文章编号: 1003-0077(2007)02-0003-06

基于词典属性特征的粗粒度词义消歧

吴云芳,金 澎,郭 涛

(北京大学 计算语言学研究所,北京 100871)

摘 要: 本文依据《现代汉语语法信息词典》中对词语多义的属性特征描述,对《人民日报》语料中155个词语共4 996个同形实例进行了粗粒度词义自动消歧实验,同时用贝叶斯算法进行了比较测试。基于词典属性特征的消歧方法在同形层面上准确率达到90%,但召回率偏低。其优点在于两个方面: 1) 不受词义标注语料库规模的影响; 2) 对特定词语意义的消歧准确率可达到100%。本文也讨论了适用于不同词类的消歧特征。

关键词: 人工智能;自然语言处理;特征;词义;词义消歧;贝叶斯分类法

中图分类号: TP391 **文献标识码**: A

Coarse-Grained Word Sense Disambiguation Using Features Described in the Lexicon

WU Yun-fang, JIN Peng, GUO Tao

(Institute of Computational Linguistics, Peking University, Beijing 100871, China)

Abstract: This paper presents a simple but effective feature-based approach to Chinese word sense disambiguation using the distributional features available from the Grammatical Knowledge-base of Contemporary Chinese. The test data is the sense-tagged corpus of People's Daily. A Naïve Bayes classifier is also tried as a comparable statistical method. The feature-based approach achieves precision of 90%, which is comparable to the NB classifier. The striking advantages of the feature-based approach are 1) It is not influenced by the data size, and 2) It can disambiguate some specific words with precision of 100%. The features appropriate for different parts of speech in Chinese WSD are also discussed. This paper demonstrates that sense features described in the lexicon are worth including in WSD.

Key words: artificial intelligence; natural language processing; feature; word sense; word sense disambiguation; Naïve Bayes classifier

1 引言

词义消歧长时期以来一直是语言信息处理中的热点难题。作为一个较为典型的分类问题,各种机器学习方法都曾应用在词义消歧研究上,例如基于AdaBoost MH的分类方法[1],最大熵方法词义消歧[2],贝叶斯分类法[3],运用搭配知识进行词义消歧[4]等。但由于大规模、高质量的词义标注语料的缺乏,有指导的词义消歧研究很难取得突破性进展。机读词典是词义消歧研究中经常利用到的重要知识资源。Lesk[5]通过计算歧义词的各个词义在词典中的定义与歧义词上下文词语的定义之间的覆盖度,选择覆盖度最大的作为正确的词义,实验结果达到了50%-70%的准确率。Yarowsky[6]根据Roget的词语语义类来进行词义消歧,当词语意义涉及多个主题时,算法效果不甚理想。本文依据《现代汉语语法信息词典》[7](文中简称《语法词典》)来进行词义消歧,不同于传统词典的定义训释方式,《语法词典》采用复杂特征集的形式描述一个词语的句法组合功能,再现了词语可能出现的上下文环境,而上下文环境乃是词义消歧的最终凭借。

收稿日期: 2005-11-04 **定稿日期**: 2006-12-20

基金项目: 国家973计划资助项目(2004CB318102)

作者简介: 吴云芳(1973—),女,博士,主要研究方向为中文信息处理,现代汉语句法和语义。

本文的词义消歧分主要依据《语法词典》中的同形信息。《语法词典》对“同形”的定义和传统语言学稍有不同,区分同形主要是依据词语的语法功能,意义相近并且语法功能相同则归入一个同形,较《现代汉语词典》等传统辞书的义项区分要显得粗糙一些,可看作是一种粗粒度的词义划分。北京大学计算语言学研究所正在进行词义标注语料库的研究与建设,开发了同形标注校对辅助软件,手工标注了《人民日报》2000 年 1 月约 200 万字语料中的同形信息,这是本文算法训练、测试和评价的数据基础。

2 属性特征和消歧算法

一般认为,词语的不同意义在句法组合上会显现出差异,当今的词汇语义研究大都力主根据词语的句法分布来分析词义。《语法词典》以复杂特征集为形式手段,以词类为纲,描述了词语不同意义的句法组合特征。例如动词“保管”:

表 1 《语法词典》中“保管”的属性特征描述示例

词语	同形	释义	体谓准	动趋	动介	着了过	重叠	aabb	备注
保管	①	保藏,管理	体	趋	在	着了过	ABAB		～粮食/～图书
保管	②	担保;有把握	谓						～甜/～你不吃亏

