软件学报 ISSN 1000-9825, CODEN RUXUEW
*Journal of Software*,2013,24(10):2300−2311 [doi: 10.3724/SP.J.1001.2013.04373]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel/Fax: +86-10-62562563

# 基于依存适配度的知识自动获取词义消歧方法*

鹿文鹏[1,2], 黄河燕[1]

[1](北京理工大学 计算机学院 北京市海量语言信息处理与云计算应用工程技术研究中心,北京 100081)

[2](山东轻工业学院 理学院,山东 济南 250353)

通讯作者: 鹿文鹏, E-mail: luwpeng@bit.edu.cn, http://cs.bit.edu.cn

**摘 要**: 针对困扰词义消歧技术发展的知识匮乏问题,提出一种基于依存适配度的知识自动获取词义消歧方法.该方法充分利用依存句法分析技术的优势,首先对大规模语料进行依存句法分析,统计其中的依存元组信息构建依存知识库;然后对歧义词所在的句子进行依存句法分析,获得歧义词的依存约束集合;并根据 WordNet 获得歧义词各个词义的各类词义代表词;最后,根据依存知识库,综合考虑词义代表词在依存约束集合中的依存适配度,选择正确的词义.该方法在 SemEval 2007 的 Task#7 粗粒度词义消歧任务上取得了 74.53%的消歧正确率;在不使用任何人工标注语料的无监督和基于知识库的同类方法中,取得了最佳的消歧效果.

**关键词**: 词义消歧;依存句法分析;知识获取;依存适配度

**中图法分类号**: TP391 **文献标识码**: A

中文引用格式: 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.软件学报,2013,24(10):2300−2311. http://www.jos.org.cn/1000-9825/4373.htm

英文引用格式: Lu WP, Huang HY. Word sense disambiguation based on dependency fitness with automatic knowledge acquisition. Ruan Jian Xue Bao/Journal of Software, 2013,24(10):2300−2311 (in Chinese). http://www.jos.org.cn/1000-9825/4373.htm

## Word Sense Disambiguation Based on Dependency Fitness with Automatic Knowledge Acquisition

LU Wen-Peng[1,2], HUANG He-Yan[1]

[1](Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Application, School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081, China)

[2](School of Science, Shandong Polytechnic University, Ji'nan 250353, China)

Corresponding author: LU Wen-Peng, E-mail: luwpeng@bit.edu.cn, http://cs.bit.edu.cn

**Abstract**: A word sense disambiguation (WSD) method based on dependency fitness is proposed to solve the problem of knowledge acquisition bottleneck in the development of WSD techniques. The method achieves automatic knowledge acquisition in WSD by taking full advantage of dependency parsing. First, a large-scale corpus is parsed to obtain dependency cells whose statistics information is utilized to build a dependency knowledge base (DKB); then, the ambiguous sentence is parsed to obtain the dependency constraint set (DCS) of ambiguous words. For each sense of ambiguous word, sense representative words (SRW) are obtained through WordNet. Finally, based on DKB, dependency fitness of all kinds of SRW on DCS is computed to judge the right sense. Evaluation is performed on coarse-grained English all-words task dataset of SemEval 2007. Compared with unsupervised and knowledge-based methods which don't utilize any sense-annotated corpus, the proposed method yields state-of-the-art performance with $F_1$-measure of 74.53%.

**Key words**: word sense disambiguation; dependency parsing; knowledge acquisition; dependency fitness

一词多义的现象在自然语言中十分普遍.词义消歧即指根据多义词所处的上下文环境来确定其词义,它是

* 基金项目: 国家自然科学基金(61132009); 国家重点基础研究发展计划(973)(2013CB329303)
收稿时间: 2012-06-06; 修改时间: 2012-07-24; 定稿时间: 2013-01-07

自然语言处理的一项重要的基础环节,对机器翻译、信息检索、文本分类、自动文摘等应用都具有直接的影响[1].

词义消歧包括有监督、无监督和基于知识库的方法[2].有监督的词义消歧方法消歧正确率高,但需要人工标注语料.由于难以获得足够大的标注语料,导致其无法应用于大规模词义消歧任务.无监督词义消歧方法不需要任何标注语料,甚至不需要任何词典资源.它主要通过对歧义词的上下文词语进行聚类而对词义进行分类,实质上是一种词义辨析(word sense discrimination)方法.基于知识库的词义消歧方法根据歧义词的上下文环境,利用知识库(如机读词典、本体库、搭配库等)来判断歧义词的词义.它的正确率通常不及有监督的方法,但其可使用丰富的各类知识库,具有较高的消歧覆盖率,能够满足大规模词义消歧任务的需求.在特定任务或特定领域内,已有相关研究证明,基于知识库的方法的效果甚至能够超过有监督的方法[3,4].鉴于基于知识库的方法是唯一可真正应用于大规模词义消歧任务的方法及其在 SemEval 评测中表现出的良好效果,该方法逐渐受到研究者的重视[3−7].

基于知识库的词义消歧方法的效果优劣,严重依赖于其所拥有的知识的规模与质量.知识获取瓶颈(knowledge acquisition bottleneck)是制约其发展的关键因素[2,7],亟待解决.如何有效地获取消歧知识,是词义消歧领域的研究热点之一[8−13].

对于基于知识库的词义消歧方法,其消歧知识的获取有两条路线:一是利用统计学习的方法,自动地从语料库中挖掘消歧知识,如词语共现信息、语言模型等[7,9,10,13−15];二是采用人工或半自动的方法构建知识库或对已有知识本体进行扩充[8,16].对于路线 1,不论是词语共现信息还是语言模型,均未考虑词语之间的句法、语义关系,仅仅根据在一定范围内词语共现或相邻来获取消歧知识,难免会受到一些近距离噪声词的干扰,影响获取的质量;对于路线 2,借助人工处理,该路线可获得少量高质量的消歧知识,但是词义消歧所需要的知识规模巨大,依靠该路线来完全获取这些知识并不现实.

