汉语动宾搭配的自动识别研究[1]

高建忠

联想研究院信息工程研究室

自然语言处理研究组，北京 100085

gaojz@legend.com

摘要： 本文面向自动句法分析的实际需要，对大规模真实文本中动宾搭配的自动识别问题进行探索性研究，提出了“概念＋词语”匹配模型，并在借鉴国外相关研究经验的基础上提出“词语+词语相似度”计算模型。通过开放测试检验并比较了两种算法的效率，初步显示了实验成果的研究价值。

关键词： 句法分析，动宾搭配，相似度，部分句法分析

Automatic Identification of Verb-Object Phrase for Chinese Language

Jianzhong Gao

NLP Group, Information Engineering Lab.

Legend, Beijing 100085

gaojz@legend.com

ABSTRACT: In this paper, we proposed two efficient Chinese Verb-Object phrase identification algorithms. Based on the following processing strategies and schemes: 1) To combine verb concept with noun as recognition pattern, 2) To use similarity between nouns for this task. It obtained satisfactory performance in the experiments of the automatic Verb-Object phrase identification on Chinese real texts.

KEYWORDS: Verb-Object Phrase, Similarity, Partial Parsing

1. 引言

动宾结构在SVO型语言里普遍存在，是句内的核心成分和名副其实的“骨架”，它实际上映射了整个句子的轮廓。如能在自动句法分析过程中首先准确识别出动宾结构，我们就有可能在此基础上进行一些后续分析，从动词出发，向左搜索各种状语，逼近句子的主语成分；从宾语出发，向左搜索各种修饰成分，逼近动词，或向右搜索其他成分，从而为实现完全的句法分析奠定一定的研究基础。

本文所指的动宾搭配，主要是从位置关系出发，对只带一个体词性宾语的动词和在句子中位于其后的体词性名词之间是否能构成动宾关系进行判定，满足条件的，即认为是合法的动宾搭配。

1 本研究得到北京语言文化大学陈小荷教授的悉心指导，谨致谢忱。

就处理范围而言，我们并不是面向全部动宾搭配。动词和名词的集合是经过甄选后形成的。其中，动词的集合我们没有考虑只能带小句宾语的动词、带兼语的动词和带双宾语的动词。宾语都是真宾语。就述宾意义关系而言，宾语既可指动作的施事、受事，也可以是动作凭借的工具，还可以是动作的结果，或位移的终点。动量宾语、时量宾语和数量宾语（即准宾语）未在本文考虑范围之内。

2. 动宾搭配识别模型介绍

我们直观地认为，表示具体意义的名词，尽管在语言使用上能产性强，无法穷尽获取，但对这类名词，使用适合的语义知识就能相对理想地加以解决。而对表示抽象意义的名词，我们可以通过统计等方式进行处理。比如，就“吃”而言，我们可以比较准确地对其宾语进行定性描写；而就“实现”而言，对它可能的宾语，如“志向”、“志愿”、“目的”、“民主”、“打算”、“政策”、“口号”、“路线”等进行描写就不那么轻松了。

基于以上这种考虑，并结合所选语料的性质，我们把名词范围调整到以具有抽象意义的名词为主。具有抽象意义的名词在新闻体语料中大量存在，选择处理由这一类名词构成的动宾搭配也符合语料的实际。我们的初衷是，在避免使用任何描写抽象意义名词的语义资源的前提下，通过构造一定的计算模型，达到对由这一类名词构成的动宾搭配的自动识别和标注。为此，我们设计构造了两种计算模型。

模型（一） 基于动词概念的语义限制模型

一般而言，构成组合型语义关系知识的双方都应是语义层面上的。我们采用了非完全的语义限制模型，构造了动词概念和具体名词之间的组合知识。以名词为出发点，利用名词对能与之构成动宾搭配的动词的概念的规约关系，检验某一动词是否能和该名词组合为合法的动宾对子。当某一动词满足该名词对所搭配动词的语义期待时，即为合法的。我们以《现代汉语辞海》的动宾搭配为工作平台，以《知网》为概念资源，构造语义限制模型。

