词义归纳综述

孙玉霞[1,2] 曲维光[1,2,3] 狄 颖[1,2] 周俊生[1,2]
(南京师范大学计算机科学与技术学院 南京 210023)[1]
(江苏省信息安全保密技术工程研究中心 南京 210023)[2]
(南京大学计算机软件新技术国家重点实验室 南京 210023)[3]

摘 要 对于很多自然语言处理任务,如机器翻译、信息检索等,使用词义来进行相关表征其效果要比单纯使用词语的好得多。由于词义消歧需要大量标注语料、存在词义缺失等问题,词义归纳受到越来越多的关注。介绍了近年来词义归纳的一些相关工作和发展,并从词义归纳概述、相关技术、评估方法这3个方面进行了详述,最后对词义归纳工作进行了总结和展望。

关键词 词义归纳,向量空间,图,评估方法
中图法分类号 TP391 **文献标识码** A

Review of Word Sense Induction

SUN Yu-xia[1,2] QU Wei-guang[1,2,3] DI Ying[1,2] ZHOU Jun-sheng[1,2]
(School of Computer Science and Technology, Nanjing Normal University, Nanjing 210023, China)[1]
(Jiangsu Research Center of Information Security & Privacy Technology, Nanjing 210023, China)[2]
(State Key Lab. for Novel Software Technology, Nanjing University, Nanjing 210023, China)[3]

Abstract For many natural language processing tasks, such as machine translation, information retrieval, using word sense other than word itself as feature can perform much better. However, word sense disambiguation requires a large number of marked corpuses, at the same time, there are some problems hindering its application, for example the absence of some word senses. Therefore, people pay more attention to word sense induction. This paper introduced the related works and development of WSI from three aspects, which include the introduction of WSI, the related WSI methods and the evaluations. At last, we summaried and outlooked these works.

Keywords Word sense induction, Feature space, Graph, Evaluation

随着互联网的普及,网络成为人们日常交流的重要工具。由于网络的便利性,其信息传播具有传播速度快、流传范围广等特点,因此大量网络语言以星火燎原之势进入人们的生活。网络语言使得词语的用法更加多样化,如"稀饭"在网络语中表达的是喜欢的含义,而不是传统意义上的粥。同时由于新事物的出现,各种新生词语也不断出现,如"职粉",这是职业粉丝的简称,他们专门为参选艺人拉票、搞活动、策划形象等。但是由于信息呈现爆炸式增长,仅靠人工方法难以应付网络上层出不穷的词语新生用法,因此迫切需要计算机来帮助语言工作者从海量语料中自动获取词语的新生用法。而词义归纳就是其中一种重要的技术,它能够从大量语料中自动获取多义词在该语料中的词义,从而协助语言工作者进行进一步的工作。

本文第1节对词义归纳进行了概述,包括词义归纳的定义、分类、研究现状、研究意义等;第2节详细阐述了词义归纳的相关技术;第3节详细介绍了词义归纳的评估方法以及一些相关的评估工作;最后,对词义归纳工作进行了总结和展望。

1 词义归纳概述

1.1 词义归纳的定义

词义归纳(word sense induction, WSI)又称词义发现(word sense discovery)或词义区分(word sense discrimination),是从语料库中自动识别出多义词的词义,也就是给定多义词以及包含该词语的上下文集合,根据上下文所包含的信息,使用无监督的学习方法自动获取该多义词在语料中的词义用法。词义归纳认为,多义词所属上下文信息对其词义有表征作用。

到稿日期:2013-05-20 返修日期:2013-07-11 本文受国家自然科学基金(61272221),江苏省社科基金(12YYA002),国家社科基金(11CYY030,10CYY021)资助。

孙玉霞(1987—),女,硕士,主要研究方向为自然语言处理,E-mail:eve_hello@126.com;曲维光(1964—),男,博士,教授,博士生导师,主要研究方向为自然语言处理、计算语言学、语言工程、人工智能;狄 颖(1988—),女,硕士,主要研究方向为自然语言处理;周俊生(1972—),男,博士,副教授,硕士生导师,主要研究方向为自然语言处理、信息抽取、机器学习。

1.2 词义归纳方法的分类

Harris 提出分布假设，该假设认为具有相似词义的词语总是出现在相似的上下文中。例如，词语“苹果”，当出现在电子设备相关的上下文中时，一般代表苹果这个品牌，如：(1)苹果宣布 1 月 13 日起在中国大陆发售 iPhone。(2)苹果正在开发苹果电视机。而它若出现在食物相关上下文中，则代表苹果这种水果，如：(1)苹果又甜又脆，非常好吃。(2)苹果富含生物类黄酮，对人体好处多多。

词义归纳方法则是基于上述假设进行研究的。目前，词义归纳中用到的方法主要分为基于特征向量的方法和基于图的方法，但是随着机器学习方法在自然处理领域的流行，基于统计模型的方法在词义归纳的研究中也越来越深入。

基于特征向量的方法是使用特征向量对包含多义词的实例进行表征，然后使用聚类算法对特征向量集合进行聚类，获得多个簇。每个簇对应多义词的某一个词义，获得的簇的个数就是该多义词的词义数。

基于图的方法将待聚类的元素和特征在图空间中进行定义，然后使用基于图的聚类算法来进行词义归纳。依存树、共现图、相似度矩阵等都是图的不同结构，词义归纳中主要涉及到了共现图和相似度矩阵。当基于共现图时，目标多义词的共现特征作为结点，特征之间的共现性度量边权值。当基于相似度矩阵时，目标多义词的实例作为结点，实例之间的相似度度量边权值。

1.3 词义归纳研究现状

1.3.1 国外研究现状

国外词义归纳研究始于 20 世纪 90 年代末，大量的聚类方法运用到其中，发展较为成熟。在此同时，为了给词义归纳提供一个统一的评估平台，SemEval 组织了词义归纳任务。SemEval 原名为 Senseval，它是一个致力于词义消歧系统评测的国际组织，通过组织评测及相关活动来研究词义消歧技术。它由 ACL-SIGLEX 赞助的小型委员会管理。1997 年，Senseval 成立，并于 1998、2001 和 2004 年成功举办了 Senseval-1、2、3。之后由于 Senseval 中除词义消歧外有关语义分析的任务越来越多，Senseval 委员会决定把评测名改为 SemEval，并于 2007[41]、2010[40] 年举办了 SemEval-1、2，其包含了词义归纳任务，使得自然语言工作者能够更好地对词义归纳技术进行探讨和研究。

在国外的研究中，主要的词义归纳方法包括如下几种：

1. 基于特征向量的方法。这种方法通过特征选择来构建特征向量空间，然后使用聚类算法来进行词义归纳。其中，特征选择和聚类算法是关键。在特征选择方法中，共现词语、N-gram、词性是最基本的特征形式，Zhang[58] 等人按照一定比例对基本特征进行了组合。除此之外，Elshamy[33] 使用主题特征，Kern[18] 将短语结构和句法依存关系作为特征，Jurgen 使用随机索引词语空间模型[30] 来构建特征空间。在上下文表征方法中，有一阶上下文向量和二阶上下文向量两种表征方式，前者适合于大规模语料，后者适合于小规模语料。在聚类算法中，主要涉及到 K-means、层次聚类、Repeated Bisection[12]（简称 RB）等算法，并且使用了期望最大化（Expectation Maximization，简称 EM）迭代算法，其中 K-means 性能最优。为了获得更好的结果，Wang 等人[57] 对 K-means 算法进行了改进，提出了基于最大距离的初始中心点选择方法。在层次聚类算法中，涉及到的主要是凝聚型的，包括 UPGMA、McQuitty's Similarity Analysis[44]、The Sequential Information Bottleneck algorithm（简称 sIB）[50]，sIB 是凝聚型聚类算法 Information Bottleneck（IB）的一个变形，它是基于信息论的，能够保证收敛到局部最大信息。Pantel 和 Lin[51] 根据词义归纳特性提出了 Clustering By Committee（简称 CBC）算法。

2. 基于图的方法。这种方法将待聚类的元素和特征在图空间中进行定义，然后使用基于图的聚类算法来进行词义归纳。依存树、共现图、相似度矩阵等都是图的不同结构，词义归纳中主要涉及到了共现图和相似度矩阵。当基于共现图时，目标多义词的共现特征作为结点，特征之间的共现性度量边权值，结点的歧义性是关键，因此共现词语、搭配[1]、搭配和共现词语相结合[2]、三元组[16]、多元组、超边等均用来构造节点，以降低结点的歧义性。当基于相似度矩阵时，目标多义词的实例作为结点，实例之间的相似度度量边权值，Agirre[48] 和 Klapaftis[46] 等人分别对其进行了研究，并取得了较好的结果。常用的图聚类算法包括 Chinese Whisper（CW）[8]、Normalised MinCut、谱聚类[5]、Markov Clustering（简称 MCL）[33]、Hierarchical Random Graphs[22] 等，其中 CW 使用得最多，性能也较优。