“词语、同形、体谓准……”等都是属性名(Attribute),“保管、①、谓……”等是相对应的属性值(Value)。表 1 清晰地展示出了“保管①”和“保管②”在句法组合上的差异,藉此差异可在语料文本中正确辨别出同形。例如下面语料中的句子:

(1) 你替我保管着,下午我来付款。

“保管①”的属性“着了过 = 着了过”,“保管②”的属性“着了过 = 否”,由此可判定例句(1)中当是保管①。对于一个词语的多个同形条目,同一个属性字段相异的取值即构成同形词之间的区别特征(Distinguish Features)。例如对于“保管”,“着了过 = 着了过”构成“保管①”区别于“保管②”的一个属性特征,“体谓准 = 谓”构成“保管②”区别于“保管①”的一个属性特征。

定义 1 词语 W 可区分为 n 个同形 S_1, S_2, … S_n $(n > 1)$,同形 S_i 用复杂特征集来描述 $S_i\begin{bmatrix} f_1 = v_1 \\ f_2 = v_2 \\ \cdots\cdots \\ f_m = v_m \end{bmatrix}$ $(m \geq 1)$,词语 W 的不同同形 S_i, S_j 存在相同的属性特征 f_k,设 $S_i(f_k = v_{ki})$, $S_j(f_k = v_{kj})$,若 $v_{ki} \neq v_{kj}$,则称 $f_k = v_{ki}$ 是 S_i 对 S_j 的区别特征,对应的 $f_k = v_{kj}$ 是 S_j 对 S_i 的区别特征。

在实际的算法设计中,有必要区分肯定性区别特征(Positive Distinguish Feature)和否定性区别特征(Negative Distinguish Feature),属性值为“是”或者有具体取值的为肯定性区别特征,属性值为“否”或者留空的为否定性区别特征。例如对于“保管”,“着了过 = 着了过”是“保管①”区别于“保管②”的一个肯定性区别特征,而“着了过 = 空”是“保管②”区别于“保管①”的一个否定性区别特征。肯定性区别特征可以帮助识别同形,如上文例句(1)所示,而否定性区别特征却不具备这种功能,请看下面的例句:

(2) 报纸归个人,保管、阅读更仔细。

“保管”在句子中没有后接“着了过”,但这并不能就证明“保管”是意义②。《语法词典》可以这样但不一定时时处处都这样。

定义 2 在《语法词典》中区别特征 $f_k = v_{ki}$ 是肯定性区别特征当且仅当 v_{ki} 不为空。基于《语法词典》描述的属性特征生成相应文件,记录词语不同同形的肯定性区别特征。例如“保管”:

(3) 保管.txt

保管① 体谓准 = 体,动趋 = 趋,动介 = 在,着了过 = 着了过,重叠 = ABAB

保管② 体谓准 = 谓

基于词典属性特征进行词义消歧的基本思路是,检查待消歧的目标多义词所在的上下文是否满足词典中特定同形的属性特征约束,若满足则确定为该同形的意义。上下文语境是词义消歧的知识来源,语境范围的选取会影响到消歧的效率。本文以多义词所在句子作为上下文语境范围。

定义 3 设标点符合集 Sen_Mark = {分号(;),句号(。),问号(?),感叹号(!),冒号(:)},mark ∈ Sen_Mark 标示左边一个句子的结束和右边一个句子的开始。

词典属性特征词义消歧算法描述

(a) 依据《语法词典》,对每一个多义词 W,比较不同同形的属性特征进而找出相互之间的肯定性区

别特征,对每一个同形 S_i,以 $f_k = v_{ki}$ 的形式列出其肯定性区别特征,对每一个多义词 W 生成一个属性特征文件 W_Lex_Rule(如上文"保管.txt");

(b) 在语料文件中定位目标多义词 W,以句子范围作为其上下文语境 C;

(c) 对 W 的不同同形赋值 $S_i.Score = 0$;

(d) 检索文件 W_Lex_Rule,提取同形 S_i 的肯定性区别特征,判断语料中 W 所在的上下文 C 是否满足约束条件,若满足,则 $S_i.Score = S_i.Score + 1$;

(e) 若文件 W_Lex_Rule 中属性特征列表非空,重复(d);