模型（二）利用词语相似度的统计模型

该模型属基于分布理论的模型。基于分布理论的模型，其依据是词语分布假设：一个词的语义和语法功能决定了它和其他词的组合关系。我们借鉴了 Ido Dagan 等（1999）所介绍的基于词语分布的相似度模型（Simlarity Based Models）之一——混同概率模型（Confusion Probability model）。该模型计算公式为：$Pc(w_1'|w_1)=\sum_{w2}\frac{P(w_2|w_1)}{P(w_2)}\cdot P(w_2|w_1')P(w_1')$。其中，$w_1$、$w_1'\in V_1$，$w_2\in V_2$，（$V_1$，$V_2$分别为词的集合），词对（$w_1,w_2$）$\in V_1\times V_2$，$P(w_2|w_1)$是在给定$w_1$的情况下$w_2$的条件概率，相应地，$P(w_2|w_1')$是在给定$w_1'$的情况下$w_2$的条件概率。混同概率表示词语$w1'$可替换$w1$的概率。其值域为[0，$0.5\max_{w2}P(w_2)$]。$Pc(w_1'|w_1)$反映出相对于某一特定词集（$V_2$）而言，词语$w_1'$、$w_1$在分布上的相似性，值越高，相似性越大，$w_1$能够被$w_1'$替换的概率越大。

应用混同概率，目的在于解决真实文本的数据稀疏问题。对于未现事件，可以根据它和已现

事件之间的相似度（Pc）高低作出判断。我们考察的是名词之间的相似度。举一简例：假如当前文本为“…建立…机制”，“建立 机制”为未知的动宾搭配，而“建立 体制”是已知的动宾搭配。此时，我们就用“体制”、“机制”的相似度信息进行判断，即 Pc(机制|体制)。

3. 研究的当用资源

语料方面，本研究选用的语料来源于北京语言文化大学语言信息处理研究所的“现代汉语研究语料库”，训练语料为 2500 万字，测试语料为 50 万字。语料仅作了分词处理。本研究的核心资源是语言知识。我们以《现代汉语语法信息词典》、《现代汉语辞海》和《知网》作为本研究的语言知识库。三种资源分属语法、搭配、语义知识，有机结合共同作用于本研究。

在抽取出《现代汉语辞海》动宾搭配对后，我们利用知网中的运动类概念体系对这些搭配对中的动词进行概念标注。在全部 52310 个对子中，33008 个的动词在知网中是单概念的。有 1607 个对子的动词未在《知网》中出现，因此实际应标动词概念的对子数为 50703 对。余下的 17695 个对子便是有歧义动词。对于有歧义动词，我们对概念进行了人工选择标注，这样在相当程度上可以保证知识源的质量和可用性。

4. 动宾搭配的自动获取

（一）、动宾搭配的自动获取

我们采用统计方法，以《人民日报》1996 年全年新闻语料（计约 2500 万字）为工作平台，获取动宾搭配数据。具体地说，统计每个动词和出现在它后面的名词的同现次数。

我们的获取目标是，由定义好的动词集合、名词集合中的动词、名词构成的所有可能的动宾搭配实例。因此首先要根据动词表和名词表对输入句中的词进行词性判断，经简单词性标注的句子至多有四种可能的词性标记形式：“V”，“N”，“*”和标点本身。判断词性之后，需要对动词和名词的上下文进行局部分析（或称部分句法分析）。这种分析主要是为了排除不大可能进入动宾搭配的动词或名词，从而最大限度降低统计误差，减少伪组合。这一步的主要工作是进行筛选和连续名词序列的捆绑。这样，从全部语料中共获取了 475801 个动宾搭配实例。按照同现次数排序，检查了前 1000 个高频组合（同现次数在 1713～78 次之间），正确率为 78%。我们根据语法信息词典中对体宾动词“前名”、“后名”两属性的描写，从全部获取的组合中抽出具备两属性之一的动词所在的对子，对其进行考察，人工排除了其中频度大于 4 的伪组合。

（二）、相似度模型的参数求解

根据相似度公式，需要对出现在名词集合中的词两两间进行相似度计算。名词集由 1240 个名词组成，由此生成的名词相似度表有 1537600 个名词对。相似度计算，用极大似然法进行概率估计。计算词对的条件概率 $P(w_2|w_1)$ 和 $P(w_2|w_1')$，我们使用的是获取的动宾搭配的统计频度。单个词概率 $P(w2)$ 和 $P(w_1')$，使用该词在全部语料中的出现频度。

相似度计算结果，以名词“政策”为例，前10个最高相似度名词对子及其数据形式如右表所示。从表中可以看出，就名词“政策”而言，其相似度计算结果比较符合我们的期待。最相似词是它本身，在其他相似词中，诸如“措施”、“制度”、“精神”、“原则”、“方针”、“规定”、“标准”、“办法”等词都是比较理想的相似词。

