面向搭配知识库建设的汉语词义区分研究*

朱虹 刘扬 俞士汶

北京大学信息学院计算语言学研究所 北京 100871

北京大学计算语言学教育部重点实验室 北京 100871

E-mail: {zh,liuyang,yusw}@pku.edu.cn

摘 要：针对当前搭配资源在描述框架和知识获取方面的不足，本文将词义研究和搭配研究结合起来，设计并实现基于搭配的汉语词义区分方法，期望利用词语的搭配特征区分词语的词义，同时，自动获取可区分词义的相关搭配知识。评测结果表明，利用最小描述长度选取最优聚类结果获取的搭配词集具有明显的词义区分能力，在新的具有词义区分功能的搭配知识描述框架下，可广泛用于未来的大规模搭配知识库建设。

关键词：搭配；词义区分；层次聚类；最小描述长度

Chinese Word Sense Discrimination Facilitating Construction of Collocation Resources

Zhu Hong, Liu Yang, Yu Shiwen

Institute of Computational Linguistics, School of EE and CS, Peking University, Beijing 100871

Key Laboratory of Computational Linguistics (Peking University), Ministry of Education, CHINA, Beijing 100871

E-mail: {zh,liuyang,yusw}@pku.edu.cn

Abstract: This paper puts forth a collocation-based method for automatically discriminating Chinese word senses and acquiring the typical collocates for these senses. The method employs a dependency parser to extract collocation from real text, clusters all collocations by their context features, and then generates the best clustering results based on Minimum Description Length. The results show that our method can well acquire the discriminating collocations and facilitate construction of the large-scale collocation resources under a new representation framework.

Keywords: collocation; word sense discrimination; hierarchical clustering; Minimum Description Length

1 引言

搭配是自然语言处理的重要研究内容[1]，能够在很大程度上帮助自然语言的理解和处理。搭配相关的资源和技术已经被广泛应用于机器翻译、信息检索、文本内容理解等不同领域。汉语方面的搭配知识库主要包括《现代汉语实词搭配词典》、《现代汉语辞海》、《现代汉语搭配词典》等搭配词典和徐睿峰等人开发的搭配标注资源[2]。其中，搭配词典大都采用层次结构的搭配知识描述框架，即分层描述词语在不同词义、前后位置以及搭配结构下的搭配知识。例如《现代汉语搭配词典》对“紧张”的描述为：“[紧张] 兴奋不安：紧张…：~状态/~心理……”。徐睿峰等人开发的搭配标注资源则细化了搭配类型，并在真实文本中标注了 3,643 个中心词的 23,581 个搭配。不过，这些搭配知识库的设计和构建还存在一些亟待解决的问题。

首先，搭配词典的词义划分标准不明。例如“健康”在《现代汉语词典(第 5 版)》（以下简称《现汉》）中有两个词义，而《现代汉语辞海》将这两个词义合并到了一起，不区分这两个词义下的搭配内容。再如“高贵”在《现汉》中有三个词义，而《现代汉语辞海》将其中两个词义合并为一个词义，并且没有定义第三个词义。一般认为只有在搭配模式完全相同的情况下词义才

*本文承国家 973 课题（2004CB318102），国家自然基金项目（60775031），国家社科基金项目(08BYY060)和全国优秀博士学位论文作者专项资助项目（200514）的资助。

可以合并，但“健康”和“高贵”的不同词义在搭配内容上都有很大的差别，并不能简单地合并。

其次，“典型”搭配不“典型”，很难保证词义区分的功能。提供能够区分词义的搭配知识是搭配词典的功能之一。但是，目前搭配词典记录的一些搭配是没有词义区分功能的。例如修饰“健康”的副词“很、非常、十分”，修饰“健康”的形容词“基本、完全”，以及需要更多语境才能区分“健康”词义的搭配词“情况、状况、因素”等等。目前搭配词典并没有把它们单独区分开来，仍将它们当成典型搭配来记录。并且有的词典还只在某些定义下记录这些搭配，极易误导搭配词典的使用者。

再次，搭配词典和标注资源的实例很少，更新速度很慢，缺乏自动获取特定词义下典型搭配的方法。其中搭配词典的知识主要是通过语言学家的观察和总结。而搭配标注资源虽然标注了大量的搭配信息，但还不能与词语的词义信息相关联。

