基于混合特征的汉语形容词词义区分研究

朱虹 贾玉祥 刘扬

（北京大学计算语言学研究所，北京 100871）

{zh,yxjia,liuyang@pku.edu.cn}

摘 要： 词义知识获取问题是词义消歧、词义知识库建设、语料库建设等不同研究的瓶颈问题。本文提出的基于混合特征的词义区分方法，通过发现并抽取易于获取的词义特征，结合 EM 迭代算法，能够很好地对汉语高频形容词实现词义区分。比较于不同的特征组合方式，实验结果证明，形容词主体名词特征和词形特征的使用对于汉语形容词的词义区分是有效的。

关键词： 形容词；词义区分；特征选择；EM 算法

Chinese Adjective Sense Discrimination with Multiplex Features

ZHU Hong, JIA Yuxiang, LIU Yang

(Institute of Computational Linguistics, Peking University, 100871)

{zh,yxjia,liuyang@pku.edu.cn}

Abstract: Word knowledge acquisition is the bottleneck for many fields like word sense disambiguation, word knowledge base construction, corpora construction. In this paper, we propose a word sense discrimination algorithm for Chinese high-frequency adjectives with multiplex easy-acquirable features into one frame. The results show that our feature selection is effective for Chinese adjective word sense discrimination.

key words: adjective; word sense discrimination; feature selection; EM algorithm

1 引言

自然语言的语义分析是实现自然语言理解的必要手段，其中面向信息处理用的词义分析一直是自然语言处理的焦点和难点。传统的词义研究主要关注词义的发展和演变，词典对于词义的定义多是描述解释性的，很难反应词语在真实语料中的词义情况。并且词义粒度过细，缺少新词义或者领域相关的词义[1]，已经成为词义消歧（Word Sense Disambiguation）、词汇语义知识库建设等研究的瓶颈[2]。因此，面向信息处理的自动词义区分（Word Sense Discrimination）成为了解决词义知识获取问题的重要研究课题。

自动的词义区分基于著名的分布假设，即词语的意义可以由它的上下文确定并描述。它从大规模真实文本中提取词语的不同用法特征，用聚类方法将词语分类，不同的类用来代表不同的语义。不同于词义消歧，词义区分不使用标注语料库，并且事先不确定每个词语有几个词义，以及所有词语能形成多少意义。词义区分可以为词义消歧提供知识源，还

基金资助：本文相关研究得到全国博士学位论文作者专项资金资助项目（200514）的支持。

能帮助大规模语料的词义标注，以及有效地辅助机器词典的构建。此外，它还被用于文档分类、信息检索等不同领域。

词义区分在汉语中已经有所研究，但是效果没有英语的好。究其原因，首先，汉语的词义区分通常在已经完成了切分标注的语料上进行，切分标注的错误会直接影响到词义区分的效果。其次，特征的发现和提取一直是词义区分的重点和难点，很多研究都建立在对语料进行句法分析之上，这样可以大大提高特征选取的有效性，减少噪声。但是目前还很难在汉语中找到一个准确率很高的句法分析器。因此如何在不引入其他词义资源的情况下，有效地发现并提取词语的语义特征，最终实现词义的自动区分，成为了本文主要解决的问题。

本文针对汉语中的高频形容词，提出了一种基于混合特征的词义区分方法，该方法选用易于获取并且能够自动获取的特征进行词义区分，不需要引入其他的词义资源以及人的介入。我们首先在第二章中重点描述特征的选择与抽取方法。选用的特征有：基于大规模新闻语料的形容词的分布信息，基于网络语料的形容词的主体名词信息，形容词的词形信息。其中，词语的分布信息根据上下文窗口的不同，又分为 bigram（二元，目标词的前后各一词距离）搭配信息和句子范围内的搭配信息，作对比研究之用。我们融合这些不同类型的特征，计算得到所有目标词的初始分类，然后将初始分类作为先验知识，应用于 EM 迭代算法实现词语的自动词义区分。具体算法在第三章中详述。第四章中我们将介绍我们的实验数据，提出我们的评价方法，并且比较不同特征组合形成的实验结果。实验结果证明形容词主体名词信息和形容词词形信息等混合特征的使用可以有效提高形容词词义区分的准确率和召回率。最后，我们在第四章中总结全文，并提出了对未来工作的设想。

