·人工智能及识别技术· 文章编号：1000-3428(2017)09-0210-04 文献标志码：A 中图分类号：TP181

基于依存句法分析的多特征词义消歧

史兆鹏，邹徐熹，向润昭

（合肥工业大学 计算机与信息学院，合肥 230000）

摘 要： 词义消歧在机器翻译、信息检索、语音语义识别等方面具有重要作用。为提高消歧质量，细化特征粒度，提出一种多特征词义消歧方案。通过依存句法分析提取上下文中多义词及义项的词性、依存结构、依存词等特征，细化特征粒度，并根据多特征构造权值函数，选择权值最大的义项作为多义词的义项。实验结果表明，与单一特征词义消歧相比，采用依存句法分析的多特征词义消歧方案细化了特征粒度，提高了消歧准确率。

关键词： 词义消歧；依存句法；细化特征；多特征；权值

中文引用格式： 史兆鹏，邹徐熹，向润昭. 基于依存句法分析的多特征词义消歧[J]. 计算机工程，2017，43(9)：210-213.

英文引用格式： SHI Zhaopeng，ZOU Xuxi，XIANG Runzhao. Multi-feature Word Sense Disambiguation Based on Dependency Parsing Analysis[J]. Computer Engineering，2017，43(9)：210-213.

Multi-feature Word Sense Disambiguation Based on Dependency Parsing Analysis

SHI Zhaopeng，ZOU Xuxi，XIANG Runzhao

（School of Computer and Information，Hefei University of Technology，Hefei 230000，China）

【Abstract】 Word Sense Disambiguation（WSD）plays an important role in machine translation，information retrieval and speech semantic recognition. In order to improve the quality of disambiguation and refine the feature，a multi-feature granularity WSD scheme is proposed. The extraction of parts of speech，dependency structure and dependent words is used to detail feature grain by dependency parsing. The weight function is constructed according to the multiple features as the classifier，and the meaning with the largest weight is chosen as the sense of the polysemous word. Experimental results show that compared with single feature WSD，the multi-feature WSD scheme based on dependency parsing refines the feature and improves the accuracy of disambiguation.

【Key words】 Word Sense Disambiguation（WSD）；dependency parsing；detailed feature；mulit-feature；weight

DOI： 10.3969/j.issn.1000-3428.2017.09.037

0 概述

一词多义是歧义产生的根源。根据多义词所在上下文确定多义词的义项称为词义消歧（Word Sense Disa mbiguation，WSD）。词义消歧在机器翻译、文本分类、信息检索、语音识别、语义网络构建等方面都具有重要意义[1]。

文献[1]给出了词义消歧的任务：词语 W 具有 *N* 个义项，W 在特定上下文 C 中只有 S’为正确义项，词义消歧的任务就是确定上下文 C 中词语 W 的义项为 S’。

目前，词义消歧以机器学习中各种分类方法为代表，如最大熵值法[2]、决策树[3]、Naive-Bayes[4-5]、支持向量机[6]等。与手工分类方法相比，这些词义消歧方法效率上取得了明显提升，但由于必须通过标注语料学习，因此难以实现大规模的知识学习和词义消歧。

为了适应真实的应用场景，无导词义消歧开始引起关注。文献[7]利用《同义词词林》，在大规模语料库中获取同义词集中单义词的同现实词，按照同现实词的分辨能力对它们加权，构造分类器，实现一种代价最小的无导词义消歧。文献[8]把待消歧的多义词所在的上下文视为查询，把同多义词具有相同或相似语义词语的上下文视为文档，采用查询文档的方法，实现一种基于向量空间模型中义项词语的无导词义消歧。文献[9-11]也分别进行了无导词义消歧。

近年来，许多学者采用依存句法分析进行词义消歧。文献[12]通过依存句法分析，统计多义词依存元组特征，获得依存约束集合；根据 WordNet 获取词义代表词；计算词义代表词在约束集合中的依存适配度完成词义消歧。文献[13]通过对多义词上下

基金项目： 国家自然科学基金(61272540)。

作者简介： 史兆鹏(1990—)，男，硕士研究生，主研方向为自然语言处理、网络安全；邹徐熹、向润昭，硕士研究生。

收稿日期： 2016-07-24 **修回日期：** 2016-09-28 **E-mail：** 874257213@qq.com

文结构进行句法分析获取句法分析树,提取依存结构作为消歧特征,使用朴素贝叶斯模型作为分类器进行词义消歧。

分析词义消歧方法,发现目前词义消歧多以单特征提取、匹配为主,采用的特征有词性特征、依存结构特征、词汇共现特征等。特征粒度过粗,容易出现特征相同的情况,例如提取依存结构特征时如果多义词结构特征与 2 个或以上义项结构特征相同,那么多个义项难以区分。

