利用语义特征生成搭配

赵晨光　蔡东风
沈阳航空工业学院自然语言处理实验室　沈阳　110034
E-mail:zcg09@hotmail.com

摘要:本文提出了一种在建立搭配模板库的基础上，通过对搭配词对进行基于语义相似性的替换，衍生出更多搭配的设计思想，初步实验结果表明这是一种有效的扩充词语搭配库的方法。

关键词:知网，词语搭配，语义特征

Using Semantic Character for Generating Collocations

ZHAO Chen-guang　CAI Dong-feng
NLP Lab,ShenYang Institute of Aeronautic Engineering　ShenYang　110034
E-mail:zcg09@hotmail.com

Abstract: This paper proposes to generate more collocations by replacing the collocation pairs based on similarity of semantic character after creating collocation template. The rudimental experiment result shows that it's an efficient strategy of extending word collocation store.

Keywords: HowNet, word collocation, semantic character

1 引言

汉语词汇浩如烟海，词语的搭配方式和范围更是多种多样，尤其在文学、艺术领域，词语搭配的自由度更大，在特定的语境中，往往出现许多异想不到的异常搭配，然而就搭配的规范性而言，似乎又有规律性可循。

通常，词语搭配的自动抽取主要基于统计的方法[7]，在其抽取算法中，确定搭配候选的观察窗口后，应用了互信息(mutual information)和方差两个统计特征[4][8][9]。但是,这种方法是就纯统计意义而言的，只有进一步引进语言学特征，才能对搭配词语的语义关系进行全面、准确的描述[5]。当然，我们也应看到一个成分与另一个成分能否搭配，除了语义因素外，还包括句法因素，语音因素，常识因素等。具体到两个实际使用的语言成分能否搭配是多方面因素共同作用的结果。任意两个成分 x 和 y，它们能不能搭配，我们认为无论选择靠语义范畴来限定，还是用句法范畴来限定，或以语用范畴限定，只要有办法组织起一套明确的范畴体系，尽可能全面、准确地描述语言成分的搭配知识，就是好的选择，至于所选的范畴可以有意无意地淡化。

应该说，词语的搭配其语义知识所描述的搭配限制比句法知识描述的搭配限制更为严格，在语义层面上探讨汉语的组合规律，对汉语的构词研究和词组组合规律研究有重要意义。但我们也看到语义知识的获取又是非常困难的工作，语义知识很难描述到足够的深度和广度[3]。然而，目前《知网》为我们提供了一部比较详尽的语义知识词典[1]，这为基于语义来生成词语搭配提供了有利的工具。《知网》是以词语所代表的概念作为描述对象，以揭示概念间及概念本身属性间的关系为基本内容的常识性知识库。其中，描述概念的最基本单位是义原，义原间存在 8 种关系：上下位关系、同义关系、反义关系、对义关系、属性-宿主关系、部件-整体关系、材料-成品关系、事件-角色关系，由此，构成一个义原层次体系。

在《知网》中对于概念的定义是采用 DEF 语义表达式，DEF 描述了词语详尽的语义特征，如：生日：DEF=time|时间，day|日，@ComeToWorld|问世，$Congratulation|祝贺

描述式以逗号进行分割，第一个描述式代表了该词语的最基本的语义特征，其中带有符号的描述式表示了语义描述式之间或概念间的关系。由此可见，我们可以利用 DEF 表达式中罗列的语义特征来刻画词语的语义关系网，从而找到词语间的内在联系。如：

节日：DEF= time|时间，Festival|节日，@Congratulation|祝贺

比较“节日”和“生日”的语义特征非常相似，因此，考虑在词语搭配生成时，若已有的搭配词库中存在”庆贺生日”这样的搭配，基于语义特征的相似性，可将“生日”替换为“节日”，衍生出多种可能的搭配，而绝不会受到语料领域、规模的限制。