为了适应基于知识库的词义消歧的需要,针对现有方法的不足,本文提出一种基于依存适配度的知识自动获取词义消歧方法.本文的主要贡献体现在以下几个方面:

- 提出一种基于依存句法分析的消歧知识自动获取方法.该方法通过依存句法分析获得语料中词语之间的依存关系;利用全部依存元组的统计信息构建依存知识库,作为消歧知识.该方法借助于依存句法分析技术,可以准确地寻找词语之间的依存关系,有效地避免噪声词的干扰,保证了知识获取的质量;
- 提出一种基于依存适配度的词义消歧方法.该方法以自动获取的依存知识作为消歧依据,综合考虑多种语义关系,计算歧义词各个词义的词义代表词(同义词、近义词、反义词、上位词、中心词)在上下文环境(依存约束集合)中的依存适配度,推测歧义词的词义;
- 本文首次提出以依存句法分析技术为主线,自动获取消歧知识;并以此为依据,通过依存适配度进行词义消歧的方法.该方法的消歧路线简明、有效,具有较高的消歧正确率,可适用于大规模词义消歧任务.

本文第 1 节介绍词义消歧的相关工作.第 2 节详细说明所提出的基于依存适配度的知识自动获取词义消歧方法.第 3 节给出实验及分析.最后对全文进行总结.

## 1 相关工作

现有的词义消歧方法可划分为有监督、无监督和基于知识库的方法,国内的卢志茂[1]、王瑞琴[17]、吴云芳[18]以及国外的 Navigli[2]、Agirre[6]均曾从不同时期、不同角度进行过相关的综述,各类方法的特点在本文引言部分已简要提及,这里不再赘述.本节着重从消歧模型和消歧知识的获取来说明基于知识库的词义消歧的相关工作.

基于知识库的词义消歧方法主要包含基于词义定义重合、选择限制、结构化关系这 3 种消歧模型[2].结构化关系模型是目前的研究热点,作为其典型代表,图模型因可深入挖掘多种结构关系,具有良好的消歧效果,而倍受关注.Navigli 提出基于结构化语义互连(structural semantic interconnections,简称 SSI)的词义消歧方法[5],它集成多种不同的语言知识(包括 WordNet[19]、WordNet Domain[16]、标注语料库及搭配词典等)作为消歧依据;

构建上下文无关文法来描述全部的语义关联模式;而后以迭代的方式,每次从词义未决词集中选择至少与已决词义概念存在一个语义关联的词来消歧,直至再无可利用的语义关联.在 SemEval 2007 评测中,SSI 方法的消歧效果超过了有监督的词义消歧方法,取得了最佳评测成绩[4].但 SSI 方法在知识获取及参数确定时需人工对齐处理,并且需要使用标注语料库,这使得部分研究者对其并不完全认同[10,20,21].Agirre 提出基于 Personalized PageRank 的词义消歧方法[3,20],它以 WordNet 和 eXtended WordNet[22]作为知识来源;将 WordNet 中的概念作为节点,将 WordNet 和 eXtended WordNet 中的关系作为边,构建知识图;将歧义词的上下文中所包含的实词的词义与知识图中的节点关联在一起;利用 PageRank 算法计算歧义词的各个词义概念的权重,从而确定它们的词义.杨陟卓提出基于词语距离的网络图词义消歧方法[23],该方法以中文词语搭配库、HowNet[24]作为知识来源;将歧义句中的词语及其词义作为节点,将词语、词义之间的语义、共现关系作为边,构建网络图;利用 PageRank 算法计算歧义句中词语的各个候选词义的权重.以上这些工作均侧重于对图模型进行改进而深入挖掘已有知识的内在结构化关联信息,侧重于如何构建更好的模型来更充分地利用已有知识,而忽视了制约词义消歧发展的关键——知识获取瓶颈问题.

词义消歧被认为是一个 AI 完全(AI-complete)问题[2],解决这一问题不仅需要有效的消歧模型,更依赖于大量的知识.鉴于词义消歧所需要的知识规模巨大,人工获取、完善它们并不现实.研究者尝试自动挖掘语料中所蕴含的信息作为消歧知识.Google 统计并发布了 Web-scale N-gram Corpus[25];Bergsma 将其作为消歧知识,根据词义候选词与上下文词语构成的 N-gram 模式的频次来选择正确的词义[9].N-gram 语言模型在统计时只是考虑了词语左右的邻接关系,而未考虑词语之间是否存在实质的句法、语义关系,容易受到噪声词的影响.刘鹏远提出基于双语词汇 Web 间接关联的译文消歧方法,该方法将双语词汇的 Web 搜索计数作为消歧知识,计算双语词汇的 Web 间接关联度以推测正确的译文[15].该方法根据双语词汇是否在同一个网页中共现来判断两者是否存在关联;但是 Web 页面复杂多样、内容参差不齐,这会影响关联度计算的可靠性.Chen 提出基于 TreeMatch 的词义消歧方法,该方法首先根据依存关系统计 Web 语料中词语的共现频率作为消歧知识,而后利用 TreeMatch 算法对比歧义词所在的句子与各词义的注释定义(gloss)的依存句法树的匹配程度,从而消除歧义[10].该方法在统计词语的共现频率时未区分依存关系的类型,将词语在不同类型的依存元组中的共现视为同一情形来统计.这并不妥当,混淆了依存元组类型的差别.这些已有工作在挖掘消歧知识时,粒度均尚不够细致,未能充分利用词语之间的句法、语义关系,有进一步提升的空间.

## 2 基于依存适配度的知识自动获取词义消歧方法

### 2.1 消歧思路

为了便于描述,本文首先给出以下概念说明.

**定义 1(依存元组)**. 如果词语 $w_1$ 与 $w_2$ 之间存在依存关系 $r$,且 $w_1$ 为支配词(governor),$w_2$ 为从属词(dependent),则称三者共同构成依存元组,记为 $r(w_1,w_2)$.其中,$w_1$ 与 $w_2$ 互称为对方的依存词.