2 特征选择

影响词义的因素有很多，包括词类、词语形态、词语内部结构、词语之间的结构（重叠、词序）、句法功能（能否充当、充当什么成分）、组合能力（跟某类词的组合）、表类别作用功能（指代、连接）等等。另外，特定上下文中表现出来的情态、文章的风格、语体、语域等不同因素也都影响词语的词义。这些特征可以归纳为四类：词法、句法、语义以及语境。其中句法功能、表类别作用等句法特征需要正确率高的句法分析器的支持。而人们对于语境类特征的研究尚处于起步阶段。因此，如何抽取易于获取并且反映真实词义的特征对词义区分至关重要。

2.1 上下文分布特征

基于分布假设，词语的意义可以通过周围词语发现并描述。因此，目标词在语料中的上下文信息，例如词语的特定组合关系，是词义区分的重要特征。由于没有正确率高的词法分析器的支持，我们选用了目标词的搭配（collocation，在本文中也称为“共现co-occurrence”）作为特征来描述目标词的上下文分布情况。一般认为，搭配是一种融语法、语义、习惯用法的等多种因素的现象。许多实验也证明，搭配是词义区分的有效特征。

文本选用了卡方假设检验作为形容词搭配词抽取的最终方法。为了比较不同搭配窗口对词义区分的影响，我们分别就 bigram 搭配窗口和句子范围窗口（以逗号等标点符号为界限，一般叫小句范围）作对比实验。举例来说，S1 和 S2 是“宝贵”一词的两个上下文片

断（经过了分词和词性标注），如果是 bigram 搭配窗口限定，可以从 S1 中抽取出搭配“宝贵/a 财富/n”。如果是句子范围窗口限定，可以从 S2 中抽取出搭配“宝贵/a 资源/n”，如黑体标识出来的。

S1：“……，/w 是/v 我们/r 党/n 和/c 人民/n 用之不竭/i 的/u ***宝贵/a 财富/n*** 。/w”

S2：“一个/m 国家/n 人民/n 的/u 素质/n 才/d 是/v 最/d ***宝贵/a*** 的/u 旅游/vn ***资源/n*** 。/w”

2.2 主体名词特征

属性-主体关系，特别是“形容词(a)—名词(n)”关系，是形容词的重要组合关系。属性词和主体词在意义上是相互约束的。例如，构成属性—主体关系的“白-雪”，在“雪”的定义中包含了“白”的描述，在“白”的定义中举了“雪”作为实例。有同一属性的主体词在意义上也往往是相关的。例如，“慈祥”一词的主体名词“爷爷、奶奶、老人、长辈”等都含有“老年人”的意义。为了获取这类关系，我们利用搜索引擎从大规模网页中抽取出“像 n 一样 a”、“如 n 般 a”等比喻句式。因为我们发现，这类比喻句式中的形容词和名词有很多都属于属性-主体关系。例如 S4 中的“广阔-天空”，S5 中的“美丽-童话”。

S4：“……,/w 像/p 天空/n 一样/u 广阔/a 。/w”

S5：“……天空/n 飘/v 着/u 雪花/n 那/r 也/d 如/v 童话/n 般/u 美丽/a……”

当然，这种获取方法有它的不足。一方面，许多词语很少被用在比喻句中，因此它们的信息很难获取。另一方面，依靠这类句式抽取出来的词对不一定是属性-主体关系，或者不是典型的属性-主体关系对。例如 S6 中的“熟-猪”，以及 S7 中的“和蔼-太阳”。

S6：“这/r 家伙/n 睡/v 得/u 像/v 死/b 猪/n 一样/u 熟/a 。/w”

S7：“……像/p 太阳/n 一样/u 和蔼/a……”

2.3 词形特征

汉字是表意文字。词语中的汉字不但可以是表音符号，往往也是表意符号[3]。因此，词形相似的词语，词义往往也相似。这是汉语较英语不同的地方。例如，汉语中“广阔”和“辽阔”，它们的最后一个字（后缀字）都是“阔”。在词义上，“广阔”解释为“广大宽阔”，“辽阔”解释为“广阔；宽广”（见《现代汉语词典（第 5 版）》），两个词都含有“阔”这个语素义。另外，词义相近并且词形相似的词语可以是第一个字相似，例如“宏伟、宏大”；也可以是最后一个字相同，例如“广阔、辽阔”；或者是中间某个字相同，例如“红火、火爆”。通过大量观察，我们最终选择了词语的后缀字作为词义区分的特征。

