一种基于聚类的汉语词语知识的获取方法

李盛　杨尔弘

（山西大学计算机科学系，太原，030006）

E_mail：lisheng@sxu.edu.cn

摘要：在自然语言处理中，知识的自动获取一直是一个核心问题。但如何实现知识的自动获取呢？本文在基于实例的学习方法（Example-Based Learning，简称 EBL）的基础上，提出了一种基于聚类的汉语多义词知识的自动获取方法。实验结果证明，用该方法获得的知识对汉语的词义排歧是有效的。

关键词：聚类算法，基于实例学习，知识获取，多义词，上下文

1 引言

在自然语言处理中，知识获取占有十分重要的地位，如何实现知识的自动获取是当前自然语言处理过程中的一个热点课题。汉语词语词义知识的获取，尤其是多义词词义知识的获取，一直是中文信息处理中一个瓶颈问题。我们认为，词语的多义是静态中的多义，它一旦进入动态，在具体的语句中只能有一个意义。词语的上下文（context）是词语在实际应用中的语言环境，是知识获取的来源，也就是说上下文本身就是一个词语的知识集。本文就是以这个知识集作为研究对象的。

"物以类聚，人以群分"，聚类就是按照事物的某些属性、特征，把事物聚集成类，使类间的相似性尽量小，类内的相似性尽量大。由于聚类是一种无指导的学习方法，因此汉语词语的聚类已成为自然语言处理研究范围里的一个基础研究课题。根据词语的分布环境进行聚类是目前常采用的一种研究方法，在这方面作了卓有成效的工作的诸如 Brown et al. 1992，李涓子等 1997，王宁等 2000。其中王宁等以《现代汉语辞典》中的搭配实例对为原始数据，根据词之间的组合搭配关系计算词之间的距离，对汉语中名词和形容词的聚类进行了大量实验和研究，实验通过调整加权系数在一定程度上克服了数据稀疏问题。李涓子利用词在语料库中的分布信息，将具有相似分布的词通过聚类组合成一个类来构造词义空间，并运用到词义排歧中。由于自动聚类的结果需要人去对应每个词义向量所对应的词义，因此很难完成大规模的词义排歧的任务。

当前没有一种单一的知识获取方法是广泛有效的，需要采用结合的方法，如何有效结合使用各种方法和充分利用知识源是一个新的难点。本文提出的基于聚类的汉语多义词知识的获取，是以知网(HOWNET)提供的实例为基础，充分利用知网这部资源的特性，在已有的经过分词和词性标注之后的文本语料中挖掘出隐藏在文本中、决定多义词词义的上下文知识，并运用于汉语词语的词义排歧。

2 知网和 EBL 简介

2.1 知网及其预处理

知网是一个以汉语和英语的词语所代表的概念为描述对象、以揭示概念与概念之间以及概念所具有的属性之间的关系为基本内容的常识知识库，它是一种具有网状有机结构的知识系统，它提供了词之间的聚合关系，提供了一个较详细的词间联系的网络。它采用了不易分割的意义的最小单位——义原对所

有概念进行定义，从而使词具有很强的可计算性。

知网词典（２０００）共有词条116533个，属于一种英汉双语词典。为了便于语义计算，首先对知网进行预处理。知网词典是以文本方式保存的，我们将它转换为数据表的格式。由于我们是基于汉语词语的知识获取，所以为了减小存储量、提高性能，根据实际需要只提取NO（词条号），W_C（汉语词形），G_C（词性）以及DEF（定义）项，同时去掉了W_C与DEF项完全相同，只有W_E（英文词）项不同的词条，得到65953个词条。

其数据表记录的形式化描述如下：

<词条>：：=<NO 词条号> <W_C 词形> <G_C 词性> <DEF 义项>

在知网的E_C中保存的是为那些多义词提供的实例，这些实例强调对于多义词各个义项的区别能力而不是它们的释义能力，它们的用途在于为消除歧义提供可靠的帮助。我们以这些实例作为我们的种子实例。

2.2 EBL简介

Example-Based Learning 学习方法是一种与基于范例推理（Cased-Based Reasoning，CBR）的学习紧密相关的归纳学习方法。它们的共同推理假设为：世界从总体上看存在一定的规整性，相似条件下发生的动作会产生相似的结果。因此EBL的学习过程是一种基于相似性计算的实例的简单匹配。通过EBL学习可以学到相似分布下的词语知识。实例之间的相似性可以通过相似度计算公式来得到：