(f) $Score$ 取值最大的同形 S_i 为标注结果。

3　贝叶斯词义消歧

贝叶斯分类法在词义消歧研究中得到了广泛应用[8],这个简单的模型在词义消歧取得了很好的效果。为了对词典特征方法的消歧效果有更为客观的评价,本文同时运用了贝叶斯分类方法,同样的语料进行了自动词义消歧测试。

3.1　贝叶斯算法

设词语 W 可区分为 n 个同形 S_1, S_2, …, $S_n (n>1)$, W 在上下文语境 C 中出现,分类器的目标是寻找一个 S_i,使得 $P(S_i|C)$ 最大。根据贝叶斯公式:

$$P(S_i \mid C) = \frac{P(C \mid S_i)\, P(S_i)}{P(C)}$$

对于所有的 S_i 而言, $P(C)$ 是一个常量,因此目标变为搜索 $S_{\max}$:

$$S_{\max} = \arg\max_i P(C \mid S_i)\, P(S_i)$$

设上下文特征 C 由 m 个特征组成 $C_1, C_2 \cdots C_m (m \geq 1)$,应用单纯贝叶斯假设,这些特征彼此之间关于 S_i 条件独立,则有:

$$P(C \mid S_i) = \prod_{j=1}^{m} P(C_j \mid S_i)$$

根据最大似然估计:

$$P(C_j \mid S_i) \approx \frac{Count(C_j, S_i)}{Count(S_i)}, \quad j = 1 \cdots m$$

$$P(S_i) \approx \frac{Count(S_i)}{Count(W)}$$

当由于数据稀疏 $Count(C_j|S_i)$ 为 0 时,取平滑参数 $\alpha = 0.01$。

3.2　特征 C 的定义

上下文语境范围限定为一个句子(见定义 3),特征 C 包括词性和词形两类特征:

1) 词性特征。以当前多义词为中心左右 n 个位置上的词性,以"$PI = Pi$"($-n \leq I \leq n$)的形式表示,$-I$ 表示目标词语左边第 I 个位置,$+I$ 表示目标词右边第 I 个位置。例如"$P+1 = v$",表示目标词右边第一个词的词性为动词。

2) 词形特征。以当前多义词为中心左右 n 个位置上的词形,以"$WI = Wi$"($-n \leq I \leq n$)的形式表示,例如"$W-1$ = 有",表示目标词左边第一个词是"有"。

4　实验结果考察

4.1　词义消歧准确率

实验语料为 2000 年 1 月《人民日报》语料,根据《语法词典》中的同形描述人工对语料进行了同形标注和校对。语料被随机分割成训练语料和测试语料,训练语料占 2/3,测试语料占 1/3。贝叶斯算法在训练语料中训练数据,在测试语料中进行测试,词典特征方法直接在测试语料中进行测试。实验中,我们有意去除了语料中最高频出现的前 11 个同形词"是/v、有/v、要/v、上/f、为/p、出/v、而/c、使/v、活动/v、来/v、会/v",因为这几个词意义用法复杂,《语法词典》对它们的同形划分显得有些粗糙,它们的加入会影响对消歧效果的正确评价。测试语料包含 155 个词语共 4 996 个同形实例。用准确率和召回率来评测标注效果,将最大频率词义消歧作为基准准确率(baseline)。

$$\text{准确率(precision)} = \frac{\text{标注正确的同形数目}}{\text{标注的总同形数}} \times 100\%$$

$$\text{召回率(recall)} = \frac{\text{标注正确的同形数目}}{\text{语料中总的同形数目}} \times 100\%$$

表 2　同形自动消歧结果

消歧效果	词典特征方法	贝叶斯算法	最大频率词义消歧
准确率	90.31%	90.29%	84.56%
召回率	13.25%	90.27%	84.56%

表 2 显示,词典特征方法和贝叶斯算法词义消歧的准确率基本相同,约为 90%,明显高于最大频率方法的基准准确率,但在召回率上词典特征方法

明显偏低。

图 1 显示,贝叶斯分类法随着词语出现频度的降低而准确率缓慢下滑,词典特征方法则不受词语出现频度的影响。贝叶斯算法和其他统计方法一样,处理效果依赖于语料的数据规模,词语频度越高,训练学习的数据越充分,准确率就越高。词典特征方法是依据词典中提供的语言知识进行消歧,与词语出现的频度无关,不受数据规模的影响。大规模词义标注语料库的建设已成为词义消歧研究的瓶颈,决定着统计方法消歧效率的优劣。词典属性特征的知识方法消歧为真实语料中低频词语、低频义项的消歧问题提供了解决方案,而这正是统计消歧方法最为犯难之处。

