Software Engineering and Applications 软件工程与应用, 2015, 4(4), 81-87
Published Online August 2015 in Hans. http://www.hanspub.org/journal/sea
http://dx.doi.org/10.12677/sea.2015.44011

Hans汉斯

Establishment of Relation Markers Collocation Corpus for Compound Sentences Based on Dependency Relations

Beibei Si, Jincai Yang

School of Computer Science, Central China Normal University, Wuhan Hubei
Email: jcyang@mail.ccnu.edu.cn

Received: Jul. 30th, 2015; accepted: Aug. 12th, 2015; published: Aug. 18th, 2015

Copyright © 2015 by authors and Hans Publishers Inc.
This work is licensed under the Creative Commons Attribution International License (CC BY).
http://creativecommons.org/licenses/by/4.0/
Open Access

Abstract

Compound sentences, connecting sentences and paragraph, play an important role in Chinese information processing. The research of relation word recognition is regarded as the breakthrough point for the research of compound sentences. Based on the dependency relationship in Chinese syntax and the characteristics and regularity of relation words and their collocations, this paper recognizes as well as extracts relation words automatically and established the relationship word collocation corpus with CCCS. The collocation corpus records the status of the match and use of various relation words in compound sentences, which will be advantageous to analyze the matching rule of the word collocation rule, and obtain rules for automatic relationship recognition, ultimately lay the foundation for the more accurate identification of the relation word.

Keywords

Compound Sentences, Extraction of Relation Markers, Collocation of Relation Markers, Dependency Relations

基于依存关系的复句关系词搭配库建设

司贝贝，杨进才

华中师范大学计算机学院，湖北 武汉
Email: jcyang@mail.ccnu.edu.cn

文章引用: 司贝贝, 杨进才. 基于依存关系的复句关系词搭配库建设[J]. 软件工程与应用, 2015, 4(4): 81-87.
http://dx.doi.org/10.12677/sea.2015.44011

收稿日期：2015年7月30日；录用日期：2015年8月12日；发布日期：2015年8月18日

摘　要

复句作为联系句子与篇章的桥梁，在中文信息处理中具有重要的地位，关系词的识别研究是复句研究的切入点。本文基于汉语依存句法、关系词及搭配的特征与规律、辅以关系词本体知识库，自动识别并提取关系词，建立了关系词搭配语料库。该关系词搭配库记录了各种关系词在复句中使用与搭配的状态，将有利于分析与统计关系词搭配的规律，从中获取用于关系词自动识别的规则，为关系词更准确的识别打下基础。

关键词

复句，关系词提取，关系词搭配，依存关系

1. 引言

关系词是复句中用来联接分句标明关系的词语，绝大多数具有搭配特性。关系词识别与提取是计算机分析汉语复句的一个重要步骤，目前一些学者对汉语复句关系词进行了相关研究。胡金柱等结合词性标记和关系词搭配理论[1]，提出了关系词抽取的正向选择算法，根据词性标记的特点进行剪枝，并对候选关系词进行过滤，提取正真的关系词。但算法中依赖的常用关系标记集中收录的词语的变化，对关系词识别准确率的影响波动幅度较大；李艳翠等对清华汉语树库的标注进行分析[2]，抽取树库中已标注的复句关系词及关系类别，关系词识别的准确率达 95.7%，但未考虑复句关系词的搭配情况。文献[3]提出了基于规则的连用关系词识别算法，阐述了位置相邻的关系词自动识别解决方法，但未就复句句法特征进行具体的分析，没有考虑上下文语境因素。

本文主要基于汉语依存关系，分析复句中词与词之间的关联关系，挖掘复句关系词特征，归纳总结关系词识别规则。并以 CCCS (the Corpus of Chinese Compound Sentences，汉语复句语料库，华中师范大学语言教育研究中心开发，已收有标复句 65 万余条)中复句为研究对象，利用计算机实现了从大量普通复句中自动识别并提取关系词，建立了大规模关系词搭配语料库。

2. 依存关系

依存语法是 Tesniere 于 1959 年提出的一种结构语法理论。该语法以谓词为中心的语法结构，将动词作为句子的中心词，句子中的其他成分都受这个中心动词支配，所有的受支配成分都以某种依存关系从属于其支配者。图 1 为句子“这是一座闻名中外的大桥”的依存语法分析树。