**定义 2(依存适配)**. 如果词语 $w_1$,$w_2$ 在依存元组 $r(w_1,w_2)$中共现,则称两者满足依存关系 $r$ 下的依存适配.

**定义 3(依存约束)**. 如果词语 $w$ 在依存元组 $r(w,x)$或 $r(x,w)$中出现,则称 $r(*,x)$或 $r(x,*)$为 $w$ 的依存约束.

**定义 4(词义代表词)**. 与词义 $s$ 具有某种特定的语义关系,可以完全或部分表示 $s$ 的含义、适配于 $s$ 的上下文的词语,称为 $s$ 的词义代表词,如同义词、近义词、上位词等.

词义消歧的一个基本原则为,“观其伴、知其义”.歧义词的词义可根据其所处的上下文来确定.歧义词的依存约束集合可视为其上下文.既然歧义词是满足这些依存约束集合的,那么歧义词的正确词义的词义代表词也应该满足它们,正确词义应与它们具有最高的依存适配度.这是本文消歧方法的主要出发点.基于此,本文做出如下假设:

**假设 1**. 歧义词的词义可根据其词义代表词在依存约束集合中的依存适配度来确定.

假设 1 是本文提出的消歧方法的主要依据.根据假设 1,我们可以把词义消歧问题转化为依存适配度计算问题.本文所提出的基于依存适配度的知识自动获取词义消歧方法以依存句法分析作为主线,自动构建依存知识

库,并获取歧义词的依存约束集合;综合考虑歧义词的各类词义代表词在依存约束集合中的依存适配度来判断正确的词义,其整体流程框架如图 1 所示.

(1) 首先,通过对语料库中的文本进行依存句法分析,统计依存元组信息,构建依存知识库;
(2) 对歧义词所在的句子进行依存句法分析,得到其依存元组集合;
(3) 对于某个特定歧义词,从句子的依存元组集合中筛选出其依存约束集合;
(4) 对于当前歧义词,根据语义本体,获取其各个词义的词义代表词;
(5) 根据依存知识库,依次计算各词义的词义代表词在依存约束集合中的依存适配度;
(6) 最后,根据各个词义的依存适配度,选择正确的词义输出.

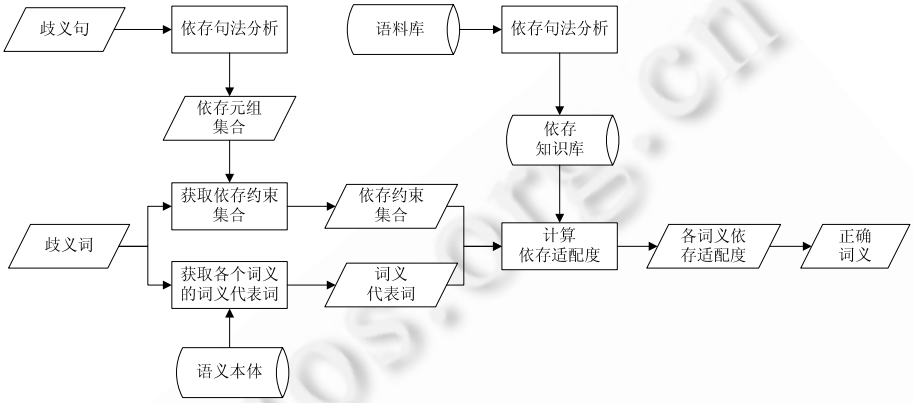


Fig.1 Framework of WSD based on dependency fitness with automatic knowledge acquisition

图 1 基于依存适配度的知识自动获取词义消歧方法的整体框架

### 2.2 依存知识自动获取方法

如本文第 1 节所述,大规模语料的统计数据(如词共现信息、语言模型等)可以作为消歧知识[9,10,15].已有的工作普遍存在知识获取粒度较为粗糙的问题.依存句法分析技术可以准确地获取语料库中词语之间的依存关系[26],为解决这一问题提供了新的思路.本文利用依存元组的统计信息构建依存知识库,具体方法如图 2 所示.本节对部分细节进行详细说明.

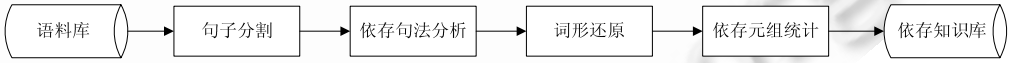


Fig.2 Method for dependency knowledge acquisition

图 2 依存知识自动获取方法

#### 2.2.1 语料库的选择

语料库是依存知识的来源,主要有两种选择:Web 语料和人工语料库[10].Web 语料具有获取方便、动态更新的优点,但其良莠不齐,会影响知识获取的质量;而人工语料库由专业人员编辑完成,质量可靠.为了保证依存知识的质量,本文选择使用人工语料库——Reuter Corpus[27].该语料库包含了路透社的 806 791 篇各类新闻报道文档.

#### 2.2.2 依存句法分析工具的选择

相关研究中,常用的句法分析工具有 Minipar[28],Stanford Parser[29],Berkeley Parser[30]等.其中,Stanford Parser 提供了 5 种类型的依存元组输出形式,可以设定是否允许对依存关系进行折叠和传递;提供了多种经过训练的 PCFG 分析模型,这为本文工作的开展提供了极大的便利.因此,本文采用 Stanford Parser 作为依存句法分析工具.Stanford Parser 可输出 52 种不同类型的依存关系,其中的部分常见依存关系见表 1.

**Table 1** Dependency relations

**表 1** 依存关系

| 依存关系名 | 依存关系含义 | 示例 | |
|---|---|---|---|
| | | 例句 | 依存元组 |
| nsubj | 名词性主语 | The baby is cute. | nsubj(cute,baby) |
| dobj | 直接宾语 | They win the lottery. | dobj(win,lottery) |
| iobj | 间接宾语 | Teachers gave students a raise. | iobj(gave,students) |
| nsubjpass | 被动名词性主语 | Dole was defeated by Clinton. | nsubjpass(defeated,Dole) |
| xsubj | 从句控制主语 | Tom likes to eat fish. | xsubj(eat,Tom) |
| nn | 名词性复合修饰语 | Oil price futures. | nn(oil,price) |
| conj | 连词 | They either ski or snowboard. | conj_or(ski,snowboard) |
| amod | 形容词 | Sam eats red meat. | amod(meat,red) |
| det | 限定词 | The man is here. | det(man,the) |
| aux | 助动词 | He should leave. | aux(leave,should) |
| auxpass | 被动助动词 | Kennedy was/got killed. | auxpass(killed,was/got) |
| xcomp | 补语从句 | He says that you like to swim. | xcomp(like,swim) |

2.2.3 依存元组的统计

受 Chen 的方法的启发[10],本文统计出现在语料库中的各类依存元组的数量信息,构建依存知识库,作为下一步进行依存适配度计算的依据.需要注意的是,部分存在明显错误的依存元组应予以剔除.Chen 的方法在统计依存元组时未对不同的依存关系类型进行区分,也未区分词语的词性信息,这并不恰当.本文处理方法与之不同,举例说明:

例句 1. “The coach is training football players.”.

