句法信息指导的汉语词义消歧

张春祥[1,2],栾 博[1],高雪瑶[1],卢志茂[3]
ZHANG Chunxiang[1,2], LUAN Bo[1], GAO Xueyao[1], LU Zhimao[3]

1.哈尔滨理工大学 计算机科学与技术学院,哈尔滨 150080
2.哈尔滨理工大学 软件学院,哈尔滨 150080
3.哈尔滨工程大学 信息与通信工程学院,哈尔滨 150001
1.School of Computer Science and Technology, Harbin University of Science and Technology, Harbin 150080, China
2.School of Software, Harbin University of Science and Technology, Harbin 150080, China
3.College of Information and Communication Engineering, Harbin Engineering University, Harbin 150001, China

ZHANG Chunxiang, LUAN Bo, GAO Xueyao, et al. Chinese word sense disambiguation directed by syntactic information. Computer Engineering and Applications, 2015, 51(5):142-145.

Abstract: The task of word sense disambiguation is to make computers choose the correct sense of an ambiguous word in a given context. It is important for problems in natural language processing, such as information retrieval, machine translation, text classification and automatic summarization. In this paper, a new method of word sense disambiguation is proposed, where syntactic information is introduced. The parsing tree of its context including the ambiguous word is built. Disambiguation features are extracted including parsing information, part of speech and word information. The Bayesian model is used to build word sense disambiguation classifier. Experimental results show that accuracy rate of disambiguation is improved and arrives at 65%.

Key words: word sense disambiguation; syntactic information; disambiguation features; Bayesian model

摘 要:词义消歧要解决如何让计算机理解多义词在上下文中的具体含义,对信息检索、机器翻译、文本分类和自动文摘等自然语言处理问题有着十分重要的作用。通过引入句法信息,提出了一种新的词义消歧方法。构造歧义词汇上下文的句法树,提取句法信息、词性信息和词形信息作为消歧特征。利用贝叶斯模型来建立词义消歧分类器,并将其应用到测试数据集上。实验结果表明:消歧的准确率有所提升,达到了65%。

关键词:词义消歧;句法信息;消歧特征;贝叶斯模型
文献标志码:A **中图分类号:**TP391.2 doi:10.3778/j.issn.1002-8331.1303-0429

1 引言

在汉语中,一个词汇可能包含几层含义。词义消歧就是根据词汇出现的特定上下文来判断它的正确含义。例如:词汇“面”在句子“这位摄影师说:作案人蒙着面。”中是一个歧义词。在《同义词词林》中,它的含义是“脸”。同样,在句子“雷锋车是港城一面鲜艳的精神文明旗帜。”中,也是歧义的。其含义是“个”。

目前,词义消歧方法大致分为有监督消歧和无监督消歧两种。朱静波提出了基于对数模型的消歧方法,把词义消歧看作是根据给定输入条件来选择具有最大概率词义的过程[1]。鲁松给出了一种基于向量空间模型的有指导学习方法,将歧义词的义项和上下文分别映射到向量空间中,通过计算上下文向量与义项向量之间的距离来进行消歧[2]。刘风成利用AdaBoost.MH对决策树产生的弱规则进行加强,经过迭代后得到了一个准确度更高的分类规则[3]。张仰森探究了《知网》的词语与义原

基金项目:黑龙江省教育厅科学技术研究项目(No.12531106)。
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;栾博(1988—),女,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;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。E-mail:z6c6x666@163.com
收稿日期:2013-03-27 **修回日期:**2013-05-17 **文章编号:**1002-8331(2015)05-0142-04
CNKI**网络优先出版:**2013-06-17, http://www.cnki.net/kcms/detail/11.2127.TP.20130617.0925.015.html

之间的关系，将从训练语料中获取的词语搭配信息转换为义原搭配信息，同时结合传统的上下文，使用隐最大熵原理来实现词义消歧[4]。陈浩利用歧义词的各个义项的同义词来构造二阶上下文向量，使用*K*-means算法进行聚类，通过计算相似度来进行消歧[5]。Samuel提出了一种无监督语义标注方法，将分布相似的词汇作为训练数据，来训练词义消歧分类器[6]。李旭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[7]。Stefano给出了一种新的术语自举获取方法，从Web上自动地抽取不同领域的术语。同时，将这些术语作为语义知识提出了一种无监督的领域词义消歧方法[8]。Buscaldi考虑了领域信息对消歧过程的影响，在概念密度的计算中添加了互领域权重调和项，以提高词义判别能力[9]。Mihalcea在语义依赖图中，使用随机行走策略对词汇序列进行消歧[10]。范冬梅通过计算信息增益，来挖掘上下文词语的位置信息，以提高Bayes模型的知识获取效率，从而改善词义分类效果[11]。Navigli利用多语联合分布来计算语义相关性，该方法利用了词典信息和多语知识库中的语义信息[12]。刘鹏远提出了一种基于双语词汇Web间接关联的无指导消歧方法，给出了词汇歧义可由双语词汇间接关联度决定的前提假设[13]。Navigli提出了一种多语联合词义消歧方法，利用多语知识库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[14]。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，充分考虑了词语距离对消歧效果的影响，即距离较近的词对歧义词的词义有较大的影响，而距离较远的词对其词义有较弱的影响[15]。