从图 1 可以看到，依存语法分析的结果没有非终结符，句中词与词之间均直接发生关系，这种关系即为依存关系。它用一条有向弧来表示，弧的方向由支配词指向从属词，弧上的标记叫做关系类型，表示该依存对中的两个词之间存在什么样的依存关系。

目前虽然已经有多种用于句法分析的语法体系，如短语结构语法、格语法、认知语法等，但依存语法直接表示词语之间的关系，不增加额外的语法符号，形式简洁、易于理解，让不同学科的研究者很容易掌握该语法形式[4]。同时，依存语法的描述侧重于反应语义关系，更倾向于人的语言直觉。在自动句法分析中，依存句法结构树这种直观地表达句中词间关系的方法，非常有利于计算机进行信息提取和语义处理。利用依存语法可以方便地构建复句的依存句法结构树，对复句开展句法分析、语义角色标注等

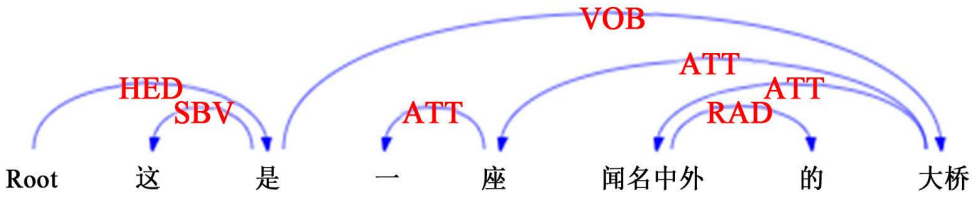

Figure 1. A sample of a dependency grammar tree
图 1. 基于依存语法的分析树例

深层分析，从语法、语义甚至语用角度对语言进行处理。

3. 关系词特征分析

正确识别并提取关系词，需要从两个方面进行研究：一是要把握关系词的基本用法；二是深入分析句间的逻辑关系。通过对上下文语境的分析，提取有用的语言知识，指导关系词识别与提取。

3.1. 依存特征

关系词大多数位于主语或谓语之前[5]。通过观察语料，考察复句句法特点和各种依存关系的分布，以之作为特征，确定句子的主语或谓语，从而确定关系词。应用较多依存关系有 SBV (主谓)、VOB (动宾)和 ADV (状中)关系。在本文所涉及的实验中，由于考察句子的主要成分，因此也采用了 SBV、ADV 和 VOB 这三种依存关系。考虑到词语的稀疏性，将每种关系推广到含有 COO (并列)连接弧的下一级词语，抽取出每个子句成分所对应词的依存关系作为特征之一。

3.2. 位置特征

根据分句定位原则，复句中的关系词分为四类[6]-[10]。对于出现在分句中的第一、二类关系词，通过依存特征进行识别与提取。第三类是出现在分句之间的关系词，用于连接联合复句中的分句，这些词的共同点是可以用逗号或分号隔开；第四类是在几个分句里重复出现的关系词，这些句子大多为表示选择关系的联合复句。因此对于这两类关系词，将句间单独出现或分句中重复出现且词性为连词的词提取出来作为其特征。

3.3. 词性特征

针对显式关系来说，连词和副词是表征句间关系的最主要的依据，因此将句子中的连词和副词提取出来作为其中的特征。

4. 关系词搭配库系统实现

4.1. 系统整体流程与结构

4.1.1. 系统整体流程图

系统整体流程图见图 2。

4.1.2. 主要数据表结构

系统中涉及的主要数据表为关系词搭配库，用于存放从复句中识别并提取出的关系词搭配序列。统计实验表明，有标复句的关系词大多 5 个以内，极少超过 5 个。因此，本文设计的搭配库结构，从复句中第一个关系词起，存放顺序的 5 个关系词，并且前标存放复句中第一个关系词，后标存放复句中最后一个关系词。超过 5 个以上的关系词一律放入备注字段。具体设计见表 1。

