面向新闻领域的主谓关系识别

杨旭　肖桐　张俐

东北大学自然语言处理实验室 沈阳 110004

Email:yangxu@ics.neu.edu.cn

The Identification Of Subject_Verbject Relation For The News Field

Yang Xu, Xiao Tong ,Zhang Li

Natural Language Processing Laboratory, Northeastern University, Shenyang, 110004

Email:yangxu@ics.neu.edu.cn

摘要： 本文提到的主谓关系，专指一个名词和动词经常共现，并且在语义上能够同时构成一个句子的主语和谓语的这样一种关系。本文提出了一种统计结合启发性规则和句法信息的方法来分析主谓关系。实验表明，该方法跟传统的统计方法相比较，F1 值得到了很大的提升。

关键词： 主谓关系，启发性规则，句法信息，统计方法

Abstract: The subject_verbject relation mentioned in this paper specifically refers the relationship between a verb and a noun which co-occurrence frequenly and can form the subject and the vebject of the same sentence semantic1y. This paper proposes a method of statistic combination of heuristic rules and syntax information to analyze subject_verbject relation. Experimental results indicate that the value of F1 promotes a lot compared with the traditional statistical methods.

Keywords: subject_verbject relation, heuristic rules, syntax information, statistical methods

1 引言

越来越多的研究表明，一个包含大量实体对应的框架以及实体之间的相互联系的大规模汉语知识库对计算机语言分析具有至关重要的作用。但是由于目前句法分析技术尚不成熟，要直接从语料库中的句子中去获取我们所要的知识框架是及其困难的。于是，在现有的条件下，我们更容易想到从词语与词语之间的关系这个角度入手去考虑这个问题。

主谓关系的获取对构建知识库是很有帮助的。本文定义的主谓关系，就是指在一个名词作为一个句子主语的前提下，另外一个动词可以同时作为这个句子的谓语动词，它不同于传统意义上的搭配，但由于这种关系比较固定，在语言中的分布比较有规律，所以它与传统意义上的搭配非常相似，本文[1]也经常使用主谓搭配来代替主谓关系这个概念。

目前搭配识别方法主流都是基于统计的，大致有如下几种方法：基于频率，基于卡方检验，基于似然比，基于互信息等方法[5, 10, 11]，以前 Sussex 大学的 Falmer[1]的工作表明，这些统计的方法的性能大体相当。在本文的这个任务中，传统的统计方法对由于语料的特殊所带来的数据稀疏问题比较敏感，而且传统的统计方法没有考虑句法信息，这也使得它们的性能大打折扣。

针对传统统计方法存在的不足，本文提出了相应的改进方法，人工加入一些启发性规则去拒绝一些错误的搭配对，同时通过引入句法信息召回一些低频但正确的搭配对。

[1] 本论文工作得到了国家教育部新世纪优秀人才计划（NCET-05-0287）和国家 863 高科技计划项目（2006AA01Z154）的资助。

本文的结构是这样的，第 1 章对本文任务背景及方法的提出做了简单介绍；第 2 章单介绍传统统计方法；第 3 章介绍本文对传统统计方法的改进，第 4 章是实验和分析，第 5 章是讨论，最后是结论和参考文献。

2 搭配概念及传统搭配抽取方法介绍

传统搭配定义

在传统的自然语言研究过程中，一些计算机和统计学文献的作者通常把搭配定义为带有特殊性质的两个或多个连续的词序列[3, 6, 10]。这些词序列往往具有句法和语义的特性，并且它的准确无歧义不能直接由它的组成部分的意思和含义直接得出。但是在语言学的主导研究中，两个词甚至不连续也可以构成一个搭配，典型语言学对搭配的衡量标准主要包括以下几个方面：非组成构词法，不可替换性，不可更改性。