3 词义区分算法

3.1 混合特征计算

本文选用的特征包括上下文分布特征、主体名词特征和词形特征，涵盖了句法、语义、词法三个方面，易于获取，并且不需要其他词义资源的介入。由于这三类特征的来源和内容形式互不相同，因此我们首先对三类特征分别作预处理，然后将它们融合起来供词义区分之用。

根据所有目标词$\bigcup_{i=1}^{N} W_i$的上下文分布特征，从中抽取特征词（目标词的典型搭配词），然后以这些特征词为分量，将目标词表示成特征向量，分量的值反映目标词与该特征词的紧密程度。一般的，分量的值可取目标词与特征词在语料中的共现频次。本文使用文献[4]中介绍的改进的互信息方法设置分量的值。见公式（1）。

$$RMI_{w,f} = \frac{F_f(w)}{F_f(w)+1} \times \frac{\min[\sum_i F_{f_i}(w), \sum_j F_f(w_j)]}{\min[\sum_i F_{f_i}(w), \sum_j F_f(w_j)]+1} \tag{1}$$

其中，$F_f(w)$表示词语 w 与特征词 f 在语料中的共现频次。该方法可以有效改善互信息对低频特征的偏向问题。

依据形容词的主体名词特征，每个形容词 w 都能对应到一个主体名词集合 N(w)。对于所有的形容词和主体名词，我们可以建立起一个形容词-主体名词矩阵，如果某名词是该形容词的主体名词，相对位置置 1，否则置 0。用式（2）表示。

$$\text{AN-Matrix}(w,n) = \begin{cases} 1 & n \in N(w) \\ 0 & n \notin N(w) \end{cases} \tag{2}$$

依据这个矩阵，每个形容词有一个主体名词为分量的 0-1 向量与之对应。利用 simple-kmeans 聚类算法，将所有目标词分类。目标词间的相似度是主体名词向量的余弦系数（cosine coefficient）值。主体名词分布情况相似的形容词，被分到同一类中。由此形成关于目标词的第一个分类$\bigcup_{i=1}^{A} S_i$，满足$\sum_{1}^{A} |S_i| = N$。

对于词语的词形特征，我们将所有目标词按照它们的后缀字分组，具有相同后缀字的词被分到同一组中。由此形成关于目标词的第二个分类$\bigcup_{i=1}^{B} S_i$。

我们将这两个分类合并形成$\bigcup_{i=1}^{A+B} S_i$，然后按照每个类的语义紧密度（Sense Tightness）将所有类倒序排列，紧密度高的类排在前面。语义紧密度的具体定义见式（3）。

$$\text{Sense\_Tightness}(S_t) = \begin{cases} \dfrac{\sum_{w_i,w_j \in S_t} \text{sim}(w_i,w_j)}{C^2_{|S_t|}} & \text{if } |S_t| < m \\ 0 & \text{if } |S_t| \geq m \end{cases} \tag{3}$$

其中，$\text{sim}(w_i,w_j)$表示类中任意两个词语分布特征向量的余弦系数值。见式（4）。

$$\text{sim}(w_i, w_j) = \frac{\sum_f RMI_{w_i f} \times RMI_{w_j f}}{\sqrt{\sum_f RMI_{w_i f}^{\ 2} \times \sum_f RMI_{w_j f}^{\ 2}}} \tag{4}$$

语义紧密度是用来衡量类的总体语义相似度的一个指标，类中词语越相似，该值就越大。由于具有相同后缀字的词语在语义上不一定相关，因此，我们通过定义语义紧密度来发现真正有助于词义区分的信息。另外，考虑到存在一些成词性特别强的字，以这些字作为后缀字的类，大小会非常大。当类的大小特别大时，我们就不再考虑使用这些类的信息，将它们的语义紧密度被视为 0。本文中我们设 m=10。

由此，我们得到了一个关于所有目标词的初始分类$\bigcup_{i=1}^{C} S_i$，它的形成融合了上下文分布特征、主体名词特征、词形特征这三类不同类型的特征。我们将这个初始分类作为 EM 迭代算法的先验知识用于词义区分。