例句 2. “The train and coach are chosen to transport the soldiers.”.

单词 coach 作为名词时,主要有两个词义:① 教练;② 长途客车.例句 1 与例句 2 分别与之对应.对两个句子进行依存句法分析和词形还原之后,可以得到如下依存元组集合:

例句 1. det(coach,the), nsubj(train,coach), aux(train,be), nn(player,football), dobj(train,player).

例句 2. det(train,the), nsubjpass(choose,train), xsubj(transport,train), conj_and(train,coach), nsubjpass(choose, coach), xsubj(transport,coach), auxpass(choose,be), aux(transport,to), xcomp(choose,transport), det(soldier,the), dobj(transport,soldier).

对于例句 1 中的 nsubj(train,coach)与例句 2 中的 conj_and(train,coach),如果不对依存关系的类型进行区分,则尽管两者的 train 与 coach 的词义各不相同,但仍会被合并为同一个元组(train,coach)而统计.这显然并不合理.此外,依存词的词性对于歧义词的词义也有指示作用.coach 与动词 train 共现时,词义倾向于“① 教练”;而与名词 train 共现时,词义倾向于“② 长途客车”.

因此,本文在统计依存元组时不仅保留了依存元组的类型信息,也为其添加了词性标记信息.定义 1(依存元组)可进一步表示为 $r(w_1,p_1,w_2,p_2)$,其中,$p_1,p_2$ 分别为 $w_1,w_2$ 的词性标记.为了便于表述,本文以下部分仍采用 $r(w_1,w_2)$的形式表示依存元组,但其中的 $w_1,w_2$ 均包含词形和词性信息.

### 2.3 词义代表词获取方法

本文将 WordNet 作为语义本体,获取词义代表词.WordNet 以同义词集(synset)来表示词义概念(concept);对名词、动词、形容词、副词分别进行组织;同义词集之间通过多种语义关系进行互连,如反义关系(antonym)、上位关系(hypernym)等.

词义概念的同义词、近义词、上位词均可表示当前词义的全部或部分语义.反义词虽然词义相互对立,但它们属于同一意义范畴,可以用来描述或支配相同的对象(如“short-long”,“quickly-slowly”),通常能够进行互换,可适用于彼此的上下文环境.WordNet 中的词义定义注释(gloss)通常为一个短句,如名词“apple”的词义 1 的注释为:fruit with red or yellow or green skin and sweet to tart crisp whitish flesh.对该注释进行依存句法分析,可获得该注释的中心词,即依存树的根为 fruit.可以看出,注释的中心词也能够代表词义的部分含义.

本文将同义词、近义词、反义词、上位词、中心词均作为词义代表词.利用 WordNet 的同义词集、近义关系(similar to)、反义关系(antonym)、上位关系(hypernym)可以方便地获取其前 4 类词义代表词.对于中心词,可以通过对注释定义进行依存句法分析,根据依存句法树而获得.需要注意的是,词义代表词不包含歧义词自身.

### 2.4 依存适配度计算方法

依存适配度考察歧义词的词义代表词或词义在依存约束元组中的适配程度,可以利用统计共现模型计算.常用的统计共现模型为点式互信息(point-wise mutual information,简称 PMI)、DICE 系数(DICE)、Phi 平方系数($\phi^2$)和对数似然比(log likelihood ratio,简称 LLR)[15].其中,以 PMI 的使用最为普遍,本文将其作为计算依据.

#### 2.4.1 词语的依存适配度计算

假定有依存元组 $r(w_1,w_2)$,为了计算词语 $w_1,w_2$ 在依存关系 $r$ 下的依存适配度,定义以下参数:

- $a=freq^r(w_1,w_2)$:依存关系为 $r$,且支配词为 $w_1$、从属词为 $w_2$ 的依存元组的总数;
- $b=freq^r(w_1,*)$:依存关系为 $r$,且支配词为 $w_1$ 的依存元组的总数;
- $c=freq^r(*,w_2)$:依存关系为 $r$,且从属词为 $w_2$ 的依存元组的总数;
- $N^r$:依存关系为 $r$ 的所有依存元组的总数.

词语 $w_1,w_2$ 在依存关系 $r$ 下的点式互信息 PMI 可由公式(1)计算[15]:

$$PMI^r(w_1,w_2)=\log\frac{N^r\times freq^r(w_1,w_2)}{freq^r(w_1,*)\times freq^r(*,w_2)}=\log\frac{N^r\times a}{b\times c} \tag{1}$$

本文将点式互信息作为词语 $w_1,w_2$ 在依存关系 $r$ 下的依存适配度,可由公式(2)计算:

$$fitness(w_1,w_2,r)=PMI^r(w_1,w_2)=\log\frac{N^r\times a}{b\times c} \tag{2}$$

公式(2)中,当 $a$=0 时,令 $fitness(w_1,w_2,r)$=0.

#### 2.4.2 词义的依存适配度计算

歧义词的某个词义在依存约束集合中的依存适配度,需要综合考虑其各类词义代表词在各个依存约束元组中的依存适配度,可根据公式(3)计算:

$$fitness(s_i,R)=\sum_{C_k\in C_{s_i}}\alpha_k\times\max_{w_{k_t}\in C_k}\sum_{r_j\in R}fitness(w_{k_t},r_j) \tag{3}$$

其中,

- $s_i\in sense(w)$,代表歧义词 $w$ 的某一个词义;
- $R$ 为歧义词 $w$ 的全部依存约束元组的集合;
- $C_{s_i}=C_{syn}\cup C_{simi}\cup C_{anto}\cup C_{hype}\cup C_{cent}$,代表词义 $s_i$ 的各类词义代表词的集合,包括同义词集 $C_{syn}$、近义词集 $C_{simi}$、反义词集 $C_{anto}$、上位词集 $C_{hype}$ 及中心词集 $C_{cent}$;
- $C_k\subset C_{s_i}$,代表某一特定类别的词义代表词的集合;
- $\alpha_k$ 代表第 $k$ 类词义代表词 $C_k$ 的权重系数;
- $r_j\in R$,代表某个依存约束元组;
- $w_{k_t}\in C_k$,代表 $C_k$ 中的某个词义代表词.