政策	政策	0.01361643
政策	措施	0.00251050
政策	制度	0.00210555
政策	精神	0.00137953
政策	原则	0.00133570
政策	方针	0.00123058
政策	立场	0.00090156
政策	规定	0.00079222
政策	标准	0.00073852
政策	办法	0.00070488

5. 动宾搭配的识别

我们已完成对所需语言知识库的构造，并求得名词相似度统计数据。为了检验这些基本数据的有效性，我们用50万字的分词语料进行动宾搭配的识别。50万字开放测试语料由68篇科技类文本和167篇政治类文本构成，共1.14MB字节。

识别算法（一）基于语义限制的动宾搭配识别 利用词语搭配和名词宾语对述语动词的语义限制条件约束选择匹配对。

（1） 扫描当前句中所有可能进入述宾结构的动词、名词（先进行筛选和名词短语捆绑），将候选动宾组合存入一个数组；

（2） 如果数组为空，转（6）；

（3） 取数组中能够与《现代汉语辞海》动宾搭配实例相匹配的候选组合作为输出，转（5），若无匹配组合，转（4）；

（4） 取数组中动词、名词在语义上能够互相满足的候选组合作为输出；

（5） 从数组中删去该组合，以及跟该组合相冲突的所有组合，转（2）；

（6） 结束。

例如，对于句子：“探索/V 建立/V 适应/V 社会主义/N 市场/N 经济/N 要求/N 的/* 现代/* 企业/N 制度/N 的/* 有效/* 途径/N ； /；”，扫描到动宾组合：①探索…要求，②探索…制度，③探索…途径；④建立…要求，⑤建立…制度，⑥建立…途径；⑦适应…要求，⑧适应…制度，⑨适应…途径。首先利用《现代汉语辞海》动宾搭配对子直接进行匹配，得到动宾对子③，删除该对子及矛盾组合①、②、⑥和⑨；第二次匹配，得到⑤，删除⑤、④和⑧；此时，全部候选对子中仅剩⑦，且该对子不在搭配表内，需调用语义信息进行判定，在“动词概念+名词”的搭配表内，搜索到与名词“要求”搭配的全部动词概念，有“提出”、“告诉”、“说明”、“实现”、“适合”、“违背”等17个；而动词“适应”的概念是“适合”，符合“要求”对所搭配动词的语义限制。此时数组已空，循环结束。得到的识别结果是：[1$探索/V [2$建立/V [3$适应/V 社会主义/N 市场/N 经济/N 要求/N$3] 的/* 现代/* 企业/N 制度/N$2] 的/* 有效/* 途径/N$1] ； /；

识别算法（二）基于相似度的动宾搭配识别 利用词语搭配和候选名词之间的相似

度信息优选动宾组合。本算法在处理流程上只在第（4）步异于算法（一）：取数组中名词相似度最高的候选组合作为输出（假如若干连续候选组合共用同一动词，则要比较各候选名词之间的相似度值高低，取最高的候选名词所在候选组合作为输出）；其余各步骤相同。

对于同样一个标注实例，通过匹配并删除矛盾的候选组合后，剩“适应…要求”，此时需调用相似度信息进行判断。首先取出动词“适应”在辞海动宾搭配表中的全部搭配名词：环境，眼前，自然，需要，需求，性情，形势，习惯，趋势，气候。然后依次用这些名词和“要求”组对。经检索相似度表，检索到的名词对共7个，相似度值排名在前300（第300对相似度值：0.00022745）以内的名词对是6对，由高到低依次是：需求—要求：0.00607915；形势—要求：0.00143646；气候—要求：0.00088242；趋势—要求：0.00055831；习惯—要求：0.00045111；环境—要求：0.00031910。如此经过相似度比较之后，算法认定“适应—要求”是合法的动宾组合实例。

6. 识别结果分析

由于两种识别算法都首先用词对匹配的方法先行识别动宾搭配，所以为了能够比较清晰地观察到语义约束和相似度各自的贡献，我们先只用辞海中动宾搭配实例进行了单纯的匹配。共处理235个文件，实验结果为：