本文提出了一种新的词义消歧方法。对歧义词汇所在的汉语句子进行句法分析，将句法信息、词性信息和词汇信息作为消歧特征。同时，采用贝叶斯分类模型进行消歧。实验结果表明：消歧的性能有所提升。

2　消歧特征获取

文本层面的消歧特征主要由一定的上下文中的词来表示，它包括局部词汇和局部词汇共现。这些特征来自于句子的表层信息，只需要对汉语句子进行基本的词语切分即可获得，而且也可以达到较高的精度。在选择单词作为消歧特征时，往往过滤掉字母和停用词，例如：“￥”、“%”、“@”、“；”、“！”和“在”等。其原因是：它们对消歧过程不能提供任何帮助。词性是词汇的语法类别，由词条的句法或形态行为决定。常用的语法类别有名词、动词和形容词等。词性层面的信息作为一种语法标记对确定词义有着较强的暗示作用。通常，选择歧义词及其左右词汇的词性作为消歧特征。句法层面的消歧特征主要指句法层次结构信息，诸如：主-谓-宾和介-宾短语等。在这些句法结构中，有些可以为词义消歧过程提供指示性信息。

通常，上下文环境中的词汇与歧义词之间的距离比较远。如果消歧窗口开设得过大，将会引起噪声和数据稀疏问题。目前，尽管句法分析的结果还不能令人满意，但是对词义消歧而言仍可能提供部分的指导信息。句法分析是利用语言学知识来确定句子的层次结构，即所包含的句法单元及其之间的相互关系，典型的技术有短语结构语法。在短语结构语法中，将句子拆分成短语并将所有的短语组织成树来描述其结构，每个短语对应树中的一个结点。结点分层组织，包括基本短语和复杂短语两种形式，其中复杂短语是基本短语的嵌套组合。本文将短语句法树用于词义消歧过程。

此处，消歧特征包括：词汇信息、词性信息、句法树中的句法信息。先根遍历句法树，提取歧义词的左兄弟结点、右兄弟结点和父结点。对于一个包含歧义词 w 的汉语句子 S，消歧特征的提取算法如下所示。

(1)使用汉语分词工具对 S 进行分词，结果为 $S_{\text{word\_seg}}$。

(2)利用汉语词性标注工具对 $S_{\text{word\_seg}}$ 的每个单词进行词性标注，结果为 $S_{\text{part\_of\_speech}}$。

(3)使用汉语句法分析工具来建立 $S_{\text{part\_of\_speech}}$ 的句法树，结果为 T_S。

(4)先根遍历句法树 T_S，获取包含歧义词 w 的结点 N。

①查找结点 N 的左兄弟结点，提取其句法标记、词性标记和词形信息。

②查找结点 N 的右兄弟结点，提取其句法标记、词性标记和词形信息。

③查找结点 N 的父亲结点，提取其句法标记、词性标记和词形信息。

(5)融合所提取的句法标记、词性标记和词形信息形成消歧特征。

对于含有歧义词“面”的汉语句子，消歧特征的提取过程如下所示：

S：这位摄影师说：作案人蒙着面。

$S_{\text{word\_seg}}$：这/r 位/q 摄影师/ng 说/vg：/wo 作案人/ng 蒙/vg 着/used 面/ng。/wj

T_S：S[SS[BNP[BMP[这/r 位/q]摄影师/ng]说/vg]：/wo SS[作案人/ng VO[VP[蒙/vg 着/used] 面/ng]]。/wj]