2 搭配的抽取

在利用语义知识生成搭配之前，首先需要建立一个搭配模板库，并以其为基础，通过语义替换生成更多的搭配。我们利用统计的方法找到搭配模板。在已经过分词和标注的语料库中(分词系统采用了中科院计算所的词法分析系统)，如果任意两个词的共现率高，这说明它们之间存在着关联，但并不意味着它们可以构成合理的搭配。因此，我们先在已知语料库中抽取共现率高的邻接词对作为搭配词语候选集，然后应用词性标注模板对候选词对进行过滤，生成搭配模板库。如表 1 列出了语料库中部分共现率较高的搭配候选词：

w1	w2	C(w1,w2)
一	类	1020
可以	在	971
位置	上	660
没	时间	580
对	结果	452
令	人	324
他	说	212
有	必要	188
相当	多	123
电话	号码	86

表 1 搭配词语候选集

A	N	形容词与名词构成搭配
V	N	动词与名词构成搭配
N	N	名词与名词构成搭配

表 2 词性标注模板

在经过词性标注模板滤掉不符合上述搭配模式的词对后，产生的部分结果如表 3：

V-N 搭配模板		A-N 搭配模板		N-N 搭配模板	
W1	W2	W1	W2	W1	W2
扒/v	窃/ng	全/a	屋/n	衣/ng	袋/ng
装满/v	钱/n	真/a	象/ng	电话/n	号码/n
乘/v	客车/n	左/a	脸颊/n	航空/n	公司/n
是/v	小事/n	冷/a	雨/n	市/n	街/n
盯住/v	车/n	灵巧/a	手指/n	楼梯/n	口/n
有/v	安全感/n	大/a	眼睛/n	关键/n	时刻/n
伤/v	朱利亚/nr	洼/a	洞/n	出版社/n	编辑/n
送行/v	酒会/n	大/a	声/n	老朋友/n	约翰/n
令/v	人/n	好/a	嗓子/n	女佣/n	人/n
没/v	时间/n	旧/a	信/n	时装/n	杂志/n

表 3

基于共现率和词性标注模板来建立词语搭配模板库后，接着我们需要利用《知网》表述的语义信息对模板库中的搭配词对进行语义标注，用以确定词的义项，标注方法采用自动标注和人工标注相结合的方法，这就使得模板库中的搭配词对即包含词性信息也包含相关的语义特征。然后再利用已知的语义信息在搭配模板库中的各个搭配词对中逐一进行替换，从而可以生成语料库中未曾出现的搭配。由于是从语法语义出发生成可能的搭配，所以生成的搭配并不一定都符合人们的语言使用习惯，因此还需要进行统计方面的检验。如果生成搭配在语料库或网上的使用频率超过某一指定值，才可以将生成的新搭配作为正式搭配加入到搭配库中。这种从语义出发生成的搭配，语义信息确定，不需要再进行困难的语义标注，是本方法的一个重要特征。

我们的做法是将 w1、w2 分别作为中心词，对其搭配词进行多次替换，从而生成多个不同的搭配，替换的过程实际上就是寻找语义概念相似的搭配词的过程，这里的“相似”与“相关”不同，主要指词语在上下文中是否可替换，即基于搭配范围内的替换。因此，问题归结为可替换搭配词的语义特征相似性的确定。设计思路是：先以词对 W1、W2 中的任意一个作为待替换的目标词，设以 W1 为中心词，W2 为目标词，找到该词在《知网》中的义项， w1、w2 在《知网》中是以义项（概念）的描述形式出现的，在《知网》中每个概念用记录来表示，例如：