以上述研究为基础,本文提出一种基于依存句法分析的多特征词义消歧方案,通过多特征细化特征粒度,减少特征相同情况,提高消歧准确率。

1　多特征词义消歧方案

1.1　多特征选择分析

在单特征的词义消歧方案中容易出现特征相同无法区分的情况。以依存结构作为特征进行词义消歧为例,在汉语中,“打”是一个多义词,包含多种义项。一种义项为“买”,一种义项为“编织”。假设对于多义词“打”查询上下文得到“打醋”,则提取到“打”的依存结构为 v + VOB + n;查询义项“买”上下文得到“买酒”,则提取到的依存结构为 v + VOB + n;查询义项“编织”上下文得到“编织毛衣”,则提取到的依存结构为 v + VOB + n;义项“买”“编织”提取到的依存结构都为 v + VOB + n,两义项提取到的依存结构与多义词“打”提取到的依存结构相同,则“打醋”中的“打”无法区分是义项“买”,还是义项“编织”。因此,仅通过依存结构特征进行词义消歧,难以确定多义词的真实义项。

考虑到词性特征对于多义词义项的确定具有指导作用:多义词在上下文中的词性应与多义词在上下文中表达的义项词性相同。因此,本文引入词性特征用于词义消歧。

词汇共现特征在一定程度上能够提高词义消歧准确率,例如“打”有“买”“编织”2 个义项;若“打”提取到“打醋”且“买”提取到“买醋”,则“打”的义项更倾向于“买”;若“打”提取到“打毛衣”且“编织”提取到“编织毛衣”,则“打”的义项更倾向于“编织”。然而,实际消歧中词汇并不一定共现,如在知识库有限的情况下,多义词“打”在上下文中提取到“打醋”,义项“买”提取到“打酒”,义项“编织”提取到“编织毛衣”。虽然“打醋”与义项“买”中的“买酒”,义项“编织”中的“编织毛衣”的依存词不同,但是“打醋”中“打”的依存词“醋”与“买酒”中“买”的依存词“酒”都是食用品,而“编织毛衣”中“编织”的依存词“毛衣”是衣物,非食用品;“醋”与“酒”的相似度要高于“醋”与“毛衣”的相似度。因此,可以采用依存词特征进行词义消歧,通过计算“醋”与“酒”、“醋”与“毛衣”的相似度,选择相似度更高的“醋”与“酒”所对应的义项“买”作为“打醋”中“打”的义项。

综合以上分析,本文选择词性特征,依存结构特征,依存词特征进行多特征词义消歧。

1.2　多特征获取

由于依存句法分析能够找到与多义词存在直接语义关系的信息,去除无关信息,因此本文从依存句法分析树抽取多义词的各个特征,依存句法分析采用哈工大依存句法分析工具[14]。首先对待消歧多义词所在上下文进行分词,词性标注,依存句法分析,构建对应的依存句法分析树。遍历依存句法分析树,确定待消歧多义词所在节点;以多义词所在节点为中心,向上和向下开设窗口,查找其父节点与子节点。抽取多义词节点的词性信息得到词性特征;抽取父节点、子节点及多义词节点的词性和关系类型信息,得到依存结构特征;抽取父节点、子节点词汇信息,得到依存词特征。

对于含有待消歧多义词“把握”的上下文,多义词“把握”的特征获取过程如下所示。

上下文:说实话,我连烧开水也没把握。

依存句法分析结果:|0 说实话 i--ADV--4|1,wp--WP--0|2 我 r--SBV--4|3 连 u--ADV--4|4 烧 v--HED---1|5 开水 n--VOB--4|6,也 d--ADV--7|7 没 v--COO--4|8 把握 n--VOB--7|9。wp--WP--4|。