3.2 EM 迭代算法

EM 是一个使用期望最大化的方法对数据缺失的问题进行估计的算法。它分为 E（Estimate）和 M（Maximize）两个步骤。E 步骤利用当前参数空间计算概率模型中隐藏变量的期望值，M 步骤利用已经计算出来的隐藏变量的期望值重新计算参数空间的极大似然估计。

在词义区分问题中，我们的目标是利用EM算法对所有目标词作一个分类。参数空间的参数是所有的$P(C_j)$和$P(f_i \mid C_j)$，其中$\bigcup_{i=1}^{|C|} C_i$是关于目标词的分类结果，每个目标词被表示成上下文分布特征向量$(f_1, f_2, ...f_n)$，f_i是某个分布特征。隐藏变量是$P(C_j \mid (f_1, f_2, ...f_n))$，即每个目标词属于每个类的概率值。

E 步骤和 M 步骤重复进行直到满足收敛条件或者完成指定的迭代次数。本文我们设最大迭代次数为 50 次。

EM 算法的最大缺点就是可能陷入局部最优，因此参数初始值的设定非常重要。最一般的方法是令参数初始值随机化，因为假设我们对参数一无所知。本文中则利用了混合特征的分类结果来初始化参数。实验证明，我们的方法提供的先验知识对词义区分是有效的。

4 实验

4.1 实验数据准备

本文使用人民日报 1998 年上半年语料，根据北京大学的词性标注规范对语料进行了切分标注。该语料包括至少 876,811 个句子以及 6,176,565 个词次。我们从中抽取出高频形容词 775 个。为了比较不同搭配窗口对于词义区分的影响，我们分别就 bigram 搭配窗口和句子范围窗口作独立试验。其中 bigram 搭配窗口选定 3368 个特征词，句子范围选定 4936 个特征词。

4.2 评价方法

结果的评价是词义区分的最大难点之一。自动的词义区分会将语义高度相关的词语分到同一类中。但是，词义区分一般事先不知道每一个词语有几个词义，以及所有目标词的词义分布情况，因此它的结果往往和现有的词典定义相差很大。本文提出的词义区分评价方法，可以依据词典的不同词义粒度层面评价词义区分的结果。

我们的评价目标是将所有结果类（Clusters）映射到现有的语义词典定义上，判断结果类与词典义的对应情况。假设词义区分的结果为$\bigcup_{i=1}^{M} C_i$，所有 N 个目标词被分到了 M 个结果类中（式 8）。每个目标词 w 在词典中会有一个或多个定义，用函数 S 表示（式 9）。建立结果类 C 中所有目标词与它们对应的词典义的逆映射，即每个词典义 T 通过映射 W 对

应到一组含有它词义的目标词（式 10）。

$$C=\{w_k \mid 1\le k\le n\} \tag{8}$$

$$S(w)=\{T_k \mid 1\le k\le m\} \tag{9}$$

$$W_C(T)=\{w_k \mid T\in S(w_k),\ w_k\in C, 1\le k\le |C|\} \tag{10}$$

最后，我们将含有的目标词个数最多的词典义判定为类 C 的词典义（式 11）。

$$\text{Sense}(C)=\begin{cases} T_i & \text{if } |W_C(T_i)|>|W_C(T_j)|(i\neq j) \\ \text{NULL} & \text{otherwise} \end{cases} \tag{11}$$

由此，我们定义准确率和召回率，见式（12）、式（13）。

$$\text{准确率}=\frac{\text{能对应到词典义的结果类个数}}{\text{结果类的总数}} \tag{12}$$

$$\text{召回率}=\frac{\text{结果类对应到的词典义的个数}}{\text{所有的词典义个数}} \tag{13}$$

4.3 实验结果与分析

本文使用哈工大信息检索研究室提供的同义词词林(扩展版)（后简称词林）作为评价标准。该词典是一部类义词典，具有一个 5 级词义分类体系，共分 12 个大类，97 个中类和 1400 个小类。每个小类还根据词义的远近和相关性分成了若干词群。词群下还分若干原子词群。其中，大类是第一级，原子词群是第五级。级别越高，词义刻画越细。