根据歧义词在依存约束元组 $r_j$ 中的位置,公式(3)中的 $fitness(w_{k_t},r_j)$ 可转化为 $fitness(w_{k_t},w',r_j)$ 或 $fitness(w',w_{k_t},r_j)$,由公式(2)计算.其中,$w'$为歧义词 $w$ 在依存约束元组 $r_j$ 中的依存词.

### 2.5 消歧决策算法

根据假设 1,歧义词的正确义项由 $s=\arg\max_{s_i} fitness(s_i,R)$ 来确定.消歧决策过程见算法 1.

**算法 1**. 消歧决策算法.

输入:歧义词 $w$、歧义词所在的句子 $Sen$、依存知识库 $DKB$、权重系数数组$\alpha$;

输出:歧义词的正确词义 $s$.

Step 1. 初始化当前句子的依存元组集合 $R_{sen}$=∅,歧义词的依存约束元组集合 $R_w$=∅;

Step 2. 对句子 *Sen* 作依存句法分析,获取其中全部依存元组,保存至 $R_{sen}$;

Step 3. 由依存元组集合 $R_{sen}$ 提取包含歧义词 *w*,且 *w* 的依存词为实词(名词、动词、形容词、副词)的元组,保存至歧义词的依存约束集合 $R_w$;

Step 4. for $s_i \in sense(w)$

置词义 $s_i$ 的综合适配度 $fitness_{(s_i,R_w)}$ 为 0;

根据 WordNet,获取词义 $s_i$ 的各类词义代表词 $C_{syn}$,$C_{simi}$,$C_{anto}$,$C_{hype}$,$C_{cent}$,剔除歧义词 *w* 自身及重复词后,统一加入词义代表词集 $C_{s_i}$ 中;

for $C_k \subset C_{s_i}$

置第 *k* 类词义代表词 $C_k$ 的综合适配度 $fitness_{(C_k,R_w)}$ 为 0;

for $w_{k_t} \in C_k$

置 $C_k$ 的第 *t* 个词义代表词的综合适配度 $fitness_{(w_{k_t},R_w)}$ 为 0;

for $r_j \in R_w$

if 歧义词 *w* 是依存约束元组 $r_j$ 的支配词(设 *w*′为其依存词)

根据 *DKB*,由公式(2)计算 $fitness(w_{k_t},w',r_j)$, 保存至 $fitness_{(w_{k_t},r_j)}$;

else

根据 *DKB*,由公式(2)计算 $fitness(w',w_{k_t},r_j)$, 保存到 $fitness_{(w_{k_t},r_j)}$;

将 $fitness_{(w_{k_t},r_j)}$ 累加至 $fitness_{(w_{k_t},R_w)}$;

end

end

对于 $C_k$ 中的各个词义代表词,取最大的 $fitness_{(w_{k_t},R_w)}$ 作为当前第 *k* 类词义代表词的综合适配度,保存至 $fitness_{(C_k,R_w)}$;

end

按照权重系数数组 $\alpha$,将 $fitness_{(C_k,R_w)}$ 加权累加至 $fitness_{(s_i,R_w)}$;

end

Step 5. 对于各个候选词义,取 $s=\arg\max_{s_i} fitness(s_i,R)$,将 *s* 作为正确义项返回(如果多个词义具有相同的 $fitness_{(s_i,R_w)}$,则根据 WordNet 选择词频较高的词义).

## 3 实 验

### 3.1 数据集和评价指标

本文利用 SemEval 2007 的 Task#7 粗粒度词义消歧任务(coarse-grained English all-words task)来验证所提出方法的效果.该测试集包含来自不同领域的 5 篇文档,共包含 2 269 个消歧对象,平均词义数为 3.06,其词义内部标注一致率为 93.80%[4].

本文采用检验词义消歧效果的常用标准:覆盖率、正确率、召回率、$F_1$ 测度[2],对实验结果进行评价,具体定义如下:

设 *A* 表示待消歧的实例的个数,*B* 表示消歧系统给出词义标记的实例的个数,*M* 表示消歧系统给出正确词义标记的实例的个数,则覆盖率(coverage)、正确率(precision)、召回率(recall)、$F_1$ 测度($F_1$-measure)的计算公式分别为

$$Coverage=\frac{B}{A},\ Precision=\frac{M}{B},\ Recall=\frac{M}{A},\ F_1=\frac{2\times Precision\times Recall}{Precision+Recall} \tag{4}$$

### 3.2 实验参数设定

本文使用 Matlab 遗传算法工具箱来确定公式(3)中各类词义代表词的权重系数$\alpha_k$[31].遗传算法是一种借鉴生物界自然选择和遗传机制,模拟染色体进化的启发式随机搜索算法.它从随机初始群体开始在染色体上进行种群的多代进化,获得最适应环境的染色体作为最优解.遗传算法工具箱的优化目标为使得适应度函数最小化,因此,本文将 $f(\alpha)=1-Precision$ 作为适应度函数;将 5 个词义代表词权重系数$\{\alpha_1,\alpha_2,\alpha_3,\alpha_4,\alpha_5\}$作为适应度函数的输入变量,设定各变量的下限值为 0;其他参数均使用遗传算法工具箱的缺省设置.

对于同义词集、近义词集、反义词集、上位词集、中心词集,分别得到以下优化权重:0.31,0.77,2.02,0.24,0.28.其中,反义词的权重最高,近义词次之,其他 3 类词权重较为接近.当一个歧义词同时存在多类词义代表词时,反义词和近义词将对词义的区分起到更大的作用.

### 3.3 实验结果

为了评价本文所提出的方法的效果,本文与其他 5 种方法进行了对比,见表 2.