3. 其他方法。随着机器学习方法在自然处理领域的流行，基于统计模型的方法在词义归纳的研究中也越来越深入。Brody[34] 等人首次将贝叶斯方法运用到词义归纳中，取得了较好的成绩。但是该系统需要词义个数作为先验，Yao 等人[36] 在 Brody 基础上提出了无参数的贝叶斯模型，使用分层狄利克雷过程（the Hierarchical Dirichlet Process，简称 HDP）[37] 来进行词义归纳。HDP 能够根据语料分布自动归纳出词义数，性能与 Brody 的系统相近。Apidianaki[32] 对基于双语语料的词义归纳进行了研究，将其放在机器翻译的背景下。

1.3.2 国内研究现状

相对于国外语言，特别是英语，中文的词义归纳研究起步比较晚，相关工作也比较少，它是近几年才发展起来的。中文有自己的特点，譬如，英文的基本单元是词语，具有较好的表义性，而中文的基本单元是汉字，表义性较差，因此适用于英文词义归纳的方法并不一定适用于中文。由于这些独有的特性，使得中文自然语言处理面临挑战和机遇。

在 CIPS（Chinese Information Processing Society of China）and SIGHAN 的赞助下，2010 年举办了第一届 CIPS-SIGHAN 中文处理联合会议（CLP2010）。CLP2010 包含多个自然语言处理任务，词义归纳也是其中一项。CLP2010 的举办有效地促进了中文词义归纳的发展，并为其提供了公共的平台，从而更好地带动了中文词义归纳的研究和探索。

在中文词义归纳研究中，使用最多的是基于特征向量的方法。在特征选择中，主要使用单个汉字、共现词语、N-gram、词性等作为特征，并对其进行组合。在聚类算法中，涉及到了 K-means、层次聚类、Locally Adaptive Clustering（简称 LAC）等，并使用了 Expectation Maximization（简称 EM）迭代算法，其中 K-means 效果相对较好，层次聚类获得的簇较为不均衡。Zhang 等人[58] 使用单个汉字、词语、二元组作为特征，并对其按比例组合，进行特征构建，使用 K-means、Expec-

tation Maximization(简称 EM)、Locally Adaptive Clustering(简称 LAC)算法进行词义归纳,同时对各个结果进行组合,该系统在 CLP2010 的词义归纳评估中获得了较好的结果,F-score 值位列第一。Liu[29]对每个实例中目标词前后两个窗口内的词语进行两两组合,与目标词相结合构成三元组。每个三元组利用搜索引擎来获取共现词语,从而构建特征向量空间。

在中文词义归纳中,基于图的方法使用较少,主要是谱聚类、Chinese Wispher(简称 CW)算法。Xu 等人[54]使用除停用词外的所有词语来构建特征向量,基于该特征空间构造相似度矩阵,并使用谱聚类、k-means、层次聚类进行词义归纳,谱聚类的结果优于 K-means 和层次聚类。Zhang 等人[55]将目标词所在实例作为结点,实例之间的相似度作为边权值,使用 CW 算法进行词义归纳,但是由于评估中使用到的语料较小,不能反映实例间的固有关系,并且也不能利用簇数信息,因此该系统结果较差。

综上所述,国内有关词义归纳的研究取得了初步的成果,词义归纳研究的重要性也越来越突出。然而,在国内由于没有大规模的语料,无法体现基于图的优越性,因此词义研究的进展缓慢,较多地停留在基于特征向量的方法上。由于中文词义归纳的应用前景是非常广阔的,因此,我们需要对其进行深入的探索与研究。

1.4 词义归纳研究的意义

词义消歧(word sense disambiguation,简称 WSD)是给定多义词的词义列表,从包含该多义词的文本中识别出它对应的词义,是有监督的学习。但是有监督的词义消歧存在一些不足:

1. 需要大量的已进行了词义标注的语料来进行训练,并且词义消歧的质量依赖于训练语料的规模。然而标注语料又较难获得。

2. 词义消歧是根据字典或者其他语言资源来进行词义标注的,这些资源只包含了常用词义,对于领域相关或者新出现的词语用法,存在词义缺失问题。

3. 不同的语料对于多义词的词义颗粒度是不一样的,词义消歧只能使用统一的词义列表,而无法根据具体语料获得适合的词义颗粒度。

词义归纳则克服了词义消歧的这些缺点。它自动从未标注语料中获取多义词的词义,是无监督学习,无需利用已标注的语料来进行训练。而且多义词的词义是从语料中自动获取的,能够根据具体语料自动获取词义并调节词义颗粒度。因此,词义归纳成为当今计算语言研究中一个最重要的课题。

由于互联网的出现,信息更新越来越快,词语用法不断翻新,获取较全面的词语用法成为语言工作者的必经之路。词义归纳能够自动地从语料中获取全面、准确的词义,从而大大提高语言工作者的工作效率,譬如,协助字典编纂者进行工作,自动构造基于语料的词义分类或者对已存在的词义分类进行调整。词义归纳也能够有效带动机器翻译、信息检索、主题内容分析、文本处理、语音处理、文语转换等自然语言领域的发展,是自然语言处理中不可缺少的一个环节。

总之,词义归纳是计算语言学和自然语言处理领域的基础研究课题,提高词义归纳研究水平,对自然语言中的众多研究领域都会有重要的推动作用。

2 词义归纳的相关技术

2.1 基于特征向量的词义归纳

基于特征向量的词义归纳方法,首先从语料中进行特征提取,构建特征向量对上下文进行表征,然后使用聚类算法来进行词义归纳。

在特征选择方法中,主要涉及到共现词语、二元组、词性、主题特征、句法信息、短语结构信息等,点互信息、Fisher's left sided test[28]等用于特征过滤。上下文表征方法,主要有一阶上下文向量和二阶上下文向量两种表征方式。一阶上下文向量是使用多义词所在实例中的共现特征来表示多义词的实例;而在二阶上下文向量中,则是使用跟多义词共现的词语所对应的共现特征来表征多义词的实例。也就是说,多义词实例中的每个词语都分别使用一个共现词向量来表示,而多义词的实例则是通过对该实例中的所有词语的共现向量求平均来表示。在聚类算法中,主要涉及到 K-means、层次聚类、CBC、sIB、RB 算法等。

前期工作主要有 Pedersen 和 Bruce[10]、Schütze[11]、Purandare[44]、Purandare 和 Pedersen[13],这些工作使用共现词语作为主要特征。Schütze 系统中的共现向量是从一个独立的大规模训练语料中统计出来的,因而特征规模较大,使用 SVD 方法进行降维。Purandare[44]对 Pedersen、Bruce 的工作和 Schütze 的工作进行了结合。他使用 Pedersen 和 Bruce 工作中涉及到的简单匹配系数相似度计算方法和 McQuitty's Similarity Analysis 聚类算法,该聚类算法性能较优。同时也结合了 Schütze 的一些工作,譬如使用独立的训练语料来获取目标词上下文的共现特征,并使用二阶共现向量。除此之外,Purandare 采用了多种类型的特征,如一元特征、二元特征以及共现词语,同时添加了余弦相似度计算方法。在词义归纳任务中,余弦相似度和简单匹配系数两种方法性能相近。Purandare 和 Pedersen[13]系统地比较了无监督词义辨别技术[10,11,14,15],并对这些技术进行了扩展,构建了多种词义归纳系统。同时,使用大规模语料、小规模语料、词义种类数较多的混合语料来进行训练和测试。实验表明:在小规模语料和混合语料中,二阶上下文向量和 RB 相结合的性能较好,一阶上下文向量由于实例较短存在数据稀疏问题;在大规模语料中,一阶上下文向量和基于 UPGMA 的凝聚型层次聚类算法相结合的性能较好,因为大规模语料中获得的特征较多,在测试语料中容易匹配,所以一阶上下文向量性能更好。

与此同时,Purandare 和 Pedersen 等人使用 Perl 语言开发了 SenseClusters[21]工具,它是用于词义辨别的免费开源软件包,也可以用于邮件分类、命名实体识别等其他自然语言任务。SenseClusters 提供各种特征选择方法、各种文本表示方法、不同的聚类算法以及簇的评估方法。它调用 the Ngram Statistics Package[23]来进行特征提取,调用 SVDPACK[24]工具包来对特征空间进行降维,并提供了调用 CLUTO 的接口[22],这是一个包含多种聚类方法的聚类工具箱。Pedersen[25,26]使用 Senseclusters 参加了 2007 年和 2010 年的 SemEval,同时使用自动获取簇数技术 PK2 和 Adapted Gap Statistic,其中 PK2 获得的簇数与标准词义数相近。Niu 等人[17]使用的 Cluster Validation 方法也能获得较为准确的簇数。