$$SIM_{ij} = 1 - d_{ij} \quad (当\ d_{ij} \in [0,1]时) \qquad (2.2.1)$$

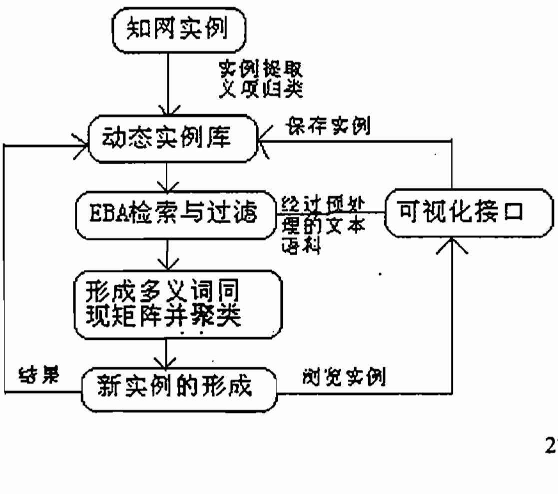
或为：$$SIM_{ij} = 1 / 1 + d_{ij} \quad (当\ d_{ij} \in [0,\infty]时) \qquad (2.2.2)$$

其中 d_{ij} 为两个实例间的距离，本文中我们把它定义为两个实例在知网中的语义距离，用欧氏距离公式来表示：

$$d_{ij} = \left[\sum_{k=1}^{N} (V_{ik} - V_{jk})\right]^{1/2} \qquad (2.2.3)$$

EBL的学习过程为：利用已有的实例，对新来的输入进行分类，根据相似性原理在已分类的实例中寻找与输入情况最相似的实例，然后把该实例所属类别作为对新实例的分类结果。

EBL学习方法在自然语言处理中具有如下优点：首先它突破了基于规则的学习方法的局限性，开发基于规则的知识系统时，规则或模型的获取是一件很繁琐的事；其次EBL学习方法简单，鲁棒性相对较好。



3 基于聚类和EBL的汉语多义词知识的获取

3.1 基于聚类和EBL的汉语多义词的知识获取

我们以HOWNET作为词典资源，以HOWNET提供的实例作为种子实例，来进行汉语多义词的知识获取。知识获取流程如上图：

各模块的功能如下：

●动态实例库的形成

动态实例库是用来存放能够区别多义词各个义项的搭配实例的。在这里我们以HOWNET提供的实例作为种子实例。实例选取原则是只选取那些对多义词义项具有表征意义的实词（名词、动词、形容词），同时对其标注词性，这样的好处是当该词在当前词性为单义词时，则不需计算处理，极大的减少了计算量。如"精神"一词作为名词时是多义词，在HOWNET中共有三个义项，当其作为"attribute|属性,strength|力量,&AnimalHuman|动物"之意时，HOWNET提供实例为"精神很足，看起来精神不好，振作精神，没有精神，精神焕发"，我们只选取"足a"、"振作v"，"焕发v"作为我们的实例，同时将其主义原加入义类集。动态实例库采用数据表的格式，其记录的形式化描述为：

<记录>:: =<NO 词条号> <W_C 词形> <G_C 词性> <DEF 义项><COWORD 搭配词集><COSENSE 义类集>

如上例在动态实例库中表示为 ：

表一 ：动态实例库中实例存储格式

023280	精神	N	Attribute|属性,strength|力量…	足a,振作v,焕发v	aValue|属性值,amount|多少..

其中义类集用来对实例进行相似度的计算。

在整个知识获取过程中，动态实例库处于不断的动态更新中，一方面利用它可以过滤掉语料中出现的已包含在实例库中的同现词，减少后面的计算量；另一方面通过不断获取新的实例，增加新的知识，从而扩大实例库。对于多义词不同词义间的共同实例，我们不加入实例库。这里我们假设当实例库在某一过程中处于相对稳定时，动态实例库已趋向稳态。

● 多义词同现矩阵的形成以及同现词的聚类

在汉语中大多数虚词的作用仅仅用于表示关联和语气，而实词间的语义搭配关系则对汉语理解起着重要的影响作用，因此我们只提取多义词当前词性下的同现实词来构成多义词同现矩阵，同时我们考虑到汉语中一些高频实词如"在"、"和"、"是"、"把"、"中"等的影响，会使得同现矩阵复杂化，因此我们不把它们加入同现矩阵。我们以两个标点符号("。"、"，"、"？"、"！"等)间的实词集作为多义词的上下文。

例句：各 rg 界 ngm 人士 ngd 踊跃 ad 支援 vgd 灾区 ngd， xp 慷慨 ad 解囊 vgd， xp 为 vgm 国 ngm 分忧 vgd， xp 表现 vgd 出 vgm 了 el 高尚 ad 的 ed 风格 ngd 和 cm 可贵 ad 的 ed 爱国 vgd **精神** ngd。

表二： 同现矩阵中数据的格式

表现	高尚	风格	可贵	爱国	精神
V	A	N	A	V	N

多义词同现矩阵形成以后，进行同现词的聚类。这里我们采用模糊聚类算法中的C-均值聚类算法进行聚类。C-均值聚类法是一种简便实用的无指导的学习算法，它可以用于已知类数的数据聚类。下面我们以"精神"一词为例来描述聚类过程。其步骤如下：

(1) 选取聚类块数k。其中k为多义词的义项个数，这里k=3。

(2) 将多义词的各个义项所对应的实例集分别作为聚类中心 c_1，c_2, c_3。

(3) 将每个样本向量 $x_i=[x_{i1},x_{i2},\cdots x_{in}]^T$(n 为向量的维数)按公式 2.2.3 归入中心为 c_i (i=1,2,3) 的类中。对于不能归入类的高频同现词通过可视化接口进行处理。

(4) 生成新实例，以新的实例集为聚类中心，重复以上过程。

(5) 当聚类中心不再变化时停止，否则重复以上过程。

● 新实例的生成

由于聚类所得搭配中包含了大量的噪音，使得学习结果并不可靠，因此需要对聚类结果进行加工处理。我们将聚类结果集表示为W()，同时利用动态实例库的义类集在HOWNET中产生一个义类词集H()，H中的词语是种子实例的同义词、近义词、对义词或反义词，取W与H的合取值，产生结果表示为E()，E中就是得到的新实例。

● 可视化接口

可视化接口，即人机交互界面，它主要有两个作用，一方面解决语料输入与结果输出；另一方面解决 W()中的没有被归类的高频词是否将作为多义词的实例加入实例库，我们的解决方法是通过人机交互界面浏览该高频词与多义词同现的上下文环境，人工决定其是否需要加入实例库并选定其所属的义项类。

3.2 试验结果及结果分析

我们取知网中的 10 个多义词在 52 兆新闻语料中进行学习，同时形成动态实例库。每个多义词都学到了新的实例，下面给出经过学习后的部分实例。

“精神”部分新的实例有：

attribute|属性,strength|力量,&AnimalHuman|动物：振奋 感动 感到 兴奋 抖擞 改观

mental|精神：文明 拥护 风貌 面貌 成果 力量 生活 影响 成效 安慰 需求 熏陶 价值

part|部件,%entity|实体,heart|心：认真 基本 爱国 高尚 求实 自觉 勤俭 腐败 腐朽 乐观 勤劳 浪漫 杰出 赤诚 严谨 奉献 弘扬 贯彻 学习 体现 全会 会议 协作 民族 主人翁 爱国主义

为了保证实例的可靠性，在学习过程中我们定义了同现度对新实例进行了测试与选取：

同现度：
$$M = flag \times \frac{Cnum}{Mnum} \times 100\% \qquad (3.2.1)$$

其中 Cnum 为新实例在语料中与多义词同现的频数，Mnum 为该多义词在语料中出现频数，flag 表示新实例是否与实例库中的义类集有语义关系，若有，flag=10，否则 flag=1。通过经验定义同现度 m=0.015，得到新实例。

从实验结果看出，本文提出的学习方法具有较强的学习能力，但通过对学习结果的分析发现存在着以下问题：

1. 种子实例规模太小，其所组成义类集明显比较狭隘，因此对于庞大的多义词同现矩阵来说，其获取存在着严重的数据稀疏。

表三：实例的学习情况表

多义词	训练实例个数	种子实例个数			训练结束实例个数		
		义项 1	义项 2	义项 3	义项 1	义项 2	义项 3
发生	2130	5	3		86	7	
精神	2044	7	5	5	9	18	77
材料	1840	4	8	2	6	15	57
健康	1848	6	5		17	41	

2. 义类层次的选择。对于集合 H()，如果义类层次给的太低，结果显得过于密集；如果给的太高，结果又显得过于稀疏。因此需要寻求一个更接近实际情况的义类层。

3. 局限于当前语料类型与规模。当多义词某一词义在语料中出现次数较少时，由这些较少的上下文同现词得到的实例不含有该类词的普遍的搭配分布信息。解决这一问题的方法是进一步扩大训练语料的规模或使用其它类型的训练语料以学习到更多的实例。

4. 因为采用的是一种自组织的学习方法，因此有些同现词可能错误的进入词义搭配实例中，成为词义的错误标识成分。

3.3 多义词词义排歧

为了测试学习结果，我们将所获得的实例应用到词义排歧中。由于排歧过程中我们将多义词所处的一个句子(可以是完整句也可以是短句，以“。”、“，”、“？”、“！”等标点符号作为分隔)作为上下文，从而可能会产生上下文同现词信息分布在多义词的不同义项的情况，因此采用了下面的排歧策略。

定义. 权值

设 (W1,W2,W3,…Wi)为一输入实词序列，DEFk 表示当前多义词的第 k 个义项，QWi 表示词 Wi 对于 DEFk 的权值,其取值分别为 0(Wi 与 DEFk 所对应实例集无任何语义关系),1((Wi 与 DEFk 所对应实例集有语义关系)，10((Wi 在 DEFk 所对应实例集中),Qdefk 表示义项 DEFk 的总权值，因此 Qdefk 可以表示为:

$$Qdefk = \sum_{i=1}^{I} QWi \qquad (3.3.1)$$

总原则：将权值最高的义项作为该多义词当前词义。

策略 1. 互信息的使用：我们这里将互信息定义为在一个句子中的两个多义词互为对方的搭配实例，如“精神作用”，“精神”与“作用”分别在对方的搭配实例库中，这种情况直接标注。

策略 2. 权值计算：计算任一个同现词对多义词每个义项的权值，从而得到各个义项的 Qdefk.

策略 3. 对于代词的处理：类似于“他们的精神”，“这些材料”等不能运用实例排歧的句子，取最近一句的词义或当前段落出现频率最高的词义进行标注。

测试过程分为封闭测试和开放测试两部分，在排歧过程中我们利用正确率来评测排歧结果。

$$正确率= \frac{正确排歧的次数}{多义词出现的总次数} \qquad (3.3.2)$$

表四：测试结果表

例词	封闭测试			开放测试		
	多义词出现次数	正确排歧次数	正确率 (%)	多义词出现次数	正确排歧次数	正确率 (%)
材料	886	844	95.2%	132	106	80.3%
发生	1205	1191	98.8%	266	232	87.2%
精神	1587	1541	97.1%	167	131	78.4%
健康	799.	763	95.5%	95	84	88.4%
平均正确率	96.7%			83.6%		

排歧结果分析：

在排歧过程中我们发现其中出现了不能标注或者标注错误的情况，主要原因为：

1. 在学习时，多义词的某个词义学习到的实例较少，从而使得上下文中同现的词并没有出现在动态实例库中或与实例的语义相似度很小以至为 0，使得排歧失败。
2. 动态实例库中的错误实例造成排歧错误。
3. 在待排歧文本中，若多义词某一词义出现很多次，而另一词义出现次数较少时，容易出现标注错误。如“发生”一词，作为“happen|发生”时出现 232 次，而作为“produce|制造”时出现 9 次，其中有 4 次被错误标注。
4. 语料的分词与词性标注错误使得排歧失败。如“代表”一词在语料中作为动词时是多义词，而往往被标注为名词。

4 结束语

本文提出了一种基于聚类的多义词词语知识的获取方法，这种方法不需要手工制定繁琐的规则，它主要通过聚类自动获取一个多义词在不同词义下的分布信息，然后通过动态实例库进行自组织学习，是一种无指导的学习方法。其中动态实例库的种子实例来源于 HOWNET，在学习过程中利用 HOWNET 可以很方便地进行基于词义的相似度的计算。

它的缺陷在于它在很大程度上受到训练语料的局限，训练语料的分词与词性标注的正确率直接影响

着知识获取的正确性；另一方面如何寻求更加高效的聚类算法仍是急需解决的一个问题。

参考文献：

[1] 史忠植著，《知识发现》，清华大学出版社，2002 年 1 月；

[2] 王宁、苑春法、黄昌宁，汉语名词和形容词的聚类算法研究，见黄昌宁、张普主编，《自然语言理解与机器翻译》，清华大学出版社，2001 年 7 月；

[3] 李涓子、姬东鸿、黄昌宁，基于组合实例的双向优化聚类，《语言工程》，清华大学出版社，1997；

[4] Julio Gonzalo，Irina Chugur and Felisa Verdejo，Sense Cluster for Information Retrieval：Evidence from Semcor and the EuroWordNet InterLingual index，ACL2000；

[5] Dekang Lin，Automati Retrieval and Clustering of Similar words；

[6] 董振东，知网，http://www.keenage.com

作者简介： 李盛，男，1977 年生，硕士研究生，主要研究方向为人工智能、中文信息处理。杨尔弘，女，1965 年生，副教授，硕士生导师，北京语言文化大学在读博士生，主要研究方向为计算语言学。

A Method for Acquisition of Knowledge About Chinese Words Based on Clustering

Li Sheng Yang Er-hong

(Dept. of Computer Science，Shanxi University Taiyuan 030006)

E-mail：lisheng@sxu.edu.cn

Abstract: Automatic acquisition of knowledge is always one of the greatest problem in NLP(natural language process). But how to achieve it? Based on EBL, we present a method for automatic acquisition of knowledge about Chinese ambiguous words based on clustering algorithm. The experiment result proves that the knowledge acquired is effective in tasks of WSD(word sense disambiguation).

Key words: clustering algorithm, Example Based Learning(EBL), acquisition of knowledge，ambiguous words, context