**Table 2** Comparison of WSD effectiveness (%)

**表 2** 消歧效果对比(%)

| 方法 | 覆盖率(C) | 正确率(P) | 召回率(R) | $F_1$ 值 |
|---|---|---|---|---|
| $BL_{MFS}$ | 100.0 | 78.89 | 78.89 | 78.89 |
| **DepFit** | **100.0** | **74.53** | **74.53** | **74.53** |
| TreeMatch | 100.0 | 73.65 | 73.65 | 73.65 |
| TKB-UO | 100.0 | 70.21 | 70.21 | 70.21 |
| SUSSZ-FR | 72.8 | 71.73 | 52.23 | 60.44 |
| UofL | 92.7 | 52.59 | 48.74 | 50.60 |

对比的系统包括 SemEval2007 中无监督和基于知识库的方法的前三名(TKB-UO,SUSSZ-FR,UofL)、TreeMatch 方法及 $BL_{MFS}$,简要介绍如下[4]:

(1) $BL_{MFS}$:该方法根据 WordNet 词频选择最常用的词义,常用作词义消歧效果比较的基准方法.因在语言中存在很强的词义选择偏好性,无监督及基于知识库的词义消歧方法的消歧效果通常难以超越该基准方法[2,32];

(2) DepFit:这是本文所提出的基于依存适配度的知识自动获取的词义消歧方法;

(3) TreeMatch:该方法为 Chen 提出的基于依存树匹配的词义消歧方法,通过比较歧义句的依存句法树与词义注释的依存句法树的匹配程度而消歧[10];

(4) TKB-UO:该方法是一种基于聚类的无监督方法,通过多次迭代聚类,判断各个歧义词的词义;

(5) SUSSZ-FR:该方法为基于优势词义(predominant sense)自动获取的词义消歧方法.首先,由不同领域的语料获取与歧义词关系最为密切的 50 个搭配词;而后,根据这些搭配词确定歧义词在当前领域的优势词义并作为消歧结果;

(6) UofL:该方法是一种基于词汇联结(lexical cohesion)的词义消歧方法.利用 WordNet 及 eXtended WordNet 中的语义关系,为歧义词构建消歧图(disambiguation graph),根据图中歧义词各词义与上下文词的词汇联结的强度来选择正确的义项.

由表 2 的实验数据来看,尽管本文方法的效果仍低于 $BL_{MFS}$,但其已超过了其他所有同类不使用任何标注语料的无监督和基于知识库的词义消歧方法.这证明了本文所提出的消歧知识自动获取、基于依存适配度的消歧方法是切实有效的.本文方法与 TreeMatch 方法相比,$F_1$ 值高出 0.88%,这主要得益于两个方面:一是本文采用了比 TreeMatch 方法更为规范的 Reuter 语料库作为知识来源,本文的知识来源质量更为可靠;二是本文在获取依存知识时进一步区分了依存元组类型和词性信息,本文获取知识的粒度更为精细.这两方面均有利于改善消歧效果.

本文所提出的方法与其他方法在不同词性上的消歧效果对比情况见表 3.从不同词性的消歧效果来看,本文方法对于副词和形容词的消歧效果最好,名词和动词的效果较差.相对于名词和动词,副词和形容词的反义词

更为丰富.第 3.2 节中,反义词获取了最高的权重,具有更好的区分词义的作用.表 3 中的数据也验证了反义词丰富的副词和形容词确实可以取得更优的消歧效果.

**Table 3** Comparison of $F_1$ measure on different POSs (%)

**表 3** 各种方法对不同词性的 $F_1$ 值(%)

| 方法 | 名词 | 动词 | 形容词 | 副词 |
|---|---|---|---|---|
| $BL_{MFS}$ | 77.44 | 75.30 | 84.25 | 87.50 |
| **DepFit** | **72.74** | **70.05** | **82.04** | **83.65** |
| TreeMatch | – | – | – | – |
| TKB-UO | 70.76 | 62.61 | 78.73 | 74.04 |
| SUSSZ-FR | 68.09 | 51.02 | 57.38 | 49.38 |
| UofL | 57.65 | 48.82 | 25.87 | 60.80 |

本文所提出的方法与其他方法在测试集 5 篇文档上的消歧效果对比见表 4.分别从不同方法对各个测试文档的消歧效果来看,本文方法仅在 D002 文档上的效果次于 TreeMatch 及 $BL_{MFS}$ 方法,整体效果优于除 $BL_{MFS}$ 之外的各种方法.从本文方法在各个测试文档的消歧效果来看,D001,D002,D003 的消歧效果要优于 D004,D005.这与本文方法的依存知识库由 Reuter 新闻语料库的获取有关.前 3 篇测试文档均来自于华尔街日报语料库(WSJ corpus),与 Reuter 语料库同属于新闻语料库;D004 来自于维基百科对词条 computer programming 的定义解释,属于计算机领域;D005 为从 Knights of the Art 一书中摘录的关于画家 Masaccio 的人物传记,属于艺术领域.显然,D004 和 D005 中的词义的领域专业性更强.本文由新闻语料库获取的依存知识库不能很好地满足其消歧要求,而导致其效果相对略差.

**Table 4** Comparison of WSD effectiveness on different articles (%)

**表 4** 各方法对不同文档的消歧效果(%)

| 系统 | D001 | | D002 | | D003 | | D004 | | D005 | |
|---|---|---|---|---|---|---|---|---|---|---|
| | P | R | P | R | P | R | P | R | P | R |
| $BL_{MFS}$ | 85.60 | 85.60 | 84.70 | 84.70 | 77.80 | 77.80 | 75.19 | 75.19 | 74.20 | 74.20 |
| **DepFit** | **82.88** | **82.88** | **75.46** | **75.46** | **73.40** | **73.40** | **72.08** | **72.08** | **71.01** | **71.01** |
| TreeMatch | 80.71 | 80.71 | 78.10 | 78.10 | 72.80 | 72.80 | 71.05 | 71.05 | 67.54 | 67.54 |
| TKB-UO | 78.80 | 78.80 | 72.56 | 72.56 | 69.40 | 69.40 | 70.75 | 70.75 | 58.55 | 58.55 |
| SUSSZ-FR | 79.10 | 57.61 | 73.72 | 53.30 | 74.86 | 52.40 | 67.97 | 48.89 | 65.20 | 51.59 |
| UofL | 61.41 | 59.24 | 55.93 | 52.24 | 48.00 | 45.60 | 53.42 | 47.27 | 44.38 | 41.16 |

如第 2.2.3 节所述,本文依存知识库包含了依存元组的类型及词性标记信息,将其记作 DK-Base.为了进一步考察两者对于词义消歧效果的影响,本文另外构建了 3 个不同的知识库:

- 第 1 个知识库在本文依存知识库的基础上剔除了依存关系类型信息,记作 DK-Relation.该知识库在统计依存元组时不再考虑依存关系的差异,只要支配词和从属词的词形与词性相同,便认为是相同的依存元组;
- 第 2 个知识库在本文依存知识库的基础上剔除了词性信息,记作 DK-Pos.该知识库在统计依存元组时不再考虑词性的差异,只要依存关系类型、支配词和从属词的词形相同,便认为是相同的依存元组;
- 第 3 个知识库在本文依存知识库的基础上同时剔除了依存关系类型和词性信息,记作 DK-Rela-Pos.该知识库在统计依存元组时不再考虑依存关系和词性的差异,只要支配词和从属词的词形相同,便认为是相同的依存元组.