造成这些问题的原因有，一方面搭配的抽取方法不够完善，在句法分析技术不成熟的情况下，基于窗口的搭配抽取方法很难获取准确性高的搭配。另一方面，词义获取方法不成熟，在词义标注语料极少的情况下，很难自动获取真实文本中词语的词义，而词典中的词典定义与词语在真实文本中的词义表现存在很大的差别。另外，最大的问题是词义研究与搭配研究相脱节，导致我们很难大量获取词义与搭配之间的信息，不能满足搭配知识库构建和自动化词义研究的需要。

搭配知识库的不健全已经严重影响了词义消歧、词聚类等相关研究的开展[3][4]，必须将词义研究与搭配研究统一起来，研究如何自动挖掘能够反映词语词义的典型搭配，并且以符合实际应用需要的描述框架记录这些信息。因此文本提出一种基于搭配的词义区分方法，在 Yarowsky 提出的“一个搭配一个词义”（One Sense Per Collocation）的著名假设下[5]，自动从语料中抽取词语的搭配，利用层次聚类和最优聚类结果选择的方法，自动获取词语的不同词义和不同词义下的典型搭配。结合本文提出的新的搭配描述框架，更好地为搭配知识库构建提供支持。

2 算法设计

本文设计的基于搭配的词义区分算法主要分三个步骤，分别是搭配的抽取、搭配的聚类和最优聚类结果的确定。

首先，本文使用基于语法分析的搭配抽取方法抽取搭配。本文选用的语料是人民日报 2000 年一年的语料，包含 14,535,097 个词次，609,892 个句子。在对语料切分和词性标注的基础上，本文对所有句子进行依存句法分析[6]，得到具有依存关系的词语对。例如句子 S1 中构成依存关系的词语对包括（高雅；健康），（活动；健康），（活动；文化）等。

S1：高雅/a 健康/a 的/u 文化/n 活动/vn 使/v 一些/NUM 低级/a 庸俗/a 的/u 东西/n 没有/v 了/u 市场/n 。/w

本文通过依存句法分析提取出与目标词（用于词义区分实验的词语）构成依存关系的所有词语，作为目标词的候选搭配词。例如目标词“健康”的候选搭配词包括“高雅、活动”等词语。然后，本文通过计算候选搭配词与目标词的点间互信息，选出互信息最高的一些搭配词作为目标词的搭配词。本文将基于这些搭配词完成目标词的词义区分。

在搭配的聚类阶段，本文使用凝聚型层次聚类算法对搭配词进行聚类。本文使用搭配词自己的搭配词作为特征，并使用了卡方假设检验方法抽取这些特征。

在最优聚类结果的判定阶段，本文将 Li 和 Abe 的方法[7]引入到词义区分问题中，基于最小描述长度原则[8]完成最优聚类结果的判定。词义区分问题对应的是聚类参数未知的问题，需要从所有可能的数据划分结果中选取一个符合应用要求的结果。不同于一般的词聚类任务[9]，基于搭配的词义区分的聚类对象不是目标词本身，而是目标词的搭配词，并且需要同时考虑搭配词的特征以及搭配词与目标词之间的信息，以此反映目标词的词义特点。

Li 和 Abe 的方法将聚类模型描述长度和特定数据描述长度的总和作为统计指标。具体来说，模型指的是搭配词聚类形成的聚类结果，数据指的聚类结果对应的搭配词与目标词的共现信息。

