精确率保持不变。其原因是:在第二组实验中,利用左右词汇单元的语义代码来指导消歧过程,具有一定的泛化能力。相对于基于词形的方法而言,在参数估计过程中减少了数据稀疏的影响,所获取的消歧特征对词义分类的效果比较好。

5 结束语

本文将语义信息引入词义消歧模型之中。在汉语句子中,以歧义词汇为中心定位其左右词汇单元。通过查阅《同义词词林》来确定左右词汇单元的语义类别,将其作为词义消歧特征。以左右词汇单元的语义类别为基础,使用贝叶斯模型来判断歧义词汇的真实语义。实验结果表明:消歧的性能有所提高。

参考文献:

[1] 杨陟卓,黄河燕.基于词语距离的网络图词义消歧[J].软件学报,2012,23(4):776-785.

[2] 范冬梅,卢志茂,张汝波.基于信息增益改进贝叶斯模型的汉语词义消歧[J].电子与信息学报,2008,30(12):2926-2929.

[3] 鲁松,白硕,黄雄.基于向量空间模型的有导词义消歧[J].计算机研究与发展,2001,38(6):662-667.

[4] Huang Zhehuang, Chen Yidong, Shi Xiaodong.A novel word sense disambiguation algorithm based on semi-supervised statistical learning[J].International Journal of Applied Mathematics and Statistics,2013,43(13):452-458.

[5] Niu Zhengyu, Ji Donghong, Tan Chew Lim.Learning model order from labeled and unlabeled data for partially supervised classification, with application to word sense disambiguation[J].Computer Speech and Language,2007,21(4):609-619.

[6] Le Anh-Cuong, Akira S, van Nam H.Semi-supervised learning integrated with classifier combination for word sense disambiguation[J].Computer Speech and Language,2008,22(4):330-345.

[7] Le Anh-Cuong, van Nam H, Akira S.Combining classifiers for word sense disambiguation based on dempster-shafer theory and owa operators[J].Data and Knowledge Engineering,2007,63(2):381-396.

[8] Huang Shilin, Zheng Xiaolin, Kang Haixiao, et al.Word sense disambiguation based on positional weighted context[J].Journal of Information Science,2013,39(2):225-237.

[9] Yeohoon Y.Unsupervised word sense disambiguation for Korean through the acyclic weighted digraph using corpus and dictionary[J].Information Processing and Management,2006,42(3):710-722.

[10] Miller T, Erbs N, Zorn H P.DKPro WSD—a generalized UIMA-based framework for word sense disambiguation[C]//Proceedings of the 51st Annual Meeting of the Association for Computational Linguistics,2013:37-42.

[11] Erk K, McCarthy D, Gaylord N.Measuring word meaning in context[J].Computational Linguistics,2013,39(3):511-554.

[12] 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.

[13] Yu Mo, Wang Shu, Zhu Conghui, et al.Semi-supervised learning for word sense disambiguation using parallel corpora[C]//Proceedings of the 8th International Conference on Fuzzy Systems and Knowledge Discovery,2011:1490-1494.

[14] Francisco T, Danushka B.A context expansion method for supervised word sense disambiguation[C]//Proceedings of the 6th International Conference on Semantic Computing,2012:339-341.

[15] Liu Pengyuan, Xue Yongzeng, Li Shiqi, et al.Minimum normalized google distance for unsupervised multilingual Chinese-English word sense disambiguation[C]//Proceedings of the 4th International Conference on Genetic and Evolutionary Computing,2010:252-255.