在这里，r是代词的标记，q代表量词，ng是名词标记，vg是动词，wo代表冒号，used是助动词标记，wj代表句号。对于汉语句子 S，其句法树的结构如图1所示。

在汉语中，“面”是一个歧义词汇。在《同义词词林》中，“面”共有9种不同的词义，表1列举了8种常用的词义类别及其汉语同义词。

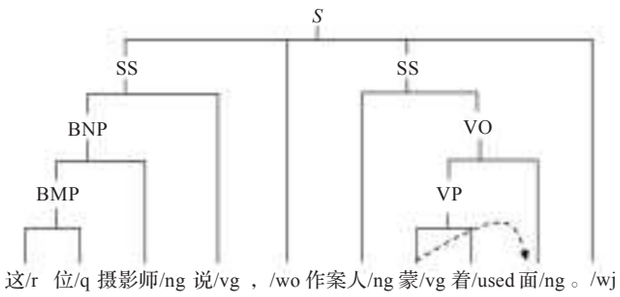

图1　汉语句子*S*的句法树

表1　“面”常用的8种词义类别及其同义词

序号	词义类	同义字
1	Bk02	头、脸、脑
2	Kb01	用脸对着、向着
3	Eb30	美丽、丑陋
4	Dd05	范围、方面、环节
5	Br01	粮食
6	Cb23	点、线、角、面、体
7	Br05	面试、杂粮
8	Bb02	粉末儿

句法树 T_S 是一棵4层结构的多叉树。在第一层中，结点 S 有4个孩子结点，包括：子句结点SS、句子停顿结点“，/wo”、子句结点SS和句子结束结点“。/wj”。在第二层中，子句结点SS有两个孩子结点，包括：基本名词短语结点BNP和词汇结点“说/vg”；子句结点SS由词汇结点“作案人/ng”和动宾短语结点“VO”组成。在第三层中，基本名词短语结点BNP由基本量词短语结点BMP和词汇结点“摄影师/ng”构成；动宾短语结点VO由动词短语结点VP和词汇结点“面/ng”组成。在第四层中，基本量词短语结点BMP由词汇结点“这/r”和词汇结点“位/q”构成。短语结点VP由词汇结点“蒙/vg”和词汇结点“着/usde”组成。此处，“面”为歧义词。根据其上下文语境，可以推断它的真实词义类别为Bk02，其汉语含义为“脸”。

如果按照常规的开设窗口的方法，以“面”为中心左右出现的词汇作为上下文候选特征词，那么“面”的左消歧特征为词汇结点“着/usde”，右消歧特征为句子结束结点“。/wj”。然而，这两个特征对于确定“面”的词义没有直接的作用。这种直线窗口所提供的消歧信息比较有限，而且容易造成数据稀疏问题。短语句法树提供了一个树状结构，可以指明歧义词汇的修饰成分。通过遍历短语句法树，可以发现：约束“面”语义的上下文语境仅仅限于动宾短语结点“VO[VP[蒙/vg 着/used] 面/ng]]”之内。进一步观察，可以看出：直接修饰“面”的语法成分有三个。第一个是“面”处于动宾短语结点VO之中，受句法标记VO的约束，表明“面”是动宾短语中的宾语成分，作名词使用，而句法标记VO是“面”的父结点；第二个是动词短语结点VP[蒙/vg 着/used]，歧义词“面”为其支配对象，而句法标记VP为“面”的左兄弟结点。结点VP包含词汇结点“蒙/vg”和词汇结点“着/used”。此处，“着/use”是一个助动词，没有任何实际的汉语意义。经过深入观察，还可以发现：词汇结点“蒙/vg”是它的核心结点。对于歧义词“面”而言，它的父结点VO，左兄弟结点VP和核心左兄弟结点“蒙/vg”直接决定了它的词义类别Bk02。所获取的消歧特征如下所示：

父特征：VO

左兄弟特征：VP(蒙/vg)

右兄弟特征：NULL

因此，句法树对于词义消歧来说能够提供一定的判别信息。

3　基于贝叶斯模型的词义消歧分类器

本文将贝叶斯分类模型应用于词义消歧过程。根据贝叶斯决策规则，正确词义在给定的上下文环境里出现的概率最大。对于歧义词汇 w，它有 m 种词义 $S_1, S_2, \cdots, S_m$。在上下文 *context* 中，它的正确词义 S 由公式(1)决定。

$$S=\arg\max_{i=1,2,\cdots,m} P(S_i|context)=\arg\max_{i=1,2,\cdots,m}\frac{P(S_i\cdot context)}{P(context)}=\arg\max_{i=1,2,\cdots,m}\frac{P(context|S_i)\cdot P(S_i)}{P(context)}\approx\arg\max_{i=1,2,\cdots,m} P(context|S_i)\cdot P(S_i) \tag{1}$$

其中，歧义词汇 w，其上下文环境 *context* 由 w 的父特征、左兄弟特征和右兄弟特征组成；$P(x)$ 是 x 出现的概率。在人工标注语料上，对参数 $P(context|S)$ 和 $P(S)$ 进行训练。使用优化后的分类器对歧义词汇 w 进行消歧，其过程如图2所示。