NO=015492
W_C=打
G_C=~毛衣,~毛裤,~双毛袜子,~草鞋,~一条围巾,~麻绳,~条辫子
W_E=knit
G_E=V
E_E=

DEF=weave|编织

其中 NO 为义项编号，W_C，G_C，E_C 分别为汉语词语、词性和例子，而我们所关心的是 DEF 语义表达式，也就是《知网》对于义项的定义，即义项本身的语义特征。在 DEF 中，包括三种形式的语义描述式：独立义原，关系义原和符号义原，其中，独立义原描述了词语的最基本语义特征。并取出该义项所对应的语义表达式 DEF，我们要检索出与该 DEF 完全匹配的义项，也就是与该词的同义义项,这些义项在很大程度上可替换 W2，作为 W1 的搭配词。譬如：在《知网》中“悦耳“、“动听”、“悠扬” 的 DEF 表达式都相同：

DEF=aValue|属性值，Sound Quality|音质，good|好，desired|良

因此，上述三个词可看作同义词，如果搭配模板库中有”悠扬歌声”的搭配词对，”悠扬”作为待替换的目标词，可由”动听”或”悦耳”对其加以替换。

当然，基于同义的替换还远远不够，譬如以下搭配：

w1	w2
打	篮球
打	网球
打	游戏

显然，W2 搭配词集中各词并非同义，但都可以与 W1 这个中心词构成搭配，不难看出,w2 词集中各词的语义特征具有一定的相似性，因此，我们考虑在《知网》中提取 W2 的语义表达式 DEF，分析 DEF 表达式的语义特征，而语义特征是通过义原描述式来刻画的，这样，我们可以把寻找可替换 W2 词语的问题归结为在《知网》中寻找与 W2 义项的 DEF 表达式在一定程度上相似的义项。

由于义项以义原来表示，所以义原的相似度计算是义项相似度计算的基础。因为所有的义原根据上下位关系构成了一个树状的义原层次体系，可采用通过语义距离计算相似度的办法。根据两个义原在层次体系中的路径距离 d，计算得到义原的相似度：

$$Sim(w1,w2)=\alpha/(d+\alpha)$$

其中，w1 和 w2 表示两个义原，α 是一个可调节参数。

一般来说，一个 DEF 含多个义原描述式，各个义原描述式作用是不同的。我们先以第一独立义原作为检索条件，检索到所有具有相同的第一独立义原的义项，并将这一部分的相似度记为 Sim1(w2',w2)，但是,仅具有相同的第一独立义原的备选词很宽泛，无法界定其是否与目标词相似，因此，还需考虑以下几部分：

(1) 其它独立义原描述式：语义表达式中除第一独立义原以外的所有其他独立义原(或具体词)，这一部分的相似度记为 Sim2(w2'，w2)

(2) 关系义原描述式：用”关系义原=基本义原”或”关系义原=(具体词)”来描述；这一部分的相似度记为：Sim3(w2',w2)

(3) 符号义原描述式：用”关系符号 基本义原”或者”关系符号(具体词)”加以描述；这一部分的相似度记为 Sim4(w2',w2)

因此，两个义项语义表达式的整体相似度记为：

$$Sim(w2',w2)=\sum \beta_i Sim_i(w2',w2)$$

其中，$\beta_i(1\leqslant i\leqslant 4)$是可调节参数，且 $\beta_1+\beta_2+\beta_3+\beta_4=1$，$\beta_1\geqslant\beta_2\geqslant\beta_3\geqslant\beta_4$，

这反映了β_1到β_4对于总体相似度所起到的作用依次递减。由于第一独立义原反映了义项的最主要的语义特征，所以将其权值设置较大。

由此可见，利用《知网》提供的语义特征的描述，可根据计算得到的语义特征的相似性，对搭配词进行基于语义的替换。

3 实验结果及分析

根据以上方法，我们首先建立了具有1000个词对的搭配模板库，其中，对搭配模板库中词对的筛选采用了人工判别的方法。然后逐一计算搭配词对中与两个词在《知网》中语义相似度高($\beta \geq 0.5$)的可替换该词的义项，几个参数的取值如下：$\beta_1=0.5$, $\beta_2=0.25$, $\beta_3=0.15$, $\beta_4=0.1$, $\alpha=1.5$