不同于传统的搭配定义，本文定义的主谓关系，具体是指在一个名词作为一个句子主语的前提下，另外一个动词可以同时作为这个句子的谓语动词，但由于本文利用搭配抽取的方法来分析主谓关系，所以本文通常也使用主谓搭配来代替主谓关系这个概念。例如：

搜救/v 人员/n 在/p 飞机/n 坠毁/v 的/u 茂密/a 森林/n 里/nd 进行/v 了/u 仔细/a 搜索/v 。/wp 在这个句子中，“人员”跟搜索就构成了一对主谓搭配，但“人员”跟坠毁就不是一对主谓搭配。

2.2 传统统计方法介绍

传统的搭配对的识别大多是直接通过比较一些统计量大小的方法去获取搭配对，常见的方法有：使用频率信息的搭配识别，基于卡方检验的搭配对识别，使用似然比或互信息的搭配识别[5, 7, 8]。

下面是这四种方法的简单说明。

方法名称	比较的统计量	简单描述
频率	O_{11}	直接把共现频率作为搭配的依据
卡方	$X^2=N(O_{11}O_{22}-O_{12}O_{21})^2/((O_{11}+O_{12})(O_{11}+O_{21})(O_{12}+O_{22})(O_{21}+O_{22})$	用 X^2 检验两个词的独立性
似然比	$SRB=xlogx(O_{21})+xlogx(O_{12})+xlogx(O_{22})-xlogx(O_{11}+O_{12})+xlogx(O_{11}+O_{21})+xlogx(O_{22}+O_{12})+xlogx(O_{21}+O_{22})+xlogx(O_{11}+O_{12}+O_{21}+O_{22})$	用似然比检验两个词的独立性
互信息	$MUI=(O_{11}/N)/(((O_{11}+O_{21})/N)*((O_{11}+O_{12})/N))$	用互信息检验两个词的独立性

表 1 四种统计方法说明

这里 W_1, W_2 表示文本中的两个词语，我们用 O_{11} 表示 W_1, W_2 共现的次数 O_{12} 表示 W_2 出现但 W_1 不出现的二元组个数，O_{21} 表示 W_1 出现但 W_2 不出现的二元组个数，O_{22} 表示 W_1，W_2 都不出现的二元组个数，N 表示文本中二元对的总数。用 xlogx(a) 表示 $a*\log_{10}(a)$ 的值；

以前 Sussex 大学的 Falmer 的工作表明[2, 4, 9]，无论使用哪种统计方法，性能大致相当，且使用频率的搭配抽取方法就能达到不错的效果。因此本文把使用频率的搭配抽取方法作为 baseline

系统，在此基础进行了改进对比实验。

3 启发性规则和句法信息

3.1 传统统计方法的问题分析

实验表明，直接使用统计的方法去获取主谓搭配，性能是不能满意的。主要原因在于，传统的统计方法对由于语料的特殊所带来的数据稀疏问题比较敏感，而且传统的统计方法没有考虑句法信息，这也使得它们的性能大打折扣。

同时，以“飞机”这个词为例，在输出的结果里面，共现频度达 7 次以上的共有 15 个词，在这 15 个词里面有 11 个是正确的，所以对于高频的搭配对来讲，直接应用统计的方法得到的结果不算太坏但也存在一部分干扰，只要想办法消除这部分干扰，就能使准确率得到更大的提升；另一方面，在那些输出的低频搭配对里面，也有相当一部份词语如“爆炸”，这些词语在新闻领域与飞机形成主谓搭配是很常见的，但由于数据稀疏的原因这个词在该语料库中只出现了三次，而像这类低频出现的主谓搭配，也应该尽量想办法找出来。