首先本文用 S 表示目标词 T 的所有搭配实例。搭配实例是搭配词在目标词的上下文中的一次出现。S 的大小是所有搭配词与目标词的共现频次之和。然后本文将目标词的搭配词聚类形成一棵层次聚类树。本文的目标是找到一条切分线，使得切分线上所有结点形成的聚类结果Γ，它对应的模型 $\hat{M} = (\Gamma, \hat{\theta})$ 的描述长度和用这个模型来描述 S 的描述长度总和最小。本文称Γ为切分树。其中 $\hat{\theta}$ 是需要估计的参数。模型数据描述总长度 $L(\hat{M}, S)$ 可以表示成切分树的描述长度 $L(\Gamma)$，参数的描述长度 $L(\hat{\theta} \mid \Gamma)$，和数据描述长度 $L(S \mid \Gamma, \hat{\theta})$ 的总和，见公式 1。其中 $L(\Gamma) + L(\hat{\theta} \mid \Gamma)$ 就是模型的描述长度。参数描述长度的计算公式见公式 2。公式 2 中 k 表示切分树的自由变量个数，本文将 k 设定为切分树对应的聚类结果的类数。数据的描述长度计算方法见公式 3。公式 3 中 $\hat{P}(n) = \hat{P}(C) / |C|$，C 是切分线上的某个类。使用最大似然估计，$\hat{P}(C) = f(C) / |S|$。

$$L(\hat{M}, S) = L(\Gamma) + L(\hat{\theta} \mid \Gamma) + L(S \mid \Gamma, \hat{\theta}) \tag{1}$$

$$L(\hat{\theta} \mid \Gamma) = \frac{k}{2} \times \log |S| \tag{2}$$

$$L(S \mid \Gamma, \hat{\theta}) = -\sum_{n \in S} \log \hat{P}(n) \tag{3}$$

为了方便起见，本文给每个切分树赋予相同的先验概率，即 $L(\Gamma)$ 对于每个模型都是一样的。因此只需要计算 $L'(\hat{M},S)=L(\hat{\theta}|\Gamma)+L(S|\Gamma,\hat{\theta})$。本文将 $L'(\hat{M},S)$ 简记为 $L'(\Gamma)$。

对于一棵深度为 d 的完全 b 树而言，寻找最优切分树的搜索量级是 $O(2^{b^{d-1}})$。为了有效地找到最优切分树，Li 和 Abe 给出了一种动态规划算法。其核心算法如图 1。其中 t 表示聚类结果树中的某棵子树，root(t)表示这棵子树的根节点。该算法的时间复杂度是 O(N*|S|)，N 表示聚类结果树中叶子节点的数量，即搭配词的数量。

本文以图 2 示意说明。假设图中是聚类结果树的一部分。最底层的是叶子节点，表示目标词的搭配词。非叶子节点表示中间聚类结果。本文的目标是找到一条使得模型数据描述总长度最小的切分线，切分线上的类就是最优聚类结果。假设当前考察的是以 C 为根节点的子树。它的两个儿子节点分别是 G 和 D。由于 D 节点下仍有子树，于是接着考察以 D 为根节点的子树。结果发现 $L'([D]) < L'([H \cup I])$，因此切分线落在 D 节点上。将 G 和 D 合并与 C 进行比较，发现 $L'([C]) > L'([G \cup D])$，因此切分线落在 G 和 D 上。如图中实线所示。重复这一过程直到生成完整的切分线。

```
算法 Find-MDL(t) := cut
  If
    t 是一个叶子节点
  Then
    Return(t)
  Else
    For each t 的子节点 t_i ,
    C_i = Find - MDL(t_i)
    C := 连接( C_i )
    If
        L'([root(t)]) < L'([C])
    Then
        Return ([root(t)])
    Else
        Return (C)
```

图 1 寻找最优聚类结果 Find-MDL 算法的伪码

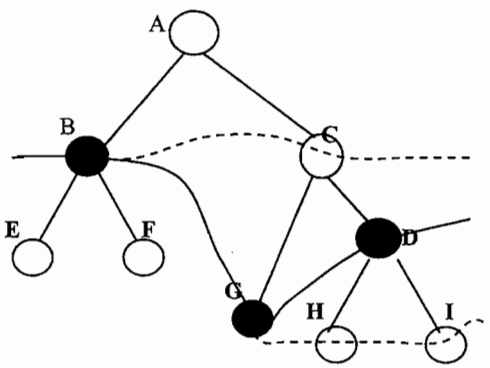

图 2 最优聚类结果的生成图示

3 实验及分析

3.1 评价方法

通过搭配抽取、层次聚类、最优聚类结果选择三个步骤，本文得到了目标词的搭配词聚类结果。我们的评价目标是判断聚类结果中的搭配词集合是否对目标词的词义有区分作用。为此，本文设计了一种基于人工相关性判断的评价方法，利用专家对搭配词集合与目标词词典定义之间的相关性信息来判断搭配词集合的词义区分能力。当然词义区分结果与现有的词典定义是有差别的，并不能简单地将词典的定义当成是标准答案。因此本文希望通过这种评价方法反映词义区分结果和词典定义之间的差异。