其中,|0 说实话 i--ADV--4|表示一个节点(Root 表示根节点,为虚节点);0 表示节点 id,“说实话”为词汇;“i”为词性标注;“--ADV--4”中“4”表示当前节点的父节点 id,“ADV”表示当前节点和父节点的关系类型为状中关系;i 表示惯用语;wp 表示标点符号关系;r 表示代名词;u 表示辅助词;v 表示动词;n 表示名词;d 表示副词;ADV 表示状中关系;WP 表示标点符号关系;SBV 表示主谓关系;HED 表示核心关系;VOB 表示动宾关系;COO 表示并列关系。

由父节点向子节点引一条有向弧,标注关系类型,构建依存句法分析树如图 1 所示。

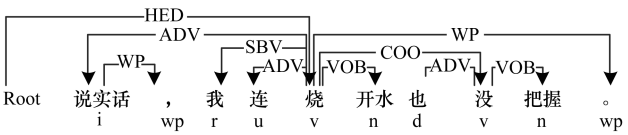

图 1　多义词“把握”依存句法分析树

查询依存句法分析树,可找到“把握”所在节点的父节点为词汇“没”所在的节点,子节点为空,抽取信息得到“把握”在该上下文中的词性特征为:“n”;依存结构特征为:v + VOB + n;依存词特征为:“没”。

多义词“把握”的义项特征获取过程如下:

根据 HowNet 知识库[15]“把握”有 4 个义项,相关义项信息为:

NO. =001644,DEF = 拿,G_C = v,E_C = ~方向盘,紧紧~

NO. =001646,DEF = 领会,G_C = v,E_C = ~事物的本质,~文件的精神实质

NO. =001648,DEF = 利用,G_C = v,E_C = ~机遇,~时间,~机会,~火候,~时机,很好地~,~好

NO. =001641,DEF = 情感,相信,G_C = n,E_C =有~,没~,有没有~,~不大,无~

其中,NO 表示义项标识;DEF 表示义项的含义;G_C 表示义项词性;E_C 表示义项实例。

以 NO. =001644 的义项为例,对 E_C 项中的短语结构进行依存句法分析(其他义项方法相同),分析结果如图 2 所示。

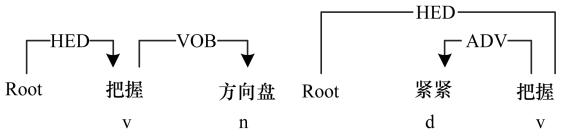

图 2 多义词“把握”的义项依存句法分析树

抽取依存句法分析树中的信息,得到 NO. =001644的义项词性特征为“v”,依存结构特征为 v + VOB + n,v + ADV + d,依存词特征为方向盘、紧紧。

1.3 词义消歧分类器

在给定的上下文中,多义词正确义项提取的多个特征应该同多义词提取的多个特征最相似,采用相似性权值计算,则正确义项的权值最大,本文的词义消歧规则如式(1)所示。

$$\max_{i=1,2,\cdots,n} Weight(w,s_i)\rightarrow s_i \tag{1}$$

其中,w 表示待消歧多义词;n 表示 w 具有的义项数;s_i 表示 w 的第 i 个义项;$Weight(w,s_i)$ 为多特征权值函数。规则描述为选择多义词 w 与义项 s_i 权值最大的义项作为多义词 w 的义项。

多特征权值计算函数如式(2)所示。

$$\begin{aligned} Weight(w,s_i) =& \frac{1}{n-m}Sim_{\text{pos}}(w,s_i) \\ &+\frac{1}{2}\left(1-\frac{1}{n-m}\right)Sim_{\text{ds}}(w,s_i) \\ &+\frac{1}{2}\left(1-\frac{1}{n-m}\right)Sim_{\text{dw}}(w,s_i) \end{aligned} \tag{2}$$

其中,$\frac{1}{n-m}Sim_{\text{pos}}(w,s_i)$ 表示词性特征权值,反映的是词性特征在词义消歧中的判别能力;n 表示多义词的义项数;m 表示与 w 词性不同的义项数。多义词 w 总共有 n 个义项,m 个义项的词性与给定上下文中 w 的词性不同,则 $n-m$ 个义项的词性与 w 的词性相同,因此,w 的正确义项必定为 $n-m$ 中的一个,所以,词性特征对于词义消歧的分辨能力为 $\frac{1}{n-m}$,n,m 根据待消歧多义词与给定上下文的不同而自动调节。$Sim_{\text{pos}}(w,s_i)$ 的表达式如式(3)所示。

$$Sim_{\text{pos}}(w,s_i)=\begin{cases}0,\text{词性特征不同}\\1,\text{词性特征相同}\end{cases} \tag{3}$$