部分实验结果如下表所示：

W1	W2	W2'	β(w2', w2)
乘	客车	汽车	0.862
乘	客车	货车	0.862
乘	客车	机车	0.862
乘	客车	卡车	0.862
乘	客车	自行车	0.653
乘	客车	三轮车	0.653
乘	客车	火车	0.653
乘	客车	车头	0.653

表4

W1	W2	W1'	β(w1', w1)
乘	火车	坐	1
乘	火车	搭	1

表5

观察上述实验结果，以搭配词对“乘客车”为例，表4是以“客车”作为待替换的目标词，在《知网》中经检索计算后得到的与目标词“客车”语义相似度高的词集W2'，其中，“客车”和“车头”的相似度较高，但“乘车头”这种搭配显然是不合理的，而后者以“乘”为目标词，其结果是较为理想的。因此，经语义替换后生成的搭配还需进行必要的验证，最好是放到网上这样一个大规模语料中，考察是否在实际语料中存在这种搭配情况，经筛选可以得到更为理想的效果。

4 总结

本文将语义相似度技术引入到词语搭配库建立的应用当中，论述了基于已有的搭配模板库，通过语义相似的替换，生成搭配的方法，并简要归纳了在研究过程中形成的一

些观念性的认识。经尝试性的实验证明，我们的做法充分利用了《知网》丰富的语义信息，得到的初步结果基本上达到了预想的目标，上述方法是可行的。这种方法不同于以往的统计，它不仅给出了词语搭配的结构信息，同时还以形式化的方式对包括中心词和搭配词在内的词汇进行语义标注，赋予了一定的语义信息。应该说，通过统计的方法我们可以得到搭配词间的共现率、互信息、离散度等特征，它是基于定量的分析，而上述方法，在语义层面上，体现了搭配词间的语义关系和组合规律，是定性分析。

词语搭配库建设是一个基础性工作，该信息库建成之后，可以应用在语义分析、语言生成、信息检索、文本自动分类、自动文摘等领域当中。

当然，在引入语义特征的情况下，不可避免地会带来噪音，生成搭配的正确率还有待于进一步验证和完善，距离能够真正生成合理而准确的搭配并有效地抑制噪音，还有相当长的路要走，这也是我们今后工作的目标。为了进一步改善生成搭配的结果，我们也在考虑是否可以将语义标准与统计标准相结合，一方面利用语义知识对搭配做以限制，另一方面，利用共现率考察搭配的分布特征。同时，可基于网络这一庞大的语料资源进行搭配的验证。当然，网络不同于常规的精选平衡语料库，它存在着噪音和大量的“垃圾”，例如，网络中有 “喝面包”、“吃牛奶”、“打足球”等搭配，这在日常的语言现象当中很可能出现，但在语义搭配上不合理，这类问题还有待于进一步探讨。

参 考 文 献

[1] 董振东，董强：《知网》，http://www.keenage.com

[2] 刘群，李素建： 基于《知网》的词汇语义相似度计算，第三届汉语词汇语义学研讨会，2002 年 5 月

[3] 詹卫东： 面向中文信息处理现代汉语短语结构规则研究，清华大学出版社，1992 年

[4] 孙宏林： 词语搭配在文本中的分布特征，1998 中文信息处理国际会议论文集，67-72

[5] 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 1997 年，29-38

[6] 孙宏林，段慧明： 面向自然语言处理的现代汉语短语信息库，《术语标准化与信息技术》，1998 年第 2 期，26-31

[7] Pavel Smrz, Pavel Rychly: Finding Semantically Related Words in Large Corpora, Proceedings of TSD, 2001

[8] Christopher D. Manning, Hinrich Schutze,: Foundations of Statistical Natural Language Processing. Cambridge, MA: The MIT Press. 1999

[9] Kenneth Ward Church and Patrick Hanks: Word association norms, mutual information, and lexicography. Computational Linguistics,16(1). 1989