Pinto[19]使用了自扩展技术来进行词义归纳。给定包含

多义词的语料，使用点互信息(Point-wise Mutual)对语料中包含的每个词语创建共现词集。点互信息如式(1)所示：

$$MI(w_1,w_2)=\log_2\frac{P(w_1,w_2)}{P(w_1)P(w_2)} \tag{1}$$

式中，$p(w_1,w_2)$是词语 w_1,w_2 共现的概率，$p(w_1)$和 $p(w_2)$是词语 w_1,w_2 出现的概率。然后，使用词语的共现词集来表征该词语，从而对每个目标实例构建特征向量。这种自动词项扩展技术能够有效提高特征规模，在各种评估方法中性能均较优。

上述基于特征向量的词义归纳方法在特征选择方面只是选择了简单的共现词语、共现搭配或词性信息。为了有效地利用上下文，更多的语言信息被加入到词义归纳的研究中，如主题特征、句法信息、短语结构信息等，从而获得更精准、有效的的上下文知识。

Elshamy[33]使用了主题特征。该文章使用潜在狄利克雷模型(Latent Dirichlet Allocation，简称 LDA)[35]获取主题特征。LDA 是基于离散数据集合的概率模型，分为 3 个等级的贝叶斯模型，如图 1 所示。

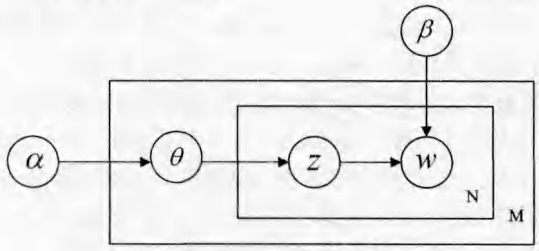

图 1　LDA 图模型

语料中包含 M 个文档，每个文档是 K 个主题的多项式分布，而这些主题分别是词语的多项式分布，使用该概率模型来生成文档 d。首先，使用带参数 α 的狄利克雷先验生成主题分布 θ_d。文档 d 包含 N_d 个词语，对于每一个词语 w_{dn}，使用参数 θ_d 从其多项式分布中提取出主题 z_{dn}。给定 β_{ij}，它是给定主题 j 选择词语 i 的概率，从该主题的词语分布中提取出词语 w_{dn}。从未标记的训练语料中获取潜在狄利克雷主题模型，从而获取每个实例的主题分布。在主题空间中，对这些实例的主题分布进行聚类。该方法在不借助任何辅助的语言工具情况下，能够挖掘潜在空间并取得了较好的性能。Kern 等人[18]将短语结构和句法依存关系作为特征。使用 the Stanford Parser 进行短语结构和句法依存关系的提取。然后利用提取出来的句法依存关系创建句法特征集，将提取出来的短语结构作为短语项特征集，并将具有相似语义的短语项进行合并。

Cruys[20]使用扩展的降维算法 NMF(Nonnegative Matrix Factorization)将词语数据和句法数据相结合，来发现潜在语义维度，根据这种潜在语义维度，可以对词语和句法关系进行分类。NMF 将矩阵 V 因式分解成两个矩阵 W 和 H，如式(2)所示：

$$V_{n\times m}\approx W_{n\times r}H_{r\times m} \tag{2}$$

式中，r 远远小于 n 和 m，从而使得实例和特征能够使用较少的组成部分(即潜在维度)来表征。该文对 NMF 进行扩展，将词语数据和句法数据相结合。首先，构造依赖关系-名词、名词共现词语-名词以及共现上下文词语-依赖关系 3 个共现频率矩阵，对其分别使用 NMF，并使用前一个因子来初始化下一个矩阵因子。具体过程如图 2 所示。

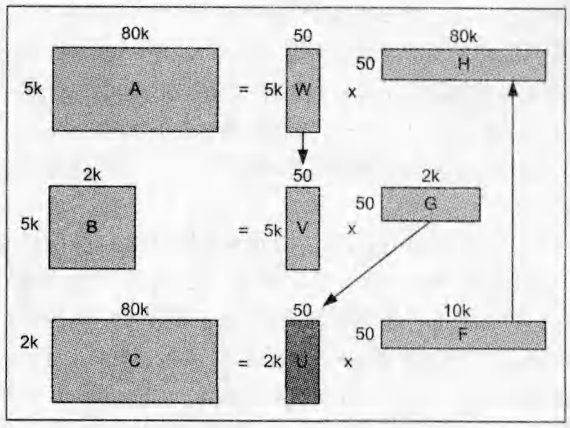

图 2　扩展的 NMF

其中，A,B,C 分别为 3 个共现频率矩阵。首先对 H 随机初始化，计算更新矩阵 W，使用 W 来初始化矩阵 V，计算更新 G，再使用 G 来初始化 U，计算更新 F，使用 F 来更新 H，不断迭代更新，直至收敛。扩展的 NMF 决定了哪些潜在语义维度对某个特定词义有决定性作用，然后使用这些维度对应的特征向量进行词义发现。并将上述方法嵌入到聚类算法 K-means 和 CBC(Clustering By Committee)[51]中来实现词义归纳的自动化。Apidianaki 和 Cruys[3]也使用了上述词义归纳方法。但是不同的是，他使用的是全局特征空间，全局特征空间指所使用的特征是从实验的所有多义词的语料中获取的，而不是从某个特定多义词语料中获取的。然后使用上述基于 NMF 的词义归纳方法将某个特定词语的多种词义通过与其他词语的词义进行对比、区分而提取出来。迄今为止的大部分词义归纳工作是局部地基于单个词语的，即每个多义词的词义归纳是单独进行的，多义词之间互不相干。

Jurgens 使用随机索引词语空间模型(Random Indexing，简称 RI)[30]来表征共现词语的高维共现空间。使用索引向量来表征一个词语。索引向量只包含 0 或 1，并且只有小部分数据为 1，数据稀疏，从而保证索引向量之间的正交性。并且每个词语对应的索引向量是固定的，维度也较小，如 5000。上下文使用多义词左右窗口中的共现词语对应的索引向量之和来表示。使用这种方法来表示上下文，词语维度较小，不需要使用降维技术。

由于中文词义归纳起步较晚，因此所涉及的技术相对比较单一。多数中文词义归纳使用的是基于特征向量的方法，特征一般选取共现词、二元组、词性等作为特征[54,31,27]，聚类算法主要涉及 K-means、层次聚类、sIB 等常见聚类算法以及 EM 迭代算法[58-59,31]。

Zhang 等人[55]使用同义词词林和 SVD 方法来解决数据稀疏问题。Wang 等人[57]对 K-means 算法进行了改进，提出了基于最大距离的初始中心点选择方法。与此同时，使用信息增益来决定特征窗口大小，并使用同义词林来解决数据稀疏问题。实验表明，该系统获得了较好的性能，并且改进的算法优于传统 K-means，F-score 提高大概 1 个百分点。

Zhang[9]等人将 K-means、EM 算法(Expectation Maximization)、局部自适应算法(Locally Adaptive Clustering，简称 LAC)的聚类结果进行集成，构成一致性矩阵 M，然后将其作为组平均凝聚型聚类算法(Group-average Agglomerative

Clustering，简称 GAAC)的输入，进行二次聚类。该文章采用一元模型和二元模型两种类型的特征，使用信息增益(Information Gain，简称 IG)来进行特征选择，并以 8∶2 的比例构成整个特征集合。然后，使用 EM、K-means、LAC 3 种算法分别对语料进行聚类。根据聚类结果，获取各个算法对应的邻接矩阵，并对其求均值，构成一致性矩阵 M，最后将其作为 GAAC 的输入，进行二次聚类。该实验在 CLP2010 提供的测试数据上进行了实验，对于单个算法，EM 获得 78.55%的 F-score 值，K-means 为 78.49%，LAC 为 78.95%，在所有参赛系统中均位居前四名。从中可以看出，上述特征选取方法能够有效地提取出有用特征。而最终的集成系统获得了 79.33%的 F-score 值，排名第一。聚类集成对多种聚类结果的一致性进行了调节，从而提供了更强健和稳定的解决方案，实验证明聚类集成的性能较佳。

2.2 基于图的词义归纳

基于图的方法将待聚类的元素和特征在图空间中进行定义，然后使用基于图的聚类算法进行词义归纳。在图空间中，结点可以是多义词的共现特征，也可以是某个实例，边权值使用结点之间的相似度度量。当结点是多义词的共现特征时，对共现图进行划分后，需根据获得的特征集合对包含多义词的实例进行词义辨别，得分最高的词义为该实例中多义词的词义。

基于共现图的词义归纳假设多义词的共现词语只跟该多义词的某一词义唯一相关。例如，“苹果又甜又脆，非常好吃。”，我们认为当“甜”、“脆”、“好吃”这些词语同“苹果”共现时，“苹果”是指水果，共现词唯一对应该词义。但是上述假设通常是不成立的，如“苹果宣布 1 月 13 日起在中国大陆发售 iPhone。”，这里共现词语“销售”并不唯一对应电子产品，“苹果”作为水果时，也会同“销售”共现。也就是说，共现词语也会存在歧义性，并不跟多义词某一个词义唯一相对，从而导致词义冲突。该问题是基于共现图的词义归纳的核心问题。