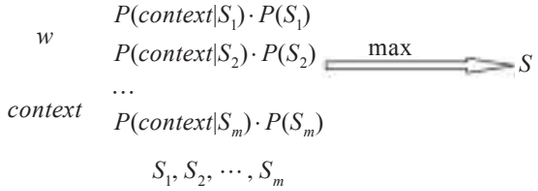

图2　歧义词 *w* 的词义决策过程

4　实验

为了衡量本文所提出方法的性能，收集了包含歧义词汇的80个汉语句子。两名人工标注者按照《同义词词林》手工标注歧义词汇的语义类，同时进行交叉校验。然后，对这80个汉语句子分别进行分词、词性标注和句法分析处理。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，其性能如表2所示。

表2　汉语分词工具、词性标注工具和句法分析工具的性能

%

	精确度	召回率	F1	准确率
汉语分词工具	85.2	81.4	83.3	—
汉语词性标注工具	—	—	—	96.61
汉语句法分析工具	78	79	78	—

针对每个歧义词汇，利用本文所提出的方法来抽取其上下文消歧特征。将这80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60个汉语句子；在测试数据集中，共包含20个汉语句子。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利用歧义词汇的左右邻接单词作为消歧特征，使用贝叶斯模型作为词义消歧分类器。实验2采用本文所提出的词义消歧分类器，即公式(1)。利用训练数据对两组实验中的分类器进行训练。使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可以看出实验2的分类准确率要高于实验1的。其原因是：在实验2中，利用了句法树来抽取歧义词汇的上下文，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

	正确分类数	错误分类数	准确率/%
实验1	11	9	55
实验2	13	7	65

5　结束语

本文提出了一种利用句法信息来指导汉语词义消歧的方法。从句法树中提取歧义词汇的上下文消歧特征，同时，使用贝叶斯模型来进行词义分类。实验结果表明：其准确率达到了65%。分类的准确率不是很高，其原因是：消歧特征包括歧义词汇的左兄弟结点、右兄弟结点和父亲结点的句法、词性以及词形信息。通过分析测试语料可以发现：在一些句子中，歧义词汇的左、右兄弟结点就是一个单词，这与常用的开设窗口的方法一样。在部分句子中，歧义词汇的左、右兄弟是短语结点，所抽出的核心左、右兄弟结点有些可以决定歧义词汇的语义，有些则不能。因此，其精度不高。下一步将认真分析结果，采取改进方案来提高性能。所使用的分词、词性标注和句法分析工具是在哈尔滨工业大学计算机科学与技术学院读博期间使用的，所用到的语料也是实验室开发标注的。由于分词规范、词性标注符号和语义标注类别与公开测试集还有不一致的地方，因此，从实验室的语料库中搜集了部分语料作为测试数据。下一步工作将对实验工具做改造，采用公开测试集进行评测。

参考文献：

[1] 朱靖波，李珩，张跃，等.基于对数模型的词义自动消歧[J].软件学报，2001，12(9)：1405-1412.

[2] 鲁松，白硕，黄雄，等.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.

[3] 刘风成，黄德根，姜鹏.基于AdaBoost.MH算法的汉语多义词消歧[J].中文信息学报，2006，20(3)：6-13.

[4] 张仰森，黄改娟，苏文杰.基于隐最大熵原理的汉语词义消歧方法[J].中文信息学报，2012，26(3)：72-78.

[5] 陈浩，何婷婷，姬东鸿.基于*k*-means聚类的无导词义消歧[J].中文信息学报，2005，19(4)：10-16.

[6] Brody S，Lapata M.Good neighbors make good senses：exploiting distributional similarity for unsupervised WSD[C]//Proceedings of the 22nd International Conference on Computational Linguistics，2008：65-72.

[7] 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.

[8] Faralli S，Navigli R.A new minimally-supervised framework for domain word sense disambiguation[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning，2012：1411-1422.

[9] Buscaldi D，Rosso P，Masulli F.Integrating conceptual density with WordNet domains and CALD glosses for noun sense disambiguation[C]//Proceedings of España for Natural Language Processing，2005：267-276.

[10] Mihalcea R.Unsupervised large-vocabulary word sense disambiguation with graph-based algorithms for sequence data labeling[C]//Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing，2005：411-418.

[11] 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.

[12] Navigli R，Ponzetto S P.BabelRelate! A joint multilingual approach to computing semantic relatedness[C]//Proceedings of the 26th AAAI Conference on Artificial Intelligence，2012：108-114.

[13] 刘鹏远，赵铁军.基于双语词汇Web间接关联的无指导译文消歧[J].软件学报，2010，21(4)：575-585.

[14] Navigli R，Ponzetto S P.Joining forces pays off：multilingual joint word sense disambiguation[C]//Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning，2012：1399-1410.

[15] 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.