假设有 N 个专家参与评测。目标词 w 的搭配词聚类结果为 $C = \{c_1, c_2, \cdots, c_n\}$，$C_i$ 是一个搭配词集合。目标词在词典中的定义集为 $S = \{s_1, s_2, \cdots, s_m\}$。

首先由专家给出每个 c 和每个 s 的“相关程度”。本文为“相关程度”定义三个级别，分别是“完全相关”，“部分相关”和“完全无关”。“完全相关”指的是 c 中任何一个搭配词与目标词搭配使用的时候，目标词能够明确体现定义 s。“部分相关”指的是 c 中大部分（超过一半专家人数）搭配词与目标词搭配使用的时候，目标词能够明确体现定义 s。“完全无关”指的是 c 中只有小部分（不超过一半专家人数）或者没有搭配词与目标词搭配使用的时候，目标词能够明确体现词典定义 s。

严格情况下，本文只将“完全相关”看成是满足相关条件的，即如果 c 与 s“完全相关”，那么 c 与 s 具有相关关系。但也可以放宽条件，将“完全相关”和“部分相关”都看成是满足相关条件的，即如果 c 与 s“完全相关”或者“部分相关”，那么 c 与 s 具有相关关系。于是本文定义函数 ***Map***(c,s)，该函数返回所有认为搭配词集 c 与词典定义 s 满足相关条件的专家人数。本文将 c 与 s 的 ***Map***(c,s) 在所有专家人数中所占的比例称为专家认为 c 和 s 满足相关条件的“认同度” ***Rel***(c,s)。见公式 4。

$$\boldsymbol{Rel}(c,s) = \frac{\boldsymbol{Map}(c,s)}{N} \tag{4}$$

本文定义函数 ***Rep***(c,s) 来表示搭配词集 c 能否表示词典定义 s。如果搭配词集 c 与词典定义 s 的认同度大于概率 P，本文就认为搭配词集 c 能够表示词典定义 s，Rep 函数返回 1，否则函数返回 0。P 的值限定在 50%到 100%之间。如公式 5。

$$\boldsymbol{Rep}(c,s) = \begin{cases} 1 & (\boldsymbol{Rel}(c,s) > P) \\ 0 & \text{otherwise} \end{cases} \tag{5}$$

最后本文用 ***Success***()函数表示搭配词集是否具有区分目标词词义的能力。即，如果搭配词集能且仅能表示一个词典定义，或者当它能表示多个词典定义时，认同度最大的两个定义的认同度之差大于等于阈值 I，那么我们认为该搭配词集有区分目标词词义的能力。前一种情况下，搭配词集唯一能表示的词典定义就是搭配词集能够区分出来的定义。后一种情况下，认同度最高的词典定义就是这个搭配词集能够区分出来的定义。见公式 6。式中，$\boldsymbol{Rel}(c,s)_{rank1}$ 表示最大的认同度数值，$\boldsymbol{Rel}(c,s)_{rank2}$ 表示第二大的认同度数值。I 在(0,P]范围内变动。特殊情况下，当 I 等于 P 时，如果搭配词集能够表示多个词典定义，那么这个搭配词集就一定不具备词义区分的能力。

$$\boldsymbol{Success}(c) = \begin{cases} 1 & (\exists | s_i \in S : \boldsymbol{Rep}(c,s_i)=1) \text{ or } ((\boldsymbol{Rel}(c,s)_{rank1} - \boldsymbol{Rel}(c,s)_{rank2}) \geq I) \\ 0 & \text{otherwise} \end{cases} \tag{6}$$

最后，本文定义词义区分结果的区分率(Discrimination)和覆盖率(Coverage)。区分率表示具有词义区分能力的搭配词集在所有搭配词集中所占的比例。覆盖率表示具有词义区分能力的搭配

词集所表示的词典定义在目标词的所有词典词义中所占的比例。见公式 7。

$$Discrimination = \frac{\sum_{c \in C} Success(c)}{|C|} \qquad Coverage = \frac{\text{具有词义区分能力的c区分出的s的种类数}}{|S|} \quad (7)$$