针对不同的知识库,本文对依存适配度计算方法及消歧算法进行了相应的调整,进行了一组实验,实验结果见表 5.

**Table 5** Effect of WSD performantce of different dependency knowledge (%)

**表 5** 不同类型依存知识库对消歧效果的影响(%)

| 依存知识库类型 | DK-Base | DK-Relation | DK-Pos | DK-Rela-Pos |
|---|---|---|---|---|
| 正确率 | 74.53 | 74.09 | 74.22 | 73.87 |
| 变化百分比 | 0 | –0.44 | –0.31 | –0.66 |

由表 5 可以看出,无论是剔除依存关系类型信息还是剔除词性信息,均会对消歧效果产生负面影响,其中,依存关系类型信息对于消歧效果的影响更为显著.这也证明了本文从依存元组类型和词性标记两个方面对知识获取进行细化是有效的.

## 4 结束语

本文提出了一种基于依存适配度的知识自动获取词义消歧方法.该方法充分发挥依存句法分析技术的优势,首先对大规模语料库进行依存句法分析,通过对依存元组的统计构建依存知识库;然后,对歧义词所在句子进行依存句法分析,获得歧义词的依存约束集合;根据 WordNet 获取歧义词的各类词义代表词;最后,根据依存知识库计算词义代表词在依存约束集合中的依存适配度,判断歧义词的词义.该方法仅需使用 WordNet 作为语义词典,无需其他任何人工标注语料,消歧方案简单、有效.在 SemEval 2007 的 Task#7 任务上,本文方法取得了同类方法的最佳消歧效果.

本文首次提出以依存句法分析技术为主线,获取消歧知识;综合考虑多种语义关系,计算词义的依存适配度,从而消除歧义的方法.这为解决知识匮乏及词义消歧问题提供了一种新的可行的思路.我们下一步的工作可从如下 3 个方面进行:第一,不断扩大依存知识库的规模,获得更完善、更可靠的消歧知识;第二,探索不同类型的依存关系的消歧权重问题,进一步提高消歧性能;第三,针对中文数据集实施本文方法,以考察其对不同语言的通用性.

**References**:

[1] Lu ZM, Liu T, Li S. The research progress of statistical word sense disambiguation. Acta Electronica Sinaca, 2006,34(2):333−343 (in Chinese with English abstract). [doi: 10.3321/j.issn:0372-2112.2006.02.027]

[2] Navigli R. Word sense disambiguation: A survey. ACM Computing Surveys, 2009,41(2):10:11−10:69. [doi: 10.1145/1459352.1459355]

[3] Agirre E, de Lacalle OL, Soroa A. Knowledge-Based WSD and specific domains: Performing over supervised WSD. In: Boutilier C, ed. Proc. of the Int'l Joint Conf. on Artificial Intelligence 2009. San Francisco: Morgan Kaufmann Publishers, 2009. 1501−1506.

[4] Navigli R, Litkowski KC, Hargraves O. SemEval-2007 task 07: Coarse-Grained English all-words task. In: Agirre E, Marquez L, Wicentowski R, eds. Proc. of the 4th Int'l Workshop on Semantic Evaluations (SemEval 2007). Morristown: Association for Computational Linguistics, 2007. 30−35. [doi: 10.3115/1621474.1621480]

[5] Navigli R, Velardi P. Structural semantic interconnections: A knowledge-based approach to word sense disambiguation. IEEE Trans. on Pattern Analysis and Machine Intelligence, 2005,27(7):1075−1086. [doi: 10.1109/TPAMI.2005.149]

[6] Agirre E, Edmonds PG. Word Sense Disambiguation: Algorithms and Applications. Berlin/Heidelberg: Springer-Verlag, 2007. 107−131.

[7] Liu PY. Research on unsupervised word translation disambiguation based on automatic knowledge acquisition [Ph.D. Thesis]. Harbin: Harbin Institute of Technology, 2008 (in Chinese with English abstract).

[8] Boyd-Graber J, Fellbaum C, Osherson D, Schapire R. Adding dense, weighted connections to WordNet. In: Sojka P, Choi K, Fellbaum C, *et al*., eds. Proc. of the 3rd Int'l WordNet Conf. Brno: Masaryk University, 2006. 29−35.

[9] Bergsma S, Lin DK, Goebel R. Web-Scale *N*-gram models for lexical disambiguation. In: Boutilier C, ed. Proc. of the 21st Int'l Joint Conf. on Artificial Intelligence (IJCAI 2009). San Francisco: Morgan Kaufmann Publishers, 2009. 1507−1512.

[10] Chen P, Ding W, Bowes C, Brown D. A fully unsupervised word sense disambiguation method using dependency knowledge. In: Ostendorf M, Collins M, Narayanan S, *et al*., eds. Proc. of the Human Language Technologies: The 2009 Annual Conf. of the North American Chapter of the ACL. Morristown: Association for Computational Linguistics, 2009. 28−36. [doi: 10.3115/1620754.1620759]

[11] Mihalcea R. Using wikipedia for automatic word sense disambiguation. In: Sidner C, Schultz T, Stone M, *et al*., eds. Proc. of the Human Language Technologies: The 2009 Annual Conf. of the North American Chapter of the ACL. Morristown: Association for Computational Linguistics, 2007. 196−203.

[12] Stevenson M, Guo YK. The effect of ambiguity on the automated acquisition of WSD examples. In: Kaplan R, Burstein J, Harper M, *et al*., eds. Proc. of the Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the ACL. Morristown: Association for Computational Linguistics, 2010. 353−356.

[13] Duan WS, Yates A. Extracting glosses to disambiguate word senses. In: Kaplan R, Burstein J, Harper M, *et al*., eds. Proc. of the Human Language Technologies: The 2010 Annual Conf. of the North American Chapter of the ACL. Morristown: Association for Computational Linguistics, 2010. 627−635.