3.2 启发性规则

过去一些统计语言学的学者也曾提到，规则和句法信息的引入，对发现语义意义上的搭配是很有帮助的[1, 5,6]，本文提到的主谓搭配就是其中的一种。但是要从一个句子中去完全准确的识别一个主谓搭配是很困难的。上面提到，在本文涉及到的任务中，要改善统计方法的性能必须从两个方面入手，首先需要想办法消除高频搭配中存在的干扰。通过分析大量的句子发现，在一个汉语句子中，虽然动词出现的位置十分灵活，一个句子中其他的一些动词也会干扰我们对谓语动词的正确判断，但是完全可以利用这个词在句中的一些上下文信息去尽量否定一些错误的搭配，本文通过分析这些错误搭配在句子中的上下文信息的一些共同点，总结出一系列的启发性规则，部分规则形式化描述如表 2。

其中 N 表示一个待审定搭配对中的名词，V 表示这个待审定搭配对中的动词，otherV 表示句子其他可能存在的动词。

本文把一个搭配对的共现频率，做为这个搭配对的基准权值，对于这个搭配对在文本中的每一次共现，我们就使用这些启发性规则去审定这个搭配，如果这个搭配被拒绝，就把这个搭配对的权值减掉 1，这样高频出现但不具备主谓关系要求的搭配对有可能被启发性规则拒绝掉。

序号	形式化描述	说明	实例
1	V+……N	动词出现在名词前面拒绝这个搭配	发出/v 警报/n
2	V+“的”	动词后面出现了一个的，拒绝这个搭配	相撞/v 的/u 问题/n
3	N+是+v	本文认为，主语后面一旦有系动词是，就拒绝后面的动词作谓语	飞机/n 是/v 垂直/a 落地/v 的/u
4	p+N+v	当名词作为介词的宾语时，拒绝这个搭配	从/p 雷达/n 上/nd 消失/v
5	N+p+V	动词前紧跟着介词，拒绝这个搭配	飞机/n 在/p 飞往/v 巴西利亚/ns 的/u 途中/nl

表 2 部分启发性规则形式化描述举例

3.3 句法信息

对于低频的搭配而言，目标就是利用句法信息，把那些低频出现但符合主谓搭配要求的搭配对给提取出来。本文利用斯坦福大学的分析器的依存分析结果提供的句法信息，提取那些低频但符合主谓关系要求的搭配对。。以下面的句子为例：

小 飞机 的 机长 接受 空军 调查 时 说 ，

斯坦福大学设计的分析器可以分析出如下一些依存关系三元组:

amod(飞机-2, 小-1)，assmod(机长-4, 飞机-2)，assm(飞机-2, 的-3)，nsubj(说-9, 机长-4)

tclaus(时-8, 接受-5)，nmod(调查-7, 空军-6)，dobj(接受-5, 调查-7)，lccomp(说-9, 时-8)

其中括号外面的英文单词表示了两个词的依存关系，通过观察发现，一旦当两个词在依存分析的结果中的出现了 nsubj 这种依存关系时，是可以作为两个词存在主谓搭配的证据。。

本文把共现频率做为一个搭配的基准权值，对于低频（本文根据语料的实际情况认为出现次数少于 8 次的为低频），每当依存分析的结果中出现一次这两个词做主谓搭配的证据时，我们就把这个搭配的权值适当提高，为了不干扰高频搭配对，本文就认为这个搭配对出现的频率越高，那么提高的权值越小。这里用 fre 表示一个搭配的频率，做为这个搭配的初始权值，对于 fre<8 的情况而言，用 fy 表示依存分析结果中能证明这个搭配是主谓搭配的三元组的个数,调整后的权值用 fre'表示，本文是这样调整它的权值的：

fre'=fre+fy*(8-fre)*0.8

4 实验和分析

4.1 测试语料及测试方法介绍

本文工作是面向空难新闻知识库构建这个任务的。本文所选取的测试语料是大约 1800 多个有关空难新闻的句子，其中的话题大多与飞机失事有关，所以本文在测试时选取了飞机，客机，乘客等 10 个词，做为测试集，这些词在本文所选取的测试语料中出现的频率较高，较能反映测试语料的特点。由于测试语料只有大约 1800 个句子所以具有很严重的数据稀疏性，所以这些代表词所对应的输出列表长度可能会有很大的差异，所以有别于传统的测试方法，本文采取了这 10 个词对应输出列表的 top%N 的正确率（precision）和召回率(recall)作为评价标准，并据此画出了 F1 值（F1=(2 *precision*recall）/(precision+recall)) 曲线。