总对子数	识别对子数	正确识别	召回率	正确率
9182	4778	4462	48.60%	93.39%

表中数据显示，单纯用《现代汉语辞海》中现成的动宾搭配实例进行识别，能够得到比较高的正确率，但召回率相对低得多，说明仅这一资源远不能覆盖大规模真实文本；但另一方面也说明了其有效性。

两种算法的VO识别结果

算法	总对子数	识别对子数	正确识别	召回率	正确率
算法（一）	9182	7156	5975	65.07%	83.50%
算法（二）	9182	11470	7803	84.98%	68.03%

不难发现，在识别效率上，两种算法呈现出“低——高”（低召回率，高准确率）和“高——低”（高召回率，低准确率）的特征。比较而言，可以直观形象地认为，算法（一）尽管识别对子数少，但“命中率”高，而算法（二）称得上是“广种薄收”。

我们对标注错误的典型实例进行了分析，具体归纳为：

（1）受名词集合定义容量的约束造成识别错误。如：[1$获得/V 电子/* 部/* 1 9 9 3/* 年/* 十/* 大/* 科技/N 成果/N$1] 奖/*；/；，“奖”未被收入名词底表，而“获得”和“成果”满足算法的条件，故误判为合法搭配。

（2）“等”的问题。集中体现在“等”的依附关系上，造成动宾结构右界识别错误。就“V+N+等…”而言，有两种层次划分一种是：“（V+N）+等…”，即“等”依附在V+N构成的NP之后；

另一种是“V+（N+等…）”，即“等”依附在N之后。例如：[1$探讨/V 了/* 产品/N 型号/N 、/、质量$1] 等/* 问题/N 。/。

（3）“V+N+的+N”的问题。“V+N+的+N”是一种典型的歧义结构，有两种可能的层次划分：“（V+N）+的+N”和“V+（N+的+N）”，前者是NP，动宾结构作NP的修饰或限定性成分；后者是VP，动词V和后面的名词构成动宾关系。这一类型的错误具有典型性。例如：必须/* 尽快/* [解决/V 质量/N] 的/* 问题/N 。/。

此外，还有名词序列的捆绑问题。我们采取了简化处理策略。

7. 结语

本文以大规模真实文本中动宾搭配的自动识别作为研究目标，以受限名词集合所定义的动宾搭配为识别范围，首先提出了“概念+词语”匹配模型，同时借鉴国外相关研究经验，提出“词语+词语相似度”处理模型，构造了两种模型的相应算法并进行了开放测试。实验表明，两种识别算法已经显示出进一步研究的价值。本研究的主要特点在于，输入语料只是作了自动分词但未标注词性的真实文本；以权威的大规模语言知识库作为底层支撑；在使用语言知识的同时，借鉴国外相关研究成果，设计实现了两种计算模型和相应算法。

实验结果已经显示，两种识别算法存在一定程度的互补性。对于进一步的研究，我们初步的打算是将这两种算法结合起来。如何实现二者的有机结合是我们目前感兴趣的问题。本文的后续研究将沿着这个思路进行。

参考文献

[1]陈小荷(1998). “一个面向工程的语义分析体系”，《语言文字应用》1998年第2期。

[2]陈小荷(1999). “动宾组合的自动获取和标注”，《计算语言学文集》，清华大学出版社。

[3]孙茂松等(1997). “汉语搭配定量分析初探”，《中国语文》1997年第1期。

[4]苑春法等(1997). “汉语语义关联网的研究”，《语言工程》，清华大学出版社，1997年。

[5]赵军等(1999).”基于复杂特征的VN结构模板获取模型”，《软件学报》1999年1期。

[6]俞士汶等(1998).《现代汉语语法信息词典详解》，清华大学出版社1998年版。

[7]胡裕树、范晓主编(1996).《动词研究综述》，山西高校联合出版社1996年版。

[8]林杏光审定，张卫国、冀小军等主编(1994).《现代汉语辞海》，人民中国出版社1994年版。

[9]张寿康、林杏光(1992).《现代汉语实词搭配词典》，商务印书馆1992年版。

[10]Andrei Mikheev(1997), “Collocation Lattices and Maximum Entropy Models”, *WVLC-5*

[11]Dekang Lin(1999), “Automatic Identification of Non-compositional Phrases”, *Proceedings of 37th Annual Meeting of the Association for Compositional Linguistics*

[12]Ido Dagan, Lillian Lee, Fernando C.N. Pereira(1999), “Similarity-Based Models of Word Co-occurrence Probabilities”, *Machine Learning*, 34, 43-69