$\frac{1}{2}\left(1-\frac{1}{n-m}\right)Sim_{\text{ds}}(w,s_i)$ 为依存结构特征权值,$Sim_{\text{ds}}(w,s_i)$ 的表达式如式(4)所示。

$$Sim_{\text{ds}}(w,s_i)=\begin{cases}0,\text{依存结构特征不同}\\1,\text{依存结构特征相同}\end{cases} \tag{4}$$

$\frac{1}{2}\left(1-\frac{1}{n-m}\right)Sim_{\text{dw}}(w,s_i)$ 为依存词特征权值,$Sim_{\text{dw}}(w,s_i)$ 的值为依存词相似度 $Sim(w_1,w_2)$ 的最大值,如式(5)所示。

$$Sim_{\text{dw}}(w,s_i)=\max_{w_1\in u_w,w_2\in u_{s_i}} Sim(w_1,w_2) \tag{5}$$

其中,u_w 表示多义词 w 在给定上下文中的依存词集;u_{s_i} 表示义项 s_i 的依存词集;w_1 表示给定上下文中的 w 的依存词;w_2 表示 s_i 的依存词;$Sim(w_1,w_2)$ 表示依存词相似度,$Sim(w_1,w_2)$ 计算采用文献[16]的计算方法,$Sim(w_1,w_2)$ 计算如式(6)~式(8)所示。

$$Sim(w_1,w_2)=\max_{i=1,2,\cdots,n,j=1,2,\cdots,m} Sim(s_{1i},s_{2j}) \tag{6}$$

$$Sim(s_{1i},s_{2j})=\sum_{k=1}^{4}\beta_k\prod_{l=1}^{k}Sim_l(p_1,p_2) \tag{7}$$

其中,$\beta_1+\beta_2+\beta_3+\beta_4=1$,$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$,$\beta_k$ 为可调因子。

$$Sim(p_1,p_2)=\frac{\alpha}{d+\alpha} \tag{8}$$

其中,s_{1i} 表示 w_1 的第 i 个义项;s_{2j} 表示 w_2 的第 j 个义项,通常 i,j 为 1。$Sim(s_{1i},s_{2j})$ 表示 w_1,w_2 义项的相似度,HowNet 知识库将每个义项映射为义原,每个义原为不能切分的语义最小单元,义项相似度通过义原相似度 $Sim(p_1,p_2)$ 计算获得。α 为可调因子,d 为 HowNet 知识库[15]的义原距离,查询 WHOLE. DAT 文件获得。根据文献[16]中参数设置建议及本文参数间大小关系分析,本文的因子设置为 $\alpha=1.6$,$\beta_1=0.5$,$\beta_2=0.2$,$\beta_3=0.17$,$\beta_4=0.13$。

2 实验结果与分析

2.1 数据集评价指标

为了衡量本文基于依存句法分析的多特征词义消歧对于特征粒度的细化作用,采用 SemEval-3 汉语词义消歧数据集进行实验,数据集包括 20 个词汇,每个词汇平均 4 个词义,共选用 379 个实例,词汇词义来自于 HowNet 知识库,具体数据分布如图 3 所示。

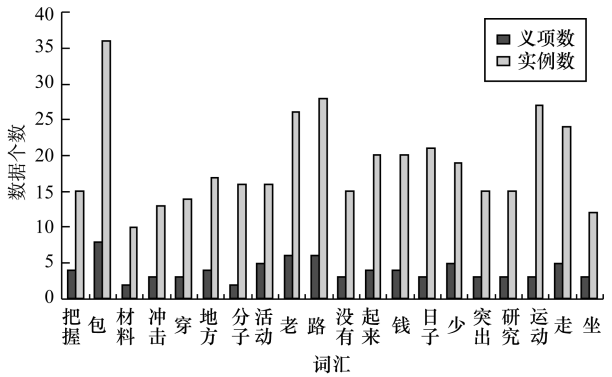

图 3　实验数据分布

采用准确率 P、召回率 R 以及 F 值作为评价指标。其中,P,R 计算如式(9)、式(10)所示。

$$P=\frac{\text{正确消歧的实例数}}{\text{完成消歧的实例数}} \tag{9}$$

$$R=\frac{\text{正确消歧的实例数}}{\text{待消歧的实例数}} \tag{10}$$

F 值为准确率和召回率的综合,F 值越大说明准确率和召回率相对较高。F 值计算如式(11)所示。

$$F=\frac{2\times P\times R}{P+R} \tag{11}$$

2.2　实验及分析

分别仅采用多义词的词性特征、依存结构特征、依存词特征进行词义消歧实验,与本文采用的基于依存句法分析的多特征词义消歧实验结果对比如表 1所示。

表 1　词义消歧结果对比　　%

特征选择	准确率	召回率	F 值
词性特征	13.47	13.47	13.47
依存结构特征	49.15	45.65	47.34
依存词特征	86.48	81.00	83.65
本文的多特征	89.11	82.06	85.44