4.2 统计方法的实验结果

本文首先用四种传统的统计方法包括频率，卡方检验，似然比检验，互信息构建了 4 套系统，我们分别计算测试结果中 10 个词所对应的输出结果中，top N%所对应的正确率和召回率，我们得到了如下 F1 值曲线：

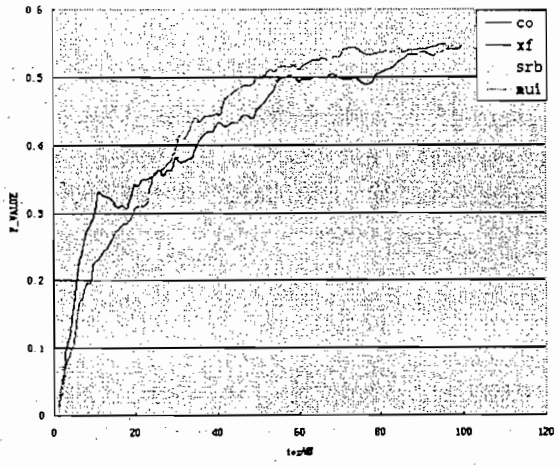

图 1　统计方法得到的 F 值曲线

其中，co 代表频率，xf 代表卡方检验，srb 代表似然比检验，mui 代表互信息，从上面的实验结果可以看出，传统的统计方法性能相差不大,基于频率的方法就已经可以达到不错的性能，但尽管如此仍然很难达到我们的要求，其原因在于，统计方法对由于语料的特殊所带来的数据稀疏的问题比较敏感，而且没有考虑句法信息，这也使得它们的性能大打折扣。

4.3　改进方法的实验结果

我们分别构建了三套系统：基于统计的基础上加入启发性规则，基于统计的基础上加入句法信息，以及在统计的基础上同时加入启发性规则和句法信息。实验得到的关于 top%N 的 F 值曲线如下：

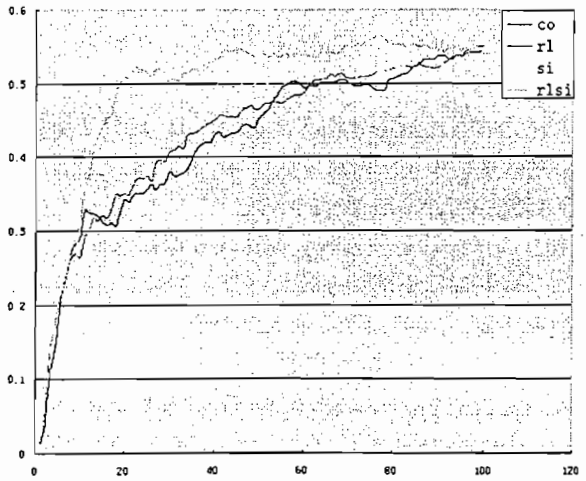

图 2 改进方法的 F 值曲线

在上面的实验结果中，co 代表基于频率的统计实验结果，rl 代表加入了启发性规则后的实验结果，si 代表加入了句法信息后的实验结果，rlsi 代表同时加入了启发性规则和句法信息后的

实验结果。分析实验结果发现，启发性规则的加入对系统性能的提高并不显著，分析其原因在于所用的启发性规则有一定的适应性，在不同的句子中还可能存在冲突的情况；句法信息的引入使得系统性能得到了明显提高，在 top%N 达到 20%时候， F 值曲线达到第一个最大值，随后不断下降，这是因为句法信息的引入时的低频的正确搭配被合理地提到了输出列表的前面部分，说明句法信息的引入对改善低频的主谓搭配抽取是很有意义的；在曲线的后半部分，系统性能趋于稳定，这是因为大部分低频主谓搭配被提到了输出列表的前面，由于数据的稀疏性，剩下的一小部分搭配对是我们系统的盲区。