Agirre[48]使用了二阶上下文向量表示方法，使用基于图的方法和最小生成树获取共现词语对应的特征向量。该方法分成两个阶段。第一个阶段，首先构建共现图，并使用基于结点中心概念的两个算法 the HyperLex algorithm[7] 和 the HITS algorithm[52]来获取各个高密度子图的中心点。然后计算共现图的最小生成树，最小生成树的根节点就是获得的簇中心，将图中的每一个结点以特定距离归属到某一个簇中心。然后根据最小生成树对目标词的每个上下文赋予一个中心获得分向量，它是对上下文中词语对应的得分向量求平均。上下文中的每个词语对应的中心向量是其到所有中心结点的距离，其归属中心对应的得分为中心点和该词语结点之间的距离，其余中心点的得分均为 0。对于某个上下文，将其归类到对应向量中得分最高的中心点。此方法获得的簇较精细，因而簇数较多。但已知 SemEval-2007 给定的簇数不多，为了获得更好的结果，使用聚类算法对二阶上下文向量进行聚类。在第二阶段中，利用得分向量计算上下文之间的相似度，构建一个上下文相似矩阵，并对其进行剪枝，最后使用马尔科夫算法进行二次聚类。该系统在 SemEval-2007 中性能远远超过其他系统。

Klapaftis 等人[46]使用目标词的实例作为结点，Jaccard 系数度量结点之间的相似度。他指出图经常呈现出层次结构而不是简单平面划分结构，提出了基于层次随机图算法(hierarchical random graph algorithm)[47]，该方法创建一个二叉树，可以反映不同层次上(也就是不同词义粒度上)的词义归纳情况。

当结点是多义词的共现特征时，结点的歧义性是研究的重点。Dorow 和 Widdows[1] 等人假定一个共现词语唯一对应多义词的一种词义，同时因为名词相对于形容词、动词等，歧义性较小，使用共现名词作为结点。如果两个结点对应的名词共现次数超过给定阈值，则使用边将其连接起来，搭配的共现次数作为对应权重。由于马尔科夫聚类算法性能依赖于间隔因子等参数，因此该文章通过对相似共现词语的局部图聚类来迭代获取词义聚类，克服了参数依赖的问题。该文章列出了一部分的聚类结果，其能够较好地与 WordNet 中的词义相匹配。

然而，单个词语作为结点容易产生词义冲突问题，因此使用搭配、超边、或者更多词语构成的组合来唯一表征多义词词义，这对产生的数据稀疏问题提供了解决方法。Klapaftis 和 Manandhar[4]提出基于超图的模型。该模型使用超边对两个或更多的共现词语进行建模，认为这些共现词语构成的组合能够有效表达某一概念或意义。超边比传统图中的边更具有代表性，因为它能够有效地获取共现词语之间共享的信息。结点对应一个词语，超边则对应一组相关的共现词语，这里限定词语的个数为 2，3，4，并给每个超边赋予权值，如图 3 所示。

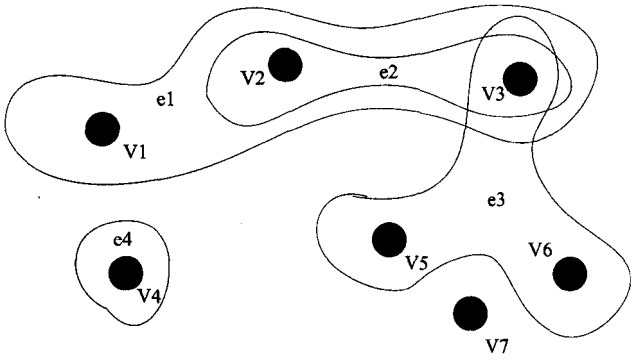

图 3 超图模型

其中，v 为结点，e 为超边。然后，使用修改的 HyperLex[7]算法来查找图中的中心结点，核心思想是获取超图中的高密度子图，这里也就是查找根结点，根结点选择的标准是包含该结点的超边数最多。然后将包含该结点的超边自动聚为一类，并从超图中分离出来。

Manandha 和 Klapaftis[38]假设两个词语构成的搭配唯一对应多义词的一个词义，由于词语搭配的词义唯一性要优于单个词语，因此能有效减缓词义冲突问题。该方法使用搭配来作为结点，边的权重使用对应结点的条件概率来进行赋值。但是这种方法构成的图是稀疏的，该文章认为互相似的两个结点应该拥有相同的邻结点，因而通过寻找图中的互相似点来获取更多的边，从而实现搭配图的平滑。最后，使用 CW (Chinese Whispers)[8]来对搭配图进行聚类，它是对基于代理的社会网络的仿真，在各种评估方法下，性能均较优，获得的词义冲突比传统的基于图的方法要小。该系统基于 SemEval-2007 提供的语料进行词义归纳，其结果如表 1 所列。

表 1　基于 SemEval-2007 语料的实验结果

Sytem	Unsupervised Evaluation				Sup.
	FSc.	Pur.	Ent.	#Cl.	Recall
UBC-AS	80.8	83.6	43.5	1.6	80.7
Col-JC	78.0	88.6	31.0	5.9	86.4
UOY	65.8	89.8	25.5	11.3	81.6

其中，Col-JC 是该文章构造的系统结果，UBC-AS 是 Agirre[48] 提出的系统结果，UOY 则是 Klapaftis 和 Manandhar[4] 提出的超图模型系统结果。从表 1 中可以看出，Col-JC 系统在各个评估指标下均取得了优异的成绩，UBC-AS 系统获得的簇数与标准簇数最为相近，而 UOY 获得的簇数较多，但是均较为纯净。

使用搭配作为结点会导致一定的数据稀疏问题，为此 Korkontzelos 等人[6] 针对这一问题进行了改进：若共现词语存在歧义，则将该词构建搭配作为结点；否则，使用该词语作为结点。首先，对于片段中的所有名词，两两构建词对，使用基于参考语料的对数似然来进行初步过滤。对于每一个词对，统计其所出现的片段集合，如果其片段集合和它所包含的两个名词对应的片段集合不相似，则认为这两个名词有歧义，该词对作为结点。使用条件概率来度量边权值。最后，使用 CW 算法来进行词义归纳。该方法归纳出来的簇数比较多，特别是在名词的词义归纳中。图 4 给出了词义归纳系统工作流程的实例。

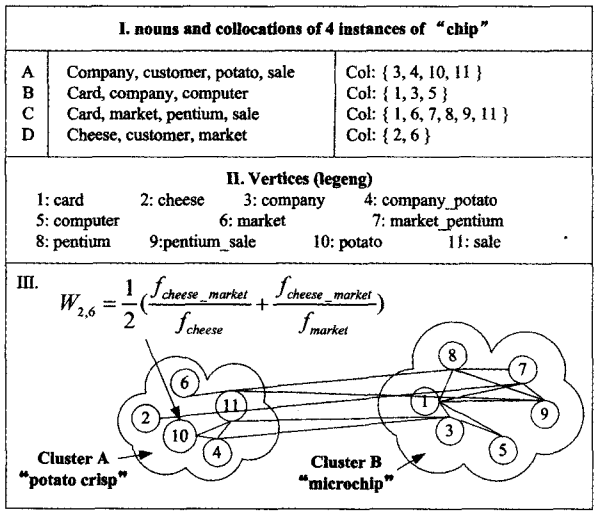

图 4　词义归纳系统工作流程实例

其中，PartI 左边给出了目标词语“chip”的 4 个片段对应的名词集合，PartII 则给出了获取到的所有结点，包含单字词结点和词对结点。其中，仅包含共现词语“company”的句子和包含“company”，“potato”的句子对应不同含义，“company”存在歧义，因此，“company_potato”作为一个结点。该方法在 SemEval-2010 中取得了较好的成绩，如表 2 所列。

表 2　基于 SemEval-2010 语料的实验结果

Sytem	VM	F-score	SRecall
UoY	15.7	49.8	62.4
KSU KDD	15.7	36.9	52.2
Duluth-WSI	9	41.1	60.5
KCDC-PCGD	7.8	59.2	59.5

其中，UoY 是该文章系统结果，其余均是基于特征向量的方法，KSU KDD 是 Elshamy[33] 提出的基于主题空间的系统，KCDC-PCGD 是 Kern[18] 使用短语结构和句法依存关系作为特征的词义归纳系统，Duluth-WSI 是 Pedersen[25] 基于词义辨别软件包 Senseclusters 的系统。从表 2 可以看出，基于无歧义结点的方法在各个评估中均取得了较好的成绩。