[14] Lin DK. Using syntactic dependency as local context to resolve word sense ambiguity. In: Cohen PR, Wahlster W, eds. Proc. of the 35th Annual Meeting of the Association for Computational Linguistics. Morristown: Association of Computational Linguistics, 1997. 64−71. [doi: 10.3115/976909.979626]

[15] Liu PY, Zhao TJ. Unsupervised translation disambiguation based on Web indirect association of bilingual word. Ruan Jian Xue Bao/Journal of Software, 2010,21(4):575−585 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3574.htm [doi: 10.3724/SP.J.1001.2010.03574]

[16] Magnini B, Strapparava C, Pezzulo G, Gliozzo A. The role of domain information in word sense disambiguation. Natural Language Engineering, 2002,8(4):359−373. [doi: 10.1017/S1351324902003029]

[17] Wang RQ, Kong FS, Pan J. Unsupervised word sense disambiguation based on WordNet. Journal of Zhejiang University (Engineering Science), 2010,44(4):732−737 (in Chinese with English abstract). [doi: 10.3785/j.issn.1008-973X.2010.04.019]

[18] Wu YF. A survey of chinese word sense disambiguation: Resources, methods and evaluation. Contemporary Linguistics, 2009, 11(2):113−123 (in Chinese with English abstract).

[19] Fellbaum C. WordNet: An Electronic Lexical Database. Cambridge: Mit Press, 1998.

[20] Agirre E, Soroa A. Personalizing PageRank for word sense disambiguation. In: Lascarides A, Gardent C, Nivre J, eds. Proc. of the 12th Conf. of the European Chapter of the ACL. Morristown: Association for Computational Linguistics, 2009. 33−41. [doi: 10.3115/1609067.1609070]

[21] Wang RQ, Kong FS. Research on unsupervised word sense disambiguation. Ruan Jian Xue Bao/Journal of Software, 2009,20(8):2138−2152 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/3566.htm [doi: 10.3724/SP.J.1001. 2009.03566]

[22] Mihalcea R, Moldovan DI. eXtended WordNet: Progress report. In: Proc. of the NAACL 2001 Workshop on WordNet and Other Lexical Resources. Morristown: Association for Computational Linguistics, 2001. 95−100. http://www.cse.unt.edu/~rada/papers. html

[23] Yang ZZ, Huang HY. Graph based word sense disambiguation method using distance between words. Ruan Jian Xue Bao/Journal of Software, 2012,23(4):776−785 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/4116.htm [doi: 10.3724/SP.J. 1001.2012.04116]

[24] Dong ZD. HowNet knowledge database. 2000. http://www.keenage.com/

[25] Brants T, Franz A. The Google Web 1T 5-gram corpus version 1.1. LDC2006T13. 2006. http://www.ldc.upenn.edu/Catalog/docs/ LDC2006T13/

[26] Lu WP, Huang HY, Zhu CY. Feature words selection for knowledge-based word sense disambiguation with syntactic parsing. Przeglad Elektrotechniczny (Electrical Review), 2012,88(1b):82−87.

[27] Rose T, Stevenson M, Whitehead M. The reuters corpus volume 1—From yesterday's news to tomorrow's language resources. In: Zampolli A, ed. Proc. of the 3rd Int'l Conf. on Language Resources and Evaluation (LREC 2002). Paris: European Language Resources Association, 2002. 827−832.

[28] Lin DK. Dependency-Based evaluation of MINIPAR. In: Proc. of the LREC Workshop on the Evaluation of Parsing Systems (LREC'98). Paris: European Language ResourcesAssociation, 1998. 234−241. [doi: 10.1007/978-94-010-0201-1_18]

[29] de Marneffe MC, MacCartney B, Manning CD. Generating typed dependency parses from phrase structure parses. In: Calzolari N, ed. Proc. of the 5th Int'l Conf. on Language Resources and Evaluation (LREC 2006). Paris: European Language Resources Association, 2006. 449−454.

[30] Petrov S, Klein D. Improved inference for unlexicalized parsing. In: Sidner C, Schultz T, Stone M, *et al.*, eds. Proc. of the Human Language Technologies 2007: The Conf. of the North American Chapter of the Association for Computational Linguistics. Morristown: Association for Computational Linguistics, 2007. 404−411.

[31] Lei YJ. Optimization Tool with Genetic Algorithm and its Applications in Matlab. Xi'an: Xidian University Press, 2005. 95−106 (in Chinese).

[32] Koeling R, McCarthy D, Carroll J. Domain-Specific sense distributions and predominant sense acquisition. In: Mooney R, Brew C, Chien LF, *et al.*, eds. Proc. of the Human Language Technology Conf. and Conf. on Empirical Methods in Natural Language Processing (HLT/EMNLP). Morristown: Association for Computational Linguistics, 2005. 419−426.

**附中文参考文献:**

[1] 卢志茂,刘挺,李生.统计词义消歧的研究进展.电子学报,2006,34(2):333−343. [doi: 10.3321/j.issn:0372-2112.2006.02.027]

[7] 刘鹏远.基于知识自动获取的无指导译文消歧方法研究[博士学位论文].哈尔滨:哈尔滨工业大学,2008.

[15] 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.软件学报,2010,21(4):575−585. http://www.jos.org.cn/1000-9825/3574.htm [doi: 10.3724/SP.J.1001.2010.03574]

[17] 王瑞琴,孔繁胜,潘俊.基于 WordNet 的无导词义消歧方法.浙江大学学报(工学版),2010,44(4):732−737. [doi: 10.3785/j.issn.1008-973X.2010.04.019]

[18] 吴云芳.词义消歧研究:资源、方法与评测.当代语言学,2009,11(2):113−123.

[21] 王瑞琴,孔繁胜.无监督词义消歧研究.软件学报,2009,20(8):2138−2152. http://www.jos.org.cn/1000-9825/3566.htm [doi: 10.3724/SP.J.1001.2009.03566]

[23] 杨陟卓,黄河燕.基于词语距离的网络图词义消歧.软件学报,2012,23(4):776−785. http://www.jos.org.cn/1000-9825/4116.htm [doi: 10.3724/SP.J.1001.2012.04116]

[31] 雷英杰.MATLAB 遗传算法工具箱及应用.西安:西安电子科技大学出版社,2005.95−106.

**鹿文鹏**(1980－),男,山东泰安人,博士生,主要研究领域为自然语言处理,词义消歧.
E-mail: luwpeng@bit.edu.cn

**黄河燕**(1963－),女,博士,教授,博士生导师,主要研究领域为自然语言处理,机器翻译.
E-mail: hhy63@bit.edu.cn