3.2 结果分析

本文对四个汉语常用多义形容词进行了基于搭配的词义区分研究，分别是“健康、紧张、清楚、大”。图 3 和图 4 显示的是当 I 值固定为 0.5 时，在不同相关条件的认定下，四个目标词的区分率和覆盖率随 P 值变化的情况。图 5 显示的是当 P 值固定为 0.5 时，四个目标词的区分率和覆盖率随 I 值变化的情况。

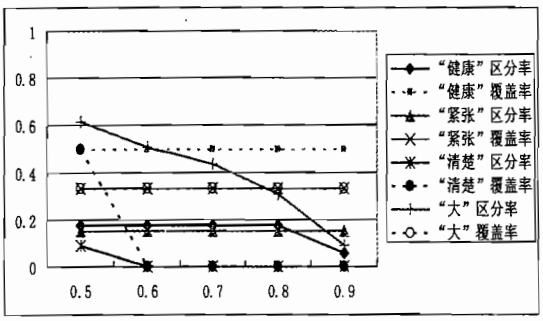

图 3 四个目标词的区分率、覆盖率（“完全相关”，I=0.5）

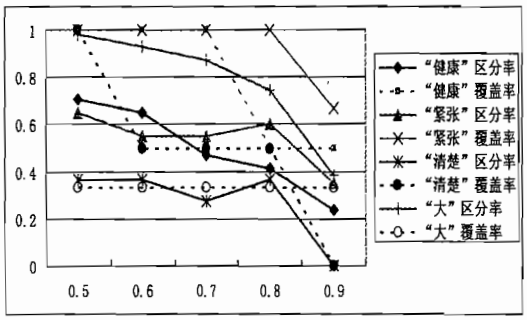

图 4 四个目标词的区分率、覆盖率（“完全相关”+“部分相关”，I=0.5）

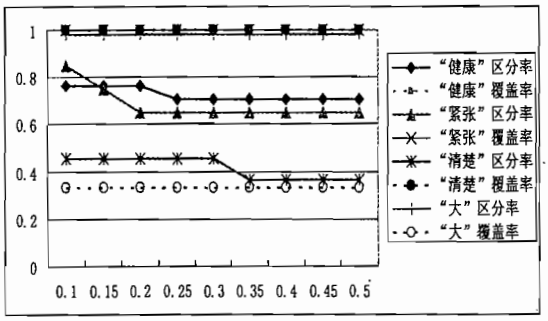

图 5 四个目标词的区分率、覆盖率（“完全相关”+“部分相关”，P=0.5）

从图中可以看出，第一，“完全相关”下的区分率和覆盖率比“完全相关”加上“部分相关”下的区分率和覆盖率低，而且差别比较大。这是因为“完全相关”的搭配词集合在所有搭配词集合中所占的比例相对较小，“部分相关”对区分率所起的作用比较明显。

第二，四个目标词的平均区分率都是比较高的。特别是“健康”和“大”。但是“大”的覆盖率普遍比较低。这主要是因为“大”在《现汉》中的两个定义“排行第一”和“用在时令或节日前，表示强调”在语料中几乎没有出现。目前的词义区分技术还很难自动地获取这些不出现或出现次数极少的词典定义，这就造成了自动的词义区分结果与词典定义之间的很大不同。目前可以缩小两者差距的方法是尽可能多地搜集平衡语料，获取低频词义的知识。

第三，总体来说，区分率和覆盖率是随着 P 值的增加而降低的。因为 P 值越高，搭配词集合能够表示词典定义的可能性就越小，能够区分词义的可能性就更小。不过区分率在局部范围可能有上下波动的情况，例如图 4 中，“紧张”在 P 等于 0.8 时的区分率比 P 等于 0.7 时的区分率高。这主要是受到不具有词义区分能力的搭配词集合的影响。例如“紧张”的搭配词集“异常/d|