Jurgens[45] 提出了社区发现方法，该方法具有两个特性：一个词语可以属于多个重叠的社区，它是发现多个词义的关键，并且可以调节词义间隔。社区是指一组紧密相连的结点，假定它对每个词项都定义了特定词义相关的上下文。它是对边进行聚类，也就是对词语搭配进行聚类。对于共享一个结点的两条边，通过计算这两条边对应结点的相似度来判定两条边的相似度。基于这种相似度计算方法，使用单连接准则不断迭代，将拥有最高相似度边的两个簇进行合并，最终形成一个系统图。可以在不同层次对系统图进行分割，获取不同颗粒度的词义。最后，给定多义词上下文，计算上下文和归纳出来的社区之间相似度，取相似度最高的社区作为该上下文中多义词的词义。该方法是 Dorow[1] 等人和 Klapaftis[46] 等人的工作的结合，即局部区域探测和层次结构获取方法的结合。

在基于图的算法中，系统参数的选择对于其性能有着重要的影响，它一般是根据经验或者有监督技术进行设定，Ioannis[53] 等人提出了基于图连通评估方法的参数选择。对基于图的词义归纳方法，他们通过对词义归纳获得的簇进行图连通情况的度量来选择参数。该文章重点对 Klapaftis 和 Manandhar[38] 的基于搭配图的词义归纳方法进行了参数的选择，使用了 Average Degree、Average Weighted Degree、Average Cluster Coefficient、Average Weighted Cluster Coefficient、Graph Entropy、Weighted Graph Entropy、Edge Density、Weighted Edge Density 8 种图连通评估方法。给定目标词和一组参数，词义归纳得到的每个簇是原始图的一个子图，然后分别使用 8 种图连通评估方法对获得的各个子图进行打分，来评估其连通性。在每一种评估方法下，对所有子图得分求均值，将其作为该评估方法下这组参数的得分。给定多组参数，依次进行连通性评估，在每一种评估方法下，均选择得分最高的一组参数。结果表明，Average degree 和 Weighted average degree 是较好的两种评估方法，它们所获得的参数性能跟通过有监督评估而获取的参数性能相似。总之，图连通评估方法能够根据获得的子图的连通性，有效识地别出性能较高的参数。

在中文词义归纳中，图的方法涉及得比较少。Xu[56] 和 He[49] 等人均使用相似度矩阵作为聚类算法的输入，采用谱聚类方法进行聚类，获得了较好的结果。Zhang 等人[55] 将目标词所在实例作为结点，实例之间的相似度作为边权值，使用 CW 算法进行词义归纳，但是由于评估中使用到的语料较小，不能反映实例间的固有关系，并且也不能利用簇数信息，因此该系统结果较差。

2.3　其它词义归纳方法

Brody[34] 将多义词的上下文词语看成是多项式词义分布的采样，将词义归纳放到概率模型中，从而构建贝叶斯词义归纳模型。这里使用潜在的狄利克雷分配模型（Latent Dirichlet allocation，简称 LDA）[35] 来进行文本生成。同时，使用多种特征类型，如局部上下文词语特征、更大范围的上下文词语特征以及搭配、n 元模型、依赖关系以及词性等，并使用一种通用框架将其组合起来，如图 5 所示。

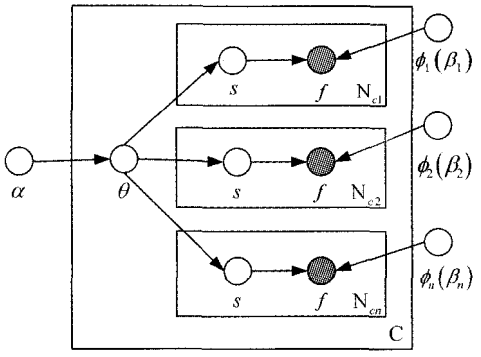

图 5 扩展的词义归纳模型

图 5 中内部矩形框代表不同的信息资源，所有的信息资源共享一个词义分布 θ，但是它们拥有各自的词义-特征多项式分布 ϕ，阴影结点代表观察到的特征 f，其可以是词语、词性、搭配或者依赖关系等。然后，使用 Gibbs 采样来获取每个实例作为某个词义的概率，选择概率最大的进行赋值。该方法性能好，而且适用性较广，也适用于文本分类等其他相关应用领域。但是 LDA 需要人工事先确定词义个数作为先验，因此 Yao 等人[36]在 Brody 基础上提出了无参数的贝叶斯模型，使用分层狄利克雷过程(the Hierarchical Dirichlet Process，简称 HDP)[37]来进行词义归纳。HDP 能够根据语料分布自动归纳出词义数，并且获得的词义归纳性能跟 Brody 的系统性能相近。

到目前为止，词义归纳方法大部分都是基于单语语料的，Apidianaki[32]对基于双语语料的词义归纳进行了研究，将其放在机器翻译的背景下。他使用了双语语料来构建双语词典，该语料分为源语料和目标语料，源语料是包含目标多义词 w 的语料，而目标语料则是源语料翻译成目标语言所构成的语料。对于多义词 w，在目标语料中获取其对应的等同体，也就是 w 在目标语料中的翻译。然后对 w 的每一个等同体获取源语言上下文，利用该上下文信息对目标词的所有等同体两两计算相似度，当相似度超过一定阈值，就将这两个等同体进行合并。将其聚类结果应用到词义消歧系统中取得了较好的效果，均远远超过 baseline 值。观察每个多义词训练语料中出现最多的等同体也是多义词对应最为频繁的翻译。

3 词义归纳评估

词义归纳是典型的聚类问题，但是由于很难全面地评估聚类结果的质量，因此选取合适的评估方法成为词义归纳的重点。可以从以下几个方面进行评估：

方法 1 人工查看每个词语对应的簇是否合适，但是这种方法很耗费人力，并且因为每个人的主观评判标准都不一致，所以比较难判断每个词语的实际用法是否与模型给出的用法一致。因而，在实际评测中并不对模型返回的簇是否合适进行主观判定，而是对语料进行黄金标准词义的标注，然后将其标注的词义和模型获得的簇进行比较。将黄金标准标注看成是类别定义，使用聚类领域的评估方法来评估聚类结果。

方法 2 将模型归纳出的词义自动映射到黄金标准词义上，并根据这种映射关系对语料进行黄金标准词义的标注，再将其标注结果和原先的人工标注结果进行比较，然后采用有监督的评估方法对映射质量进行评定。

方法 3 在某个具体应用中对其进行评估，例如词义归纳在机器翻译系统中对翻译性能的影响、在信息检索中对搜索性能的改善情况等等。这种想法很好，但是所需的系统开发也比较昂贵，较难实现。

在词义归纳的评估方法中，主要从前两个方面来进行评估，即无监督评估方法和有监督评估方法。有监督评估方法主要是使用精确率和召回率来进行评估，而无监督评估方法有 F-score 值、V-measure、Paired F-score、纯度和熵等方法。

3.1 无监督的评估方法

无监督的评估方法是将系统的聚类结果作为一个簇，将黄金标准作为类别。一个完美的聚类方法就是对每一个簇而言，存在一个类别，这个类别对应的实例集合和这个簇对应的实例集合是完全吻合的。常用的无监督评估方法有 F-score、V-measure、Paired F-score、纯度和熵等。

3.1.1 F-score

F-score 定义：给定类别 s_r，对应的实例个数为 n_r；给定簇 h_i，对应实例个数为 n_i；n_r^i 表示属于类别 s_r 和簇 h_i 的实例数。F-score 值越高，聚类效果越好。

精确率(Precision)：对于一个类别，某个簇分到该类别中的实例所占的比例。

$$P(s_r,h_i)=\frac{n_r^i}{n_r} \tag{3}$$

召回率(Recall)：对于一个簇，某一个类别分到该簇中的实例所占的比例。

$$R(s_r,h_i)=\frac{n_r^i}{n_i} \tag{4}$$

对于类别 s_r 和簇 h_i 的 F-score 定义见式(5)：

$$F(s_r,h_i)=\frac{2P(s_r,h_i)R(s_r,h_i)}{P(s_r,h_i)+R(s_r,h_i)} \tag{5}$$

对于某个类别 s_r 的 F-score 值定义见式(6)：

$$F(s_r)=\max_{h_i}F(s_r,h_i) \tag{6}$$

对于聚类整体的 F-Score 值定义见式(7)：

$$FScore=\sum_{r=1}^{c}\frac{n_r}{n}F(s_r) \tag{7}$$

其中，c 是类别数，n 是实例总数。

Klapaftis 和 Manandhar[4]指出一个好的聚类方法必须具备同质性和完整性。同质性是指对于聚成的每一个簇，都存在一个黄金标准类别，这个类别和该簇拥有相同的上下文集合；而完整性是指对于每一个黄金标准类别，都存在一个簇，这个簇和该类别拥有相同的上下文集合。传统的 F-score 方法同时评测了一个聚类方法的同质性和完整性，是对聚类方法的整体质量的评估。它偏向于获得簇数较少的系统。