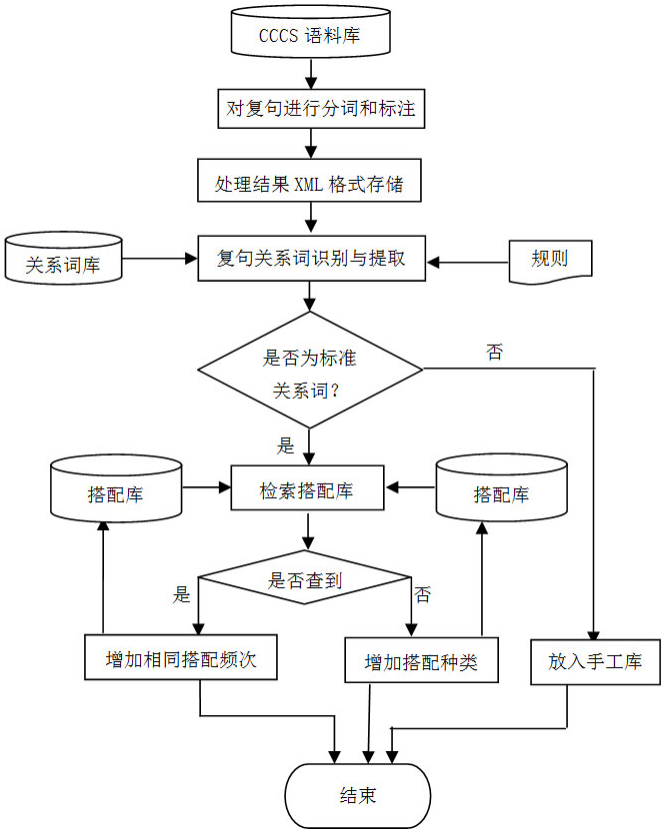

Figure 2. System flow chart
图 2. 系统整体流程图

Table 1. Storage structure of relation markers collocation database
表 1. 关系词搭配库存储结构

属性名称	字段名	数据类型	大小	描述
自动编号	Id	int		主键
前标	front_mark	varchar	20	前关系标记
中标一	mid_mark1	varchar	20	第一个中标
中标二	mid_mark2	varchar	20	第二个中标
中标三	mid_mark3	varchar	20	第三个中标
后标	back_mark	varchar	20	后关系标记
例句	eg_sentence	varchar	500	关系词搭配例句
频次	Frequence	int		关系词搭配频次
备注	Remark	varchar	255	第三个以后的中标

4.2. 系统功能模块结构

本系统功能共划分为复句标注、关系词识别与提取、关系词搭配库可视化展示三个大的模块。复句标注又划分为复句标注和标注结果 XML 格式存储两个子模块；可视化展示模块划分为搭配库展示和手工库展示两个子模块。具体如图 3 所示。

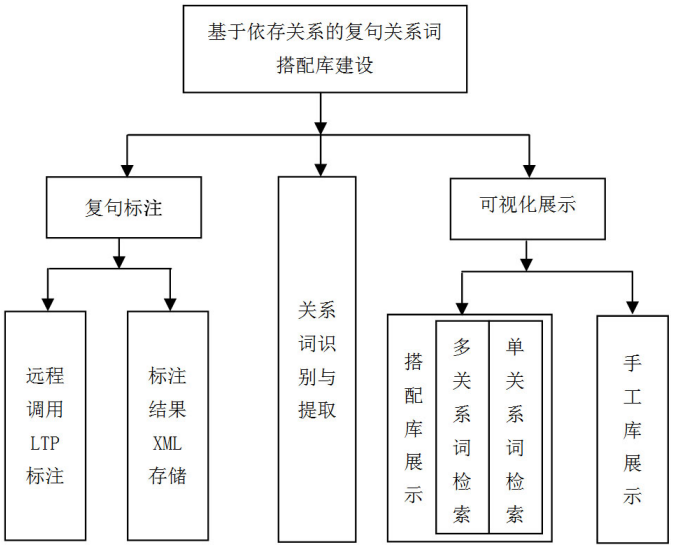

Figure 3. System function structure
图 3. 系统功能结构图

具体功能介绍如下：

(1) 复句标注模块对 CCCS 语料库 60 多万条复句进行分词、词性标注、依存关系标识。在该模块中，系统读取本地 CCCS 语料库中的复句，远程调用哈工大提供的开放的语言分析接口，对复句语料进行标注。系统对成功标注结果以 XML 格式在本地存储，同时对标注异常的情况也进行了相关处理。