5 讨论

搭配一词在我们的语言学中难以给出一个准确定义，有些词语之间本身很难说清楚是否能在句子中构成一个主谓搭配，因此迄今为止，没有一个机构为搭配识别提供一个统一的评测标准。为了方便评价本文的方法与传统方法在性能上的改进程度，本文人为地制定了一个评测标准。

本文的工作是面向知识库建设这个任务的，目的是要发现那些在语义上能经常出现在这个领域的搭配，所以对于那些语义上可能存在主谓搭配这种关系，但本文只要认为它不会经常出现在本研究领域，我们就可简单地认为它不是一个正确的搭配，但这并不影响今后工作的进一步展开。

由于本文是要发现那些对知识库构建有帮助的主谓搭配，所以对于某些万能词，如有，没有，能，不能等虽然在语义也能与相关的主语构成主谓搭配，但是对知识库建设帮助不大，在本任务中对这部分搭配词不做考虑。

最后，由于目前的句法分析器（如本文所用到的斯坦福工具）得到的句法分析结果，效果并不理想，所以本文没有用它来直接进行主谓关系的分析，而是考虑用它来间接地改善统计方法的性能。

6 结论

本文提出了一种统计结合启发性规则和句法信息的方法，进行主谓关系发现。该方法是在统计方法的基础上，利用启发性规则和句法信息对统计方法的结果进行合理，有效的筛选，该方法一定程度上有效地克服了传统的统计方法存在的不足。

其中引入启发性规则是为了排除那些出现频度比较高但在语义上不符合主谓关系标准的搭配对；引入句法信息是为了找回那些出现频度较低但符合主谓关系标准的词语。实验结果表明，该方法跟传统的统计方法相比较，F1 值得到了很大的提升。

参考文献

[1].Joachim Wermter Udo Hahn , You Can't Beat Frequency (Unless You Use Linguistic Knowledge) A Qualitative Evaluation of Association Measures for Collocation and Term Extraction,D-07743 Jena, 2006., 1-5.

[2].Prague , An Extensive Empirical Study of Collocation Extraction Methods

Pavel Pecina, Acl2005, 3-6.

[3].Paul Deane, A Nonparametric Method for Extraction of Candidate Phrasal Terms, Acl2005， 1-4.

[4]：Brigitte Krenn ,Can we do better than frequency?A case study on extracting PP-verb collocations, D-70174 Stuttgart,

Acl2001, 1-5.

[5]: Darren Pearce, A Comparative Evaluation of Collocation Extraction Techniques, 2002, 1-3.

[6]: Saim Shin, KAIST,KorTerm,Automatic clustering of collocation for detecting practical sense boundary, 1-4.

[7]: Stefan Evert and Hannah Kermes , Experiments on Candidate Data for Collocation Extraction Institut f¨ur Maschinelle Sprachverarbeitung Universit¨at Stuttgart, EACL2003, 1-3

[8]: Stefan Evert, The Statistics of Word Cooccurrences Word Pairs and Collocations aus Ludwigsburg, 30. August 2004, 1-5.

[9] Winnipeg, Manitoba, Extracting Collocations from Text Corpora, Department of Computer Science University of Manitoba, 1-4.

[10] Christopher D.Manning, Hinrich Schutze, 统计自然语言学处理基础, 苑春华, 李伟等译, 电子工业出版社, 94-116.

[11]: Aristomenis Thanopoulos, Nikos Fakotakis, George Kokkinakis ,Comparative Evaluation of Collocation Extraction Metrics, 2002, 1-6.