但是 Manandhar 和 Klapaftis[38]指出 F-score 评估方法存在一些缺陷。对于某标准类别的 F-score 值，它是取跟该类别最匹配的簇对应的 F-score 值，也就是说它只能对与标准类别最匹配的簇进行评估，而不是对获得的所有簇进行评估。因此，无法对那些因为规模小等原因而不能与黄金标准词义相匹配的簇进行评估，并且当词义归纳系统使用较细的词义颗粒度时，由于获得的簇较多，会导致较低的 F-score。

3.1.2 V-measure[43]

V-measure 是完整性和同质性的调和平均数。给定完整性 c，同质性 h：

$$V\text{-}measure=\frac{2*h*c}{h+c} \tag{8}$$

给定簇集合 H、词义类别集合 S，a_{ij} 表示既属于类别 s_i 又属于簇 h_j 的实例数，N 是实例总数，$|S|$ 是词义类别数，

$|H|$是簇数。

完整性定义：$H(K|S)$是给定类别分布下簇的条件熵。当$H(K|S)=0$时，说明属于某个类别的实例只属于一个簇，此时该聚类方法有极好的完整性。

$$c=\begin{cases}1, & \text{if } H(K)=0\\ 1-\dfrac{H(K|S)}{H(K)}, & \text{otherwise}\end{cases} \tag{9}$$

$$H(K)=-\sum_{j=1}^{|K|}\frac{\sum_{i=1}^{|S|}a_{ij}}{N}\log\frac{\sum_{i=1}^{|S|}a_{ij}}{N} \tag{10}$$

$$H(K|S)=-\sum_{i=1}^{|S|}\sum_{j=1}^{|K|}\frac{a_{ij}}{N}\log\frac{a_{ij}}{\sum_{k=1}^{|K|}a_{ik}} \tag{11}$$

同质性定义：$H(S|K)$是给定簇下类别的条件熵。当$H(S|K)=0$时，每个簇所包含的实例只属于一个类别，具有很好的同质性。

$$h=\begin{cases}1, & \text{if } H(S)=0\\ 1-\dfrac{H(S|K)}{H(S)}, & \text{otherwise}\end{cases} \tag{12}$$

$$H(S)=-\sum_{i=1}^{|S|}\frac{\sum_{i=1}^{|K|}a_{ij}}{N}\log\frac{\sum_{i=1}^{|K|}a_{ij}}{N} \tag{13}$$

$$H(S|K)=-\sum_{j=1}^{|K|}\sum_{i=1}^{|S|}\frac{a_{ij}}{N}\log\frac{a_{ij}}{\sum_{j=1}^{|S|}a_{kj}} \tag{14}$$

V-measure对每个标准类别中各个簇的构成情况进行了考察，能够更全面地对整个聚类结果进行综合评定，并且对每个簇的完整性和同质性都进行了评估，而不像F-score依赖于事后的匹配，因此能够提供更客观的评估结果。它偏向于获得簇数多于标准词义数的系统。

3.1.3 Paired F-score

在该评估方法下，聚类问题转换成了分类问题。对于每一个簇c_i，生成$C_{|c_i|}^2$个实例对，这里$|c_i|$是指簇c_i所包含的实例数。对于每一个类别g_i，生成$C_{|g_i|}^2$个实例对，这里$|g_i|$是指类别g_i所包含的实例数。$F(K)$是簇所对应的实例对集合，$|F(K)|$是其实例对总数；$F(S)$是类别所对应的实例对集合，$|F(S)|$是其实例对总数。$|F(K)\cap F(S)|$是两个集合共有的实例对数。

精确度定义见式(15)：

$$P=\frac{|F(K)\cap F(S)|}{|F(K)|} \tag{15}$$

召回率定义见式(16)：

$$R=\frac{|F(K)\cap F(S)|}{|F(S)|} \tag{16}$$

Paired-score值定义见式(17)：

$$FS=\frac{2*P*R}{P+R} \tag{17}$$

在Paired-score评估方法中，簇数较多或者少于标准词义数的系统处于不利地位。

3.1.4 纯度[42]

纯度定义：给定簇h_i，它被映射到类别s_r上，类别s_r和簇h_i拥有的共同实例是最多的。然后计算进行这种赋值后得到的正确率。纯度越大，聚类效果越好。

$$Purity(S,H)=\frac{1}{N}\sum_{i=1}^{p}\max_k|h_i\cap s_k| \tag{18}$$

其中，$|h_i\cap s_k|$是指簇h_i对应的实例集合和类别s_k对应的实例集合的交集的大小。N是指实例总数。

纯度评估了方法的同质性，但是不能评估其完整性。它评估的是每个簇对应的最匹配类别和该簇的相交情况。因此对于簇数较多的系统，每个簇的规模较小，每个簇相对较纯净，从而使得其纯净度较高。当语料中的每个对象属于一个独立的簇时，纯度达到最大值1。因此纯度并不能很好地平衡聚类结果的质量和簇数。

3.1.5 熵[42]

熵的定义：熵考虑了在每一个簇中，各个类别的分布情况。熵越小，说明类别在簇中的分布越单一，聚类效果越好，给定簇h_i，它的熵定义见式(19)：

$$E(h_i)=-\frac{1}{\log q}\sum_{r=1}^{q}\frac{n_r^i}{n_r}\log\frac{n_r^i}{n_r} \tag{19}$$

其中，q为类别数，n_r^i是指既属于簇h_i又属于类别s_r的实例数。

对于聚类整体的熵，p为簇数，定义见式(20)：

$$E=\sum_{i=1}^{p}\frac{1}{p}E(h_i) \tag{20}$$

熵评估了方法的同质性，但是不能评估其完整性。不同于F-score评估方法，熵对于所有的簇都进行了评估，评估在某一个簇中各个类别的分布情况，分布越单一，则系统性能越好。同纯度评估方法一样，熵也偏向于簇数较多的系统。

3.2 有监督的评估方法

有监督的评估方法[41]将获得的簇映射到黄金标准词义上，并根据这种映射关系对测试语料进行黄金标准词义的标注，然后将其标注结果和原先的人工标注结果进行比较，即转换成词义消歧系统，从而使用有监督的评估方法，如精确率、召回率等。该方法需要额外的映射语料来获取映射关系。它很可能将多个簇映射到同一个黄金标准词义上，从而改变簇的分布。生成簇的个数过多会使得从簇到黄金标准词义类别的映射不可靠。在该评估方法下，获得不平衡簇分布的系统往往具有更好的性能。这里，不平衡簇分布是指：少量的几个同质性的簇包含了大部分的实例，而其他大部分簇只包含少数几个实例。

3.3 词义归纳评估相关工作

为了使得不同词义归纳系统之间能够互相比较，提供公共的实验语料以及设定相应的评测平台就显得尤为重要。但是先前的很多词义归纳工作都是使用各自的语料以及评估准则来进行评估，这使得系统之间的比较变得困难，不得不重复实验。Pantel和Lin[51]基于WordNet进行有监督的词义评估，使用Lin定义的相似度方法来进行词义映射，若簇数多于词义数，则部分簇数不进行映射，若词义数多于簇数，则部分词义不进行映射，然后使用F-score值进行评估。Purandare和Pedersen的SenseClusters使用混合矩阵来进行评估，这个矩阵描述了黄金标注词义在每个归纳出的簇中的分布，然后利用这个矩阵将黄金标准词义映射到某个簇中，从而使得正确率最高。正确率是在映射中正确标注的实例数除以所有实例数之和。

SemEval-2007[41]第一个对词义归纳任务创建了一个通用框架。选取了65个动词多义词和35个名词多义词，从华尔街日报中提取包含这些多义词的实例，并使用OntoNotes[3]词义对其进行标注。OntoNotes词义要比WordNet中的词义粗糙，因此词义数相对较少。评估方法包括无监督评估方法和有监督评估方法。这里无监督评估方法使用传统的聚类评估方法F-Score值。而有监督的评估方法首先将归纳出的词义映射到黄金标准词义上，并根据这种映射关系对测试语料进行黄金标准词义的标注，然后将其标注结果

和原先的人工标注结果进行比较，即这里转换成了词义消歧系统，使用精确值和召回率来对其进行评估。

SemEval-2010[39]在2007年评估框架的基础上提出了改进。这里包含50个名词多义词和50个动词多义词。评估方法还是包含两种：无监督评估方法和有监督评估方法。但是这里无监督评估方法使用的是V-Measure和paired F-Score两种评估标准。有监督评估方法和SemEval-2007 Task里面用到的一致，但是SemEval-2007提出的评估框架中，是在同一个语料上进行词义归纳和词义评估的，而SemEval-2010是将测试数据和训练数据分开，训练数据用于词义归纳，测试数据根据归纳出的词义进行词义标注并进行评估，测试数据是未知的新实例集合，从而保证了评估的可靠性和实用性，能够更好地对各种词义归纳模型进行比较。

CLP2010创建了第一个中文词义归纳任务的通用框架。训练语料中包含50个多义词，测试语料中包含100个多义词，每个多义词提供50个实例。所有实例都从网络和新华日报、人民日报等报纸中提取，使用HowNet词义进行人工标注。在进行实验时，多义词的词义个数是给定的。在词义评估上，使用F-score作为主要的评估方法，为了进行较全面的系统评估，同时使用了纯净度、熵以及V-measure评估方法。