实验对比发现,通过词性特征进行词义消歧准确率、召回率、F 值低,原因在于同给定上下文多义词词性相同的义项通常具有多个,仅通过词性特征难以区分;通过依存结构特征进行词义消歧准确率、召回率、F 值有所提升,但仍不高,分析原因在于依存结构特征包含了词性和关系类型,因此,相对词性特征有所提升,但依存结构特征在不同义项中仍有较高重复率,所以,导致消歧效果受限;依存词特征用于词义消歧,准确率、召回率、F 值相对较高是由于本文在词汇词汇共现基础上,进一步细化了特征粒度,不但能够处理词汇共现的情况而且能够更准确处理多义词依存词与义项依存词不同的情况;最后本文基于依存句法分析的多特征词义消歧准确率,召回率、F 值均为最高,相对依存词特征准确率提高了 2.63%,召回率提高 1.06%,F 值提高 1.79%,进一步细化了特征粒度,提高了消歧的准确率。分析本文基于依存句法分析的词义消歧准确率较高的原因归结为先对单特征的粒度细化,在此基础上通过多个特征相互作用对粒度细化,提高消歧准确率。

3　结束语

本文针对词义消歧特征粒度过粗容易出现特征相同的问题,采用依存句法分析方法,提出一种多特征词义消歧方案。实验结果表明,该方案能有效改善单一特征词义消歧粒度过粗、消歧质量不高的情况。下一步将对新衍生多义词实时消歧展开研究,探究新衍生多义词义项在词义消歧上的敏感性。

参考文献

[1] 卢志茂,刘　挺,李　生. 统计词义消歧的研究进展[J]. 电子学报,2006,34(2):143-153.
[2] 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.
[3] BLACK E. An Experiment in Computational Discrimination of English Word Senses[J]. IBM Journal of Research and Development,1988,32(2):185-194.
[4] MOONEY R J. Compative Experiments on Disambiguating Word Senses:An Illustration of the Role of Bias in Machine Learning[C]//Proceedings of Conference on Emporocal Methods in Natural Language Processing. Somerset, USA: Association for Computational Linguistics,1996:82-91.
[5] 丁江伟. 隐马尔可夫模型和贝叶斯模型词义消歧对比研究[C]//全国第七届计算语言学联合学术会议论文集. 哈尔滨:哈尔滨工业大学,2003:7-13.
[6] 张振景,李新福,田学东,等. 基于 SVM 的离合词词义消歧[J]. 计算机科学,2016,43(2):239-244.
[7] 李娟子,黄昌宁. 基于转换的无指导词义标注方法[J]. 清华大学学报(自然科学版),1999,39(7):117-121.
[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] 秦　颖. 汉语词和短语的歧义消解研究[D]. 北京:北京邮电大学,2008.
[10] 王瑞琴,孔繁胜. 无监督词义消歧研究[J]. 软件学报,2009,20(8):2138-2152.
[11] 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.
[12] 鹿文鹏,黄河燕. 基于依存适配度的知识自动获取词义消歧方法[J]. 软件学报,2013,24(10):2300-2311.
[13] 张春祥,栾　博,高学瑶,等. 基于句法分析的汉语词义消歧[J]. 计算机应用研究,2014,31(1):40-42.
[14] 哈工大社会计算与信息检索研究中心. 语言技术平台云:LTP[EB/OL]. [2014-11-30]. http://www.ltp-cloud.com/.
[15] 董　强,董振东. 知网:HowNet Knowledge Database[EB/OL]. [2013-01-29]. http://www.keenage.com/.
[16] 刘　群,李素建. 基于《知网》的词汇语义相似度计算[C]//第三届汉语词汇语义学研讨会论文集. 台北,中国:[出版者不详],2002:59-76.

编辑　刘　冰