(2) 关系词识别与提取模块首先是利用复句关系词的特征，结合关系词本体知识库，将识别出的并在关系词本体知识库中出现的关系词提取出来。其次将获取的关系词搭配序列在关系词搭配库中进行模式匹配，新的关系词搭配模式则收录到关系词搭配库，对相同关系词搭配模式则记载频次。

(3) 可视化展示模块主要对建立的关系词搭配库提供基于任意关系词的检索，展示大规模复句语料库中关系词的搭配模式。

(a) 单关系词检索。检索关系词固定在前标中出现，且在中标、后标中任意出现时的搭配情况及频次。

(b) 多关系词检索。一是检索多个关系词在前标、第一个中标、第二个中标、第三个中标及后标 5 类位置上任意出现时，关系词的搭配情况及频次。二是检索某个关系词在上述五种关系标记中任意出现的情况。

4.3. 实验结果

为了验证本文的方法，我们从 CCCS 中随机选择了 1015 条复句作为样本集，根据上述系统整体流程图(图 2)进行关系词的提取，将关系词提取的结果根据关系词搭配库的表结构(表 1)进行入库操作。经统计，样本集中属于单关系词搭配的句子有 300 条，占整体的 29.5%；属于双关系词搭配的句子有 486 条，占整体的 47.8%，属于三关系词及以上搭配的句子相对较少，不足 250 条。从参与搭配的关系词数量来看，在复句中主要使用单关系词、双关系词和三关系词搭配方式，并且单关系词使用频率较高。

我们用正确率作为衡量分析结果好坏的指标。

正确率 = 1 − (属于样本集复句中的关系词但未提取 + 属于样本集复句中的关系词但只提取部分做为关系词 + 不属于样本集复句中的关系词但提取作为关系词)/样本集中符合搭配模式复句数。

测试结果见表 2。

试验中我们还考察了关系词未被正确识别的复句，其中被错误标注的复句主要集中在一些非典型关

Table 2. Experimental result
表 2. 实验结果表

关系词搭配模式	样本集中符合搭配模式数据(条)	搭配库中符合搭配模式数据(条)	正确率
5 个关系词	10	6	60.0%
4 个关系词	47	25	53.2%
3 个关系词	165	84	50.9%
2 个关系词	486	295	60.6%
1 个关系词	300	222	74%
超过 5 个关系词	7	4	57.1%

Table 3. Distribution of high-frequency words not extracted
表 3. 部分高频词未被提取分布情况

关系词	更	又	即使	以免	尤其是	同时也	是因为
错误句数	68	203	56	78	98	43	83

系词。如将复句中出现的“也、就、并、是、不是”等错误标注为关系词；还有个别关系词语由于存在组合型歧义而被错误切分为两个词，如复句中关系搭配“就是……也”中“就是”被错误切分为“就/是”两个词。对部分高频关系词未识别情况进行统计，结果见表 3。

试验结果表明，双关系词搭配正确率高于多关系词识别，但各类搭配模式的关系词提取正确率并不高，出现漏判、误判的情况，主要原因是规则不够全面、完善，且对关系词的连用和超词没有进行专门处理，造成正确率较低。此外中文分词的效果难以达到 100%的准确率，也会增加一定的影响。

5. 结束语

本文分析了复句关系词的主要特征，提取了关系词搭配序列，建立了关系词搭配语料库。通过大规模的语料库支持，展示了汉语复句中关系词的使用情况，为进一步加强对汉语复句的研究打下了坚实的基础。

本文研究尚存一些不足之处，一是关系词特征提取不够全面，未对关系词连用和超词进行处理，二是未对 CCCS 以外的语料进行分析和研究。下一步拟从以下三个方面进行开展后续研究：(1) 对关系词的构成形式、特征进行进一步研究，扩展、优化关系词识别与提取规则，提高关系词识别的精度；(2) 采用多种形式的复句语料进行试验，拓展规则的应用场景，增强规则的适应范畴；(3) 完善关系词搭配库建设，进一步分析关系词搭配规律，为复句的分析与处理奠定良好的基础。

基金项目

本文得到教育部社科基金(编号 13YJAZH117)，国家社科基金(项目批准号：14BYY093)，国家自科基金(项目编号：31371275)的支持。

参考文献 (References)

[1] 胡金柱, 舒江波, 姚双云, 等 (2009) 面向中文信息处理的复句关系词提取算法研究. *计算机工程与科学*, **10**, 90-93.

[2] 李艳翠, 孙静, 周国