在进行系统评估的时候，需要基准来进行对比。对于基准系统的设定，最常用的有两种方式：1. 将多义词的所有实例聚为一个簇(Most Frequent Sense，简称MFS)；2. 使用随机算法随机对所有实例进行聚类。然后使用评估方法对以上两种方式获得的结果进行评定，并将其作为评估的基准。

结束语 本文介绍了近年来词义归纳的一些相关工作，主要从词义归纳概述、相关技术、评估方法等4个方面较全面地介绍了词义归纳相关工作和发展。

词义归纳技术主要包含基于特征向量的方法、基于图的方法以及其他方法。基于特征向量是较为传统的方法，研究重点是特征选择和聚类算法。在特征选择中，主要涉及到共现词语、二元组、词性、主题特征、句法信息、短语结构信息等，点互信息等用于特征过滤，使用一阶上下文向量和二阶上下文向量进行实例表征。在聚类算法中，主要涉及到K-means、层次聚类、CBC、sIB、RB算法等。基于图的方法将待聚类的元素和特征在图空间中进行定义，然后使用基于图的聚类算法来进行词义归纳。依存树、共现图、相似度矩阵等都是图的不同结构，词义归纳中主要涉及到了共现图和相似度矩阵。当基于共现图时，目标多义词的共现特征作为结点，特征之间的共现性度量边权值，结点的歧义性是关键，因此共现词语、搭配、搭配和共现词语相结合、三元组、多元组、超边等均用来构造节点以降低结点的歧义性。当基于相似度矩阵时，目标多义词的实例作为结点，实例之间的相似度度量边权值。常用的图聚类算法包括CW、Normalised MinCut、谱聚类、MCL、Hierarchical Random Graphs等，其中CW使用得最多，性能也较优。除此之外，部分研究工作涉及到了统计模型，如贝叶斯模型以及无参数的贝叶斯模型等。

在词义归纳评估方法中，介绍了无监督和有监督的评估方法。其中无监督的评估方法包括F-score、V-measure、Paired F-score、纯度和熵。

尽管人们对词义归纳有了一定的研究和探索，但它是一个新兴的研究方向，在自然语言处理领域对其进行广泛研究只有十几年的时间，所以词义归纳研究中尚有许多值得深入探索的问题。本文基于词义归纳近年来的一些工作，提出一些值得进一步挖掘的研究点，希望对本领域研究者有所启发。

在词义归纳的评估方面，评估方法多种多样，如有监督的召回率、无监督的F-score值等，但这些方法存在一些缺陷，不能够较全面地评估一个系统的性能。将词义归纳应用到某个自然语言领域，如信息检索、机器翻译等，通过查看词义归纳给该领域带来的性能提升来判定词义归纳的效果，这种方法能够比较直观、有效、全面地对系统进行评测。因此，基于任务的词义归纳评估方法是词义归纳的一个重要研究方向。

对于词义归纳系统而言，选取一组合适的参数对于系统的性能也是至关重要的。一组好的参数能够有效提高系统的性能，故在选择好的词义归纳模型的同时，使用合适的参数优化和选择技术也能有效促进词义归纳的发展。其次，由于词义数的规模对于模型的词义归纳性能有着较为重要的影响，因此要实现词义归纳的自动化，必须实现词义数的自动获取。

对于中文词义归纳，目前使用的技术相对比较单一，主流方法是基于特征向量的，研究侧重于特征的选取和聚类算法的选择。中文的基本单位是汉字，但是表意的基本单位却是词语，而英文的基本单位是词语，同时表意的基本单位也是词语，所以进行中文词义归纳必须对中文的特性进行一定的研究，进而构造适合于中文的特征，才能有效改善性能。在此基础上，应不断引进国外一些先进技术，如基于统计模型的方法等。基于图的方法适用于规模较大的图空间，在国外的研究中取得了较好的成果，而目前中文词义归纳拥有的语料规模较小，从而导致图方法的结果并不理想，投入人力进行中文词义归纳语料扩展是当前的重中之重。

参考文献

[1] Dorow B, Widdows D. Discovering corpus-specific word senses [C]// Proceedings of the 10th conference of the European chapter of the ACL. 2003: 79-82

[2] van Dongen S. A cluster algorithm for graphs[R]. Technical Report INS-R0010. National Research Institute for Mathematics and Computer Science, 2000

[3] de Cruys T V, Apidianaki M. Latent Semantic Word Sense Induction and Disambiguation[C]// the Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics. 2011: 1476-1485

[4] Klapaftis I P, Manandhar S. UOY: A Hypergraph Model For Word Sense Induction & Disambiguation[C]// Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). 2007: 414-417

[5] Cai Xiao-yan, Dai Guan-zhong, Yang Li-bin. Survey on Spectral Clustering Alogrithms[J]. Computer Science, 2008, 35(7): 14-18

[6] Korkontzelos I, Manandhar S. UoY: Graphs of Unambiguous Vertices for Word Sense Induction and Disambiguation[C]// Proceeding of the 5th International Workshop on Semantic Evaluation. 2010: 355-358

[7] Veronis J. Hyperlex: lexical cartography for information retrieval[J]. Computer Speech & Language, 2004, 18(3): 223-252

[8] Biemann C. Chinese whispers-an effcient graph clustering algorithm and its application to natural language processing problems[C]// Proceedings of TextGraphs. 2006: 73-80

[9] Zhang Bi-chuan, Sun Jia-shen. Word Sense Induction using Cluster Ensemble[C]// The first CIPS-SIGHAN Joint Conference

on Chinese Language Processing(CLP2010). 2010

[10] Pedersen T, Bruce R. Distinguishing word senses in untagged text[C]// Proceedings of the Second Conference on Empirical Methods in Natural Language Processing. 1997:197-207

[11] Schütze H. Automatic word sense discrimination[J]. Computational Linguistics, 1998, 24(1):97-123

[12] Zhao Y, Karypis G. Evaluation of hierarchical clustering algorithms for document datasets[C]//Proceedings of the 11th Conference of Information and Knowledge Management (CIKM). 2002:515-524

[13] Purandare A, Pedersen T. Word sense discrimination by clustering contexts in vector and similarity spaces[C]//Proceedings of the CoNLL. 2004:41-48

[14] Pedersen T, Bruce R. Knowledge lean word sense disambiguation[C]//Proceedings of the Fifteenth National Conference on Artificial Intelligence. 1998:800-805

[15] Schütze H. Dimensions of meaning[C]//Proceedings of Super Computing. 1992:787-796

[16] Bordag S. Word Sense Induction: Triplet-Based Clustering and Automatic Evaluation[C]//Proceedings of the 11st Conference of the European Chapter of the Association for Computational Linguistics. 2006

[17] Niu Zheng-yu, Ji Dong-hong, Tan C-L. I2R: Three Systems for Word Sense Discrimination, Chinese Word Sense Disambiguation, and English Word Sense Disambiguation[C]// Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). 2007:177-182

[18] Kern R, Muhr M, Granitzer M. KCDC: Word Sense Induction by Using Grammatical Dependencies and Sentence Phrase Structure [C]//Proceedings of the 5th International Workshop on Semantic Evaluation. ACL, 2010:351-354

[19] Pinto D, Rosso P, Jimenez-Salazar H. UPV-SI: Word Sense Induction using Self Term Expansion[C]//Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). 2007:430-433

[20] Van de Cruys T. Using Three Way Data for Word Sense Discrimination[C]// Proceedings of the 22nd International Conference on Computational Linguistics (Coling 2008). 2008:929-936

[21] Purandare A, Pedersen T. SenseClusters-Finding Clusters that Represent Word Senses[M]. Department of Computer Science, University of Minnesota, 2007

[22] Karypis G. CLUTO-a clustering toolkit [R]. Technical Report 02-017. Department of Computer Science, University of Minnesota, 2002

[23] Banerjee S, Pedersen T. The design, implementation, and use of the Ngram Statistics Package[C]// Proceedings of the Fourth International Conference on Intelligent Text Processing and Computational Linguistics. 2003:370-381

[24] Berry M, Do T, O'Brien G, et al. SVDPACK (version 1.0) user's guide[R]. Technical Report CS-93-194. Computer Science Department, University of Tennessee at Knoxville, 2003

[25] Pedersen T. UMND2: SenseClusters Applied to the Sense Induction Task of SENSEVAL-4[C]// Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). 2007:394-397

[26] Pedersen T. Duluth-WSI: SenseClusters Applied to the Sense Induction Task of SemEval-2[C]//Proceedings of the 5th International Workshop on Semantic Evaluation. ACL, 2010:351-354

[27] Zhang Hao, Xiao Tong, Zhu Jing-bo. NEUNLPLab Chinese Word Sense Induction System for SIGHAN Bakeoff 2010[C]// The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010