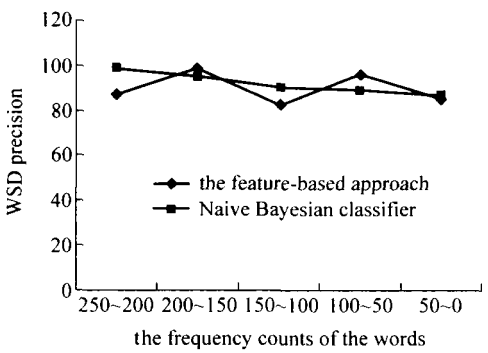

图 1 消歧准确率受词语频度的影响

图 2 显示,对于不同的词语,词典特征方法的消歧准确率要么是 0,要么是 100 %,而处于 0 - 100 % 中间状态的词语只是少数。贝叶斯算法则恰好相反,多数词语词义消歧准确率处于 0 - 100 % 的中间状态,只有少数词语准确率为 0 或 100 %。在进行消歧处理的 155 个词语中,有 117 个词语(117/155 = 75 %)利用词典属性特征方法其消歧准确率可达到 100 %,表 3 显示了部分这样的词语。对于量词 q 词类,词典属性特征方法的消歧准确率达到了 100 %,远远高于贝叶斯算法。这显示了词典特征方法对词语敏感的(Word-sensitive)的特性:当某条

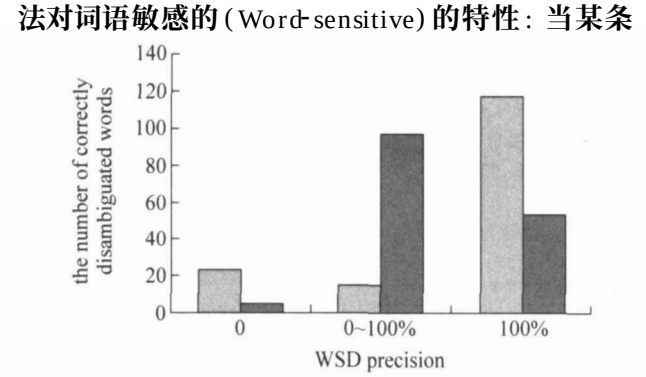

图 2 词语消歧准确率分布

(些)规则可以捕捉到某多义词的特性时,就对该多义词具有很高的消歧准确率(经常为 100 %);如果规则不能反映词语的句法语义特性,消歧准确率就干脆为 0,而处于中间状态的词语仅占少数。对于统计方法而言,使词义消歧的准确率达到 100 %是很难企及的一个目标。

表 3 部分利用词典属性特征方法消歧准确率达到 100 %的词语

词语	词类	词义消歧准确率		
	词典特征方法	贝叶斯算法	最大频率词义消歧
走	100 %	92 %	95 %
等	100 %	86 %	81 %
拿	100 %	72 %	75 %
写	100 %	67 %	64 %
带	100 %	58 %	54 %
逃	100 %	37 %	75 %
玉米	100 %	31 %	37 %
q	100 %	89 %	78 %

4.2 词义消歧特征

在上述实验结果的基础上,分析考察不同属性特征对词义消歧效率的影响,这无论对词义消歧研究,还是对词义辨析的语言学探讨,都有着重要的意义。Dang & Palmer[9], Yarowsky & Florian[10]论证指出,较之不同的算法选择,特征空间对词义消歧的性能有着更大的影响。

本文的词典属性特征方法一共用到了 65 个属性特征,其中用于动词词义消歧的属性多达 49 个,名词有 8 个,形容词没有合适的属性特征用于词义消歧,副词、语气词、量词、处所词、时间词各用到了少数几个属性特征。对 65 个属性特征进一步分析,消歧准确率为 100 %、消歧词数大于 10 的属性特征如表 4 所示。

表 4 显示,不同词类对应不同的消歧特征。量词 q 只有一个属性特征"后名"用于词义消歧,但整体消歧效率是最好的,准确率达到 100 %,而贝叶斯算法对量词的词义消歧准确率只有 89.68 %。相对应的,用于名词词义消歧效率最好的属性特征是修饰它的量词。名词表称具有空间特性的事物,其主要语法特点之一是接受数量词修饰,而量词的主要句法功能是修饰名词,因此互为搭配的名词和量词

制约消歧是符合逻辑的结果。动词后接的趋向动词可以比较有效地区分多义动词,《语法词典》动趋式分库设有 29 个属性字段,其中 16 个趋向动词都可以高效地区分出动词的义项。带趋向动词作补语是表动作行为义动词的主要语法特性,而非动作行为义动词像认知义动词一般不能带趋向补语。

表 4　消歧准确率为 100 %、消歧词数大于 10 的属性字段

词　类	属性名 (正确消歧词数)	示　例
量词 q	后名(179)	【轮】①表循环的事物、动作 ②多用于太阳、明月 例: a 一轮 ②红日喷薄而出。 b 将参加叙以第二轮 ①谈判。
动词 v	动趋(127): 下(29)、出来(25)、上去(16)、起来(9)、出去(8)、下来(7)、下去(7)、过(5)、v 来 v 去(4)、过去(4)、进来(4)、回来(3)、进去(3)、过来(1)、上来(1)、回去(1)	【拍】①用手掌打 ②拍摄 例: 能够拍 ②下黑洞照片的方法。
动词 v	介宾的后(33)	【支持】①支撑;勉强维持 ②鼓励并帮助 例: 对国有企业改革和发展的支持 ②。
动词 v	动结(32): 好 (32)	【吃】①咀嚼吃饭 ②受,挨 例: 四季有肉、吃饱吃 ①好。
名词 n	量词(32): 个体量词(28)、度量词(3)、 容器量词(1)	【机组】①一组机器 ②飞机上全体员工 例: 小浪底水利枢纽首台机组 ①发电。
动词 v	重叠形式(16): V 一 V(10)、V 了 V(6)、	【笑】①欢笑,大笑　②讥笑 例: 想了许久,笑 ①了笑:
副词 d	搭配(12)	【快】①赶快　②将要 例: 春节快 ②到了。
时间词 t	后时(12)	【当年】①过去某段时间　②同一年 例: 当年 ②7 月/t

贝叶斯分类法假设特征之间是彼此独立的,应用了一定窗口范围内不同特征的综合;而词典属性特征方法只是应用一个(或几个)显著的特征。Yarowsky & Florian[10] 根据特征运用方式的不同将词义消歧算法划分为两类: 综合型的(Aggregative)和离散型的(Discriminative)。本文的词典属性特征方法应当归属于离散型的特征运用,实验结果表明,现代汉语中的一些多义词存在着显著的少数几个特征可对其进行高效的词义消歧,而找寻这些显著特征是研究的重点和难点。

表 4 中用于词义消歧的特征大多局限于具体词形,只有时间词属性"后时"用到了词性信息 t,这主要是由于缺乏高效的句法分析器而使词典中的很多句法信息无法付诸应用,像词形如此细粒度的信息就直接导致了词典特征方法的覆盖性差,召回率低。关于这一点,后文还有论述。

5　对于规则方法的讨论

词典特征方法大致可看作是规则方法的代表,贝叶斯分类法大致可看作是统计方法的代表,基于本文的实验结果,可以来审视一下规则方法的是与非。

词典特征方法和贝叶斯算法形成鲜明对比的是,前者的召回率(仅为 13.25 %)远远小于后者(为 90.27 %),这当然主要归咎于运用于词义消歧的语言知识还太贫乏,不能覆盖真实文本中丰富复杂的语言现象。本文的具体应用中,语言知识的贫乏主要体现在以下两个方面。其一,对于同形词语之间的区别特征还把握不够,在《语法词典》中没有设置字段描述。例如动词对论元的语义选择限制可以高效地消除动词以及名词的歧义,但《语法词典》没有加以描述。《语法词典》描述的是主要词语的语法信

息,单纯语法信息还无法实现词义自动消歧这样高难的任务。其二,《语法词典》中包含了知识描述,但目前语言自动处理水平还无法充分、直接应用这些知识。当前词义消歧研究所基于的语料大多只是经过了词语切分和词性标注,而充分有效地利用句法知识需要高效的句法分析。例如动词"保管":

(4) 保管 ①保藏,管理 体谓准=体 ……
保管 ②担保,有把握 体谓准=谓 ……

对下面看似简单的句子:

(5) 这西瓜保管你满意。

要准确辨析(5)中"保管"的意思,需首先作出如下正确的句法分析:

(6) NP[这/r 西瓜/n] 保管/v ZW[你/r 满意/v]。

ZW表示主谓结构,具有谓词性功能,因此可判定(5)中的"保管"表示②义"担保"。然而,成熟高质的现代汉语句法分析器还不多见,对语料中的句子进行准确的句法分析还是计算语言学研究者奋斗的一个目标。现阶段自然语言的处理水平使已知的词义消歧知识无法付诸应用,也是造成规则方法召回率不如人意的原因。提高规则方法对真实语料的覆盖率,依然是任重而道远。

词典属性特征方法的消歧准确率为90%,其错误主要来源于两方面的原因。其一,"判断语料中*W*的上下文是否满足约束条件"时,由于缺乏句法分析,本文只作了简单粗糙的近似匹配,这就难免引发错误。例如对下面句子中的"想"进行词义消歧:

(7) a 努拉/n 婉拒/v 再三/d,/wd 却/d 想/v 出/v 变通/v 的/u 办法/n。
b 真正/d 想/v 出/v 政绩/n 的/u 干部/n

(7b) 中正确的分析应该是[想/v[出/v 政绩/n]],而不是[[想/v 出/v] 政绩/n]。类似这样的错误占了很大的比例。当汉语自动句法分析研究取得突破之后,此类问题自可迎刃而解。其二,像其他各式词典一样,《语法词典》储藏的是词语的静态知识,这种知识反映了人们对于词语用法意义的一般认识,却难以应对真实语言使用中丰富复杂的变异。

(8) 三讲,进②出了好作风。

(9) 把其中1.4万余元又全部输②进了"游戏厅"。

作为"讲求"意义解的"讲"一般不能带趋向动词"出",像"讲文明,讲礼仪",但在句(8)中却带着"出"出现。作为"输送"意义解的"输①"经常带趋向动词"进",而作"输赢"意义解的"输"一般不带趋向动词"进",但在句(9)中却带了"进"补足说明"输"的场所。如何解决语言动态使用中部分违反"语言规律"的现象,是很值得探讨的一个难题。

粗粒度词义消歧可算作是具体而微的自然语言处理的一个应用,那么,上文所分析的特征方法在词义消歧中的不足大体也可看作是自然语言处理中规则方法的普遍性缺陷。就同形消歧这个任务而言,词典特征方法表现出了不受数据规模影响以及对词语敏感的特性,而这个特性正好是像贝叶斯这样的统计方法所缺乏的。对于某些利用特征方法消歧准确率可达到100%的词语,不妨先利用特征方法进行消歧,对于特征方法无法召回的句例,再利用其他统计方法,譬如像量词q的消歧就可采取这种策略。

参考文献:

[1] 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.

[2] 李涓子.汉语词义排歧方法研究[D].清华大学计算机科学与技术系博士学位论文.1999.

[3] 卢志茂,等.神经网络和贝叶斯网络在汉语词义消歧上的对比研究[J].高技术通讯,2004,(8).

[4] 全昌勤,等.从搭配种子获取最优种子的词义消歧方法[J].中文信息学报,2005,19(1):30-35.

[5] Lesk, M. E. Automated sense disambiguation using machine-readable dictionaries: How to tell a pine cone from an ice cream cone [A]. In: Proceedings of the SIGDOC Conference [C]. 1986.

[6] Yarowsky, D. Word-sense disambiguation using statistical models of Roget's categories trained on large corpora [A]. In: Proceedings of COLING 92 [C]. 1992.

[7] 俞士汶,等.现代汉语语法信息词典详解[M].北京:清华大学出版社,2003.

[8] Niu, ZH. Y., Ji, D. H. and Tan, Ch. L.: Optimizing Feature Set for Chinese Word Sense Disambiguation [A]. In: Third International Workshop On The Evaluation of Systems for the Semantic Analysis of Text [C]. 2004.

[9] Dang, H. T. and Palmer, M.: The Role of Semantic Roles in Disambiguating Verb Senses [A]. In: Proceedings of the 43th Annual Meeting of the ACL [C]. 2005.

[10] Yarowsky, D. and Florian, R. Evaluating Sense Disambiguation Performance Across Diverse Parameter Spaces [J]. Journal of Natural Language Engineering, 2002.