我们组合不同类型的特征，作对比实验。每个特征组合说明如表 1。

表 1 不同特征组合标识说明

特征组合标识	具体内容
BD	bigram 搭配窗口下的上下文分布特征
BDM	bigram 搭配窗口下的上下文分布特征 + 词形特征
BDN	bigram 搭配窗口下的上下文分布特征 + 主体名词特征
BDMN	bigram 搭配窗口下的上下文分布特征 + 词形特征 + 主体名词特征
SD	句子范围窗口下的上下文分布特征
SDM	句子范围窗口下的上下文分布特征 + 词形特征
SDN	句子范围窗口下的上下文分布特征 + 主体名词特征
SDMN	句子范围窗口下的上下文分布特征 + 词形特征 + 主体名词特征

根据不同的特征组合，我们得到了各自的词义区分结果。例如，BDMN 特征组合下的词义区分结果中前 6 个词语类为：

C1[宝贵 珍贵 平凡]　　C4[广阔 开阔 辽阔]

C2[长 短 漫长 隐蔽 直观]　　C5[严格 严肃 过硬 严谨 苛刻]

C3[平静 冷静 安静 从容]　　C6[红火 响 清淡 兴旺]

由于词林是一个语义分级结构，因此我们分别在它的中类和小类上评价我们的词义区分结果。表 2 和表 3 分别为不同特征组合在词林中类和小类上的准确率和召回率。

表 2 不同特征组合相对于词林中类的准确率和召回率

	BD	BDM	BDN	BDMN	SD	SDM	SDN	SDMN
Precision	0.5613	0.5805	0.5869	**0.6022**	0.5200	0.5634	0.5975	0.6320
Recall	0.2432	0.3514	0.3514	**0.4595**	0.2703	0.2973	0.2973	0.3243

表 3 不同特征组合相对于词林小类的准确率和召回率

	BD	BDM	BDN	BDMN	SD	SDM	SDN	SDMN
Precision	0.2642	0.3408	0.2934	**0.3569**	0.2800	**0.3396**	0.2924	0.3383
Recall	0.1620	0.2500	0.2007	**0.2676**	0.1585	0.2148	0.1690	**0.2324**

从结果中我们看到，混合了主体名词特征和词形特征的特征组合，在准确率和召回率上比没有使用这些特征的结果高出很多。这说明我们抽取的特征的有效性。另外，中类上评价的结果比小类上的评价结果好很多。这主要因为，一方面，词义区分方法会把许多只是语义相关但不是语义相似的词语放在一个类中，例如“平凡”和“宝贵”，它们是反义词而不是同义词。它们在词林中都属于 Ed 大类，但是它们所属的小类并不相同。比较不同搭配窗口的结果，使用 bigram 搭配窗口特征在大多数情况下比句子范围窗口特征效果要好。这主要是由于句子范围窗口带入了更多的噪音，而大部分的搭配词都集中在目标词的附近。当然，我们选用的特征都有各自的缺陷，即使混合起来使用，还是不可避免地带来了许多错误信息。

5 总结和展望

本文提出了一种基于混合特征的词义区分方法。我们从大规模语料库以及大量网页中抽取出形容词的上下文分布信息、主体名词信息以及词形信息这些易于获取的信息，利用 EM 迭代算法实现汉语高频形容词的词义区分。实验证明，我们选用的特征对于汉语形容词的词义区分是有效的。今后，我们将考察其他词性词语的词义区分问题，考察更多的特征选择组合，并且将词义区分结果应用到语义词典的编纂中去。另外，我们还试图通过词义区分发现不同词性词语之间的语义关联。

参考文献

[1] Navigli R. Meaningful Clustering of Senses Helps Boost Word Sense Disambiguation Performance [A]. Proceedings of the 21st International Conference on Computational Linguistics and the 44th annual meeting of the Association for Computational Linguistics, COLING-ACL 2006, 105-112.

[2] 朱虹，刘扬. 词汇语义知识库的研究现状和发展趋势 [J]. 情报学报. 待刊。

[3] Hsieh SK, and Huang CR. When Conset meets Synset: A Preliminary Survey of an Ontological Lexical Resource based on Chinese Characters [A]. Proceedings of the COLING/ACL on Main conference poster sessions, Sydney, Australia. 2006.

[4] Pantel P, and Lin DK. Discovering Word Senses from Text [C]. Proceedings of the eighth ACM SIGKDD international conference on Knowledge discovery and data mining. Pages 613-619. Edmonton, Canada. 2002.