[28] Pedersen T, Kayaalp M, Bruce R. Significant lexical relationships[C]// Proceedings of the Thirteenth National Conference on Artificial Intelligence. 1996:455-460

[29] Liu Zhao, Qiu Xi-peng, Huang Xuan-jing. Triplet-Based Chinese Word Sense Induction[C]// The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010

[30] Kanerva P, Kristoferson J, Holst A. Random indexing of text samples for latent semantic analysis[C]// Proceedings of the 22nd Annual Conference of the Cognitive Science Society. 2000: 1036-1040

[31] 蔡科,史晓东,陈毅东,等. 基于层次聚类的中文词义归纳[J]. 心智计算,2010,4(3):159-167

[32] Apidianaki M. Translation-oriented word sense induction based on parallel corpora[C]// Proceedings of the Sixth International Language Resources and Evaluation (LREC'08). 2008

[33] Elshamy W, Caragea D, Hsu W H. KSU KDD: Word Sense Induction by Clustering in Topic Space[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. ACL, 2010:367-370

[34] Brody S, Lapata M. Bayesian Word Sense Induction[C]// Proceedings of the 12th Conference of the European Chapter of the ACL. 2009:103-111

[35] Blei, David M, Ng A Y, et al. Latent dirichlet allocation[C]// Journal of Machine Learning Research. 2003:993-1022

[36] Yao Xu-chen, Van Durme B. Nonparametric Bayesian Word Sense Induction[C]// Proceedings of the TextGraphs-6 Workshop. 2011:10-14

[37] Teh Y W, Jordan M I, Beal M J, et al. Hierarchical Dirichlet Processes[J]. Journal of the American Statistical Association, 2006, 101(476):1566-1581

[38] Klapaftis I P, Manandhar S. Word Sense Induction Using Graphs of Collocations[C]//Proceedings of the 12th Conference of the European Chapter of the Association for Computational Linguistics. 2007

[39] Manandhar S, Klapaftis I P. Semeval-2010 Task 14: Evaluation Setting for Word Sense Induction & Disambiguation Systems [C]//DEW'09: Proceedings of the Workshop on Semantic Evaluations: Recent Achievements and Future Directions. 2009:117-122

[40] Manandhar S, Klapaftis I P, Dligach D, et al. SemEval-2010 Task 14: Word Sense Induction & Disambiguation[C]// Proceedings of the 5th International Workshop on Semantic Evaluation. ACL, 2010:63-68

[41] Agirre E, Soroa A. SemEval-2007 Task 02: Evaluating Word Sense Induction and Discrimination Systems[C]// Proceedings of SemEval-2007. 2007:7-12

[42] Zhao Ying, Karypis G, Fayyad U. Hierarchical clustering algorithms for document datasets[J]. Data Mining and Knowledge Discovery, 2005, 10(2):141-168

[43] Rosenberg A, Hirschberg J. V-measure: A Conditional Entropy-based External Cluster Evaluation Measure[C]//Proceedings of the 2007 EMNLP-CoNLL Joint Conference. 2007:410-420

[44] Purandare A. Discriminating among word senses using mcquitty's similarity analysis[C]// Proceedings of the HLT-NAACL 2003 Student Research Workshop. 2003:19-24

(下转第 48 页)

Sets and Systems,1997(90):111-127

[4] Dempster A. Upper and lower probabilities induced by a multi-valued mapping [J]. The Annals of Mathematical Statistics, 1967(38):325-339

[5] Shafer G. A Mathematical Theory of Evidence [M]. Princeton University Press,1976

[6] Smets P. The combination of evidence in the transferable belief models [J]. IEEE Transaction on Pattern Analysis and Machine Intelligence,1990,12(5):447-485

[7] Smets P,Kennes R. The transferable belief model [J]. Artificial Intelligence,1994,66(2):191-234

[8] Yager R R. On the Dempster-Shafer framework and new combination rules [J]. Information Sciences,1987,41(2):93-138

[9] Dubois D,Prade H. On the unicity of Dempster rule of combination[J]. International Journal of Intelligent System, 1986(1): 133-142

[10] Smsrandache F,Dezert J. Advances and application of DSmT for information fusion (Vol. 1) [M]. Rehoboth:American Research Press,2004

[11] Liu W R. Analyzing the degree of conflict among belief functions [J]. Artificial Intelligence,2006,5:909-924

[12] 孙全,叶秀清,顾伟康. 一种新的基于证据理论的合成公式[J]. 电子学报,2000,28(8):117-119

[13] 邓勇,王栋,李齐,等. 一种新的证据冲突分析法[J]. 控制理论与应用,2011,28(6):839-844

[14] 杨善林,朱卫东,任明仑. 基于可变参数优化的相关证据合成方法研究[J]. 管理科学学报,2003,6(5):12-16

[15] Pawlak Z. Rough sets [J]. International Journal of computer and Information Sciences,1982(11):341-365

[16] Pawlak Z. Rough Sets:Theoretical Aspects of Reasoning about Data,System Theory,Knowledge Engineering and Problem Solving[D]. Kluwer,Dordrecht,1991,9

[17] Qian Y H,Liang J Y,Yao Y Y,et al. MGRS:A multi-granulation rough set [J]. Information Sciences,2010(180):949-970

[18] Qian Y H,Liang J Y. Rough set method based on multi-granulations[C]//Proceedings of 5th IEEE Conference on Cognitive Informatics. 2006(I):297-304

[19] Yao Y Y,Lingras P J. Interpretations of belief functions in the theory of rough sets [J]. Information Sciences,1998,104(172): 81-106

[20] Wu W Z,Leung Y,Zhang W X. Connections between rough set theory and Dempster-Shafer theory of evidence [J]. International Journal of General Systems,2002,31(4):405-430

[21] Liang J Y, Li R, Qian Y H. Distance: A more comprehensible perspective for measures in rough set theory [J]. Knowledge-Based Systems,2012(27):126-136

[22] 杨风暴,王肖霞. D-S证据理论的冲突证据合成方法[M]. 北京:国防工业出版社,2011

[23] MacQueen J B. Some methods for classification and analysis of multivariate observations[C]//Proceeding of the 5th Berkeley Symposium on Mathematical Statistics and Probability. 1967: 281-297

(上接第 32 页)

[45] Jurgens D. Word Sense Induction by Community Detection[C]// Proceedings of the TextGraphs-6 Workshop. 2011:24-28

[46] Klapaftis I P,Manandhar S. Word Sense Induction & Disambiguation Using Hierarchical Random Graphs[C]//Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing. 2010:745-755

[47] Clauset A,Moore C,Newman M E J. Hierarchical Structure and the Prediction of Missing Links in Networks[J]. Nature,2008, 453(7191):98-101

[48] Agirre E,Soroa A. UBC-AS:A Graph Based Unsupervised System for Induction and Classification[C]//Proceedings of the 4th International Workshop on Semantic Evaluations (SemEval-2007). 2007:346-349

[49] He Zheng-yan, Song Yang, Wang Hou-feng. Applying Spectral Clustering for Chinese Word Sense Induction[M]. The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010

[50] Slonim,Friedman,Tishby. Unsupervised Document Classication Using Sequential Information Maximization[C]//Proceedings of the 25th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. 2002

[51] Pantel,Patrick,Lin De-kang. Discovering word senses from text [C]//Proceedings of the Eighth ACM SIGKDD International Conference on Knowledge Discovery and Data Mining. 2002: 613-619

[52] Kleinberg J M. Authoritative sources in a hyperlinked environment[J]. Journal of the ACM,1999,46(5):604-632

[53] Korkontzelos I, Klapaftis I, Manandhar S. Graph Connectivity Measures for Unsupervised Parameter Tuning of Graph-Based Sense Induction Systems[C]//Proceedings of the NAACL HLT Workshop on Unsupervised and Minimally Supervised Learning of Lexical Semantics. 2009:36-44

[54] Jia Yu-xiang,Yu Shi-wen,Chen Zheng-yan. Chinese Word Sense Induction with Basic Clustering Algorithms [C] // The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010

[55] Zhang Zhen-zhong, Sun Le , Li Wen-bo. ISCAS: A System for Chinese Word Sense Induction Based on K-means Algorithm[C]// The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010

[56] Xu Hua, Liu Bing, Qian Long-hua, et al. Soochow University: Description and Analysis of the Chinese Word Sense Induction System for CLP2010[C]// The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010

[57] Wang Li-sha,Dou Yan-zhao,Sun Xiao-ling,et al. K-means and Graph-based Approaches for Chinese Word Sense Induction Task[C]// The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010

[58] Jin Peng, Sun Rui, Zhang Yi-hao. A Knowledge based Method for Chinese Word Sense Induction[C]//Genetic and Evolutionary Computing(ICGEC). 2010:248-251

[59] Jin Peng, Zhang Yi-hao, Sun Rui. LSTC System for Chinese Word Sense Induction[C]//The first CIPS-SIGHAN Joint Conference on Chinese Language Processing(CLP2010). 2010