更加/d|日益/d|日趋/d”在三个定义下的认同度分别是“0.7、0.75、0.9”。如果 P 等于 0.7，由于 0.75 和 0.9 都大于 0.7，该搭配词集合能够同时表示两个定义，因此该搭配词集不具有区分词义的能力。但当 P 等于 0.8 时，该搭配词集合只能表示一个定义，于是被判定为具有词义区分的能力，区分率反而上升了。因此 P 值一般不应设的太高，否则就会产生评价的错误。

第四，在 P 值固定不变的情况下，区分率和覆盖率变化都不是很大。这说明 I 值对评价结果影响不大。这也说明大多数搭配词集都只有一个词义的认同度大于 50%。不过为了保证搭配词集合区分词义的典型性，在实际操作中，I 值应该设的比较高。

通过实验本文获取了目标词不同词义下的典型搭配，例如“紧张”词义 1 下的典型搭配词集合“难免/v|心里/s|单调/a|显得/v|太/d|过于/d|麻烦/a|惶恐/a|压抑/a”等，词义 2 下的“作息时间/n|只怕/d|调试/v|装载/v|升高/v|缘何/r”等，词义 3 下的“财力/n|运力/n”、“电力/n|水资源/n”、“供应/v|供水/v”等。同时，本文也获得了目标词的不具有词义区分能力的搭配词集合。例如“紧张”的“异常/d|更加/d|日益/d|日趋/d”、“非常/d|比较/d|相当/d”、“空前/a|愈发/d”等等。

为了更好地服务于词义研究，本文对原有层次结构的搭配知识描述框架进行了修改，将词义区分能力放到描述框架中。例如“紧张”一词在新的描述框架中的内容如图 6 所示。

[紧张]
具有词义区分能力的
词义 1　麻烦/a 惶恐/a 压抑/a 恐惧/a 失调/v……
词义 2　作息时间/n 只怕/d 调试/v 装载/v 升高/v……
词义 3　财力/n 运力/n 电力/n 供应/v 供水/v……
不具有词义区分能力的
异常/d 更加/d 日益/d 非常/d 比较/d 相当/d 空前/a……

图 6 “紧张”在新的搭配知识描述框架中的内容

4 结论

本文针对现有搭配词典和标注资源的不足，将词义研究和搭配研究结合起来，设计并实现了基于搭配的汉语词义区分方法，根据词语的搭配特征区分词义，最后得到能够用于词义区分的搭配知识。由于受到依存句法分析器的限制，本文还不能自动获得搭配的结构类型信息。因此本文将在下一步工作中，丰富搭配描述内容，改进搭配抽取方法获得更多搭配信息，例如增加前后位置信息、搭配类型信息、频次信息等。进一步优化词义区分算法、扩大语料，并将名词和动词纳入研究的范围，为构建大规模的基于真实文本的搭配知识库奠定基础。

参 考 文 献

[1] 孙茂松, 黄昌宁, 方捷. 汉语搭配定量分析初探. 《中国语文》第 1 期，29-38 页. 1997.

[2] Xu, Ruifeng. Qin Lu, Sujian Li. The Design and Construction of A Chinese Collocation Bank. In Proceedings of the 5th International Conference on Language Resources and Evaluation, 2006: 1880-1885.

[3] 全昌勤, 何婷婷, 姬东鸿, 刘辉.从搭配知识获取最优种子的词义消歧方法. 中文信息学报..2005, 19(1).

[4] 闻扬, 苑春法, 黄昌宁. 基于搭配对的汉语形容词—名词聚类. 中文信息学报. 2000, 14(6).

[5] Yarowsky, David. One sense per collocation. In Proceedings of the ARPA Human Language Technology Workshop. 1993.

[6] Duan, Xiangyu. Jun Zhao. Ungreedy Methods for Chinese Deterministic Dependency Parsing. Proceedings of Twenty-second Conference of Association for Artificial Intelligence Student Session. 2007: 22-26.

[7] Li, Hang and Naoki Abe. Generalizing Case Frames Using a Thesaurus and the MDL Principle. Computional Linguistics. 1998. 24(2): 217-244.

[8] Rissanen, Jorma. 1983. A universal prior for integers and estimation by minimum description length. The Annals of Statistics, 11(2):416–431.

[9] 王锦, 陈群秀. 汉语述语形容词机器词典机器学习词聚类研究. 中文信息学报, 2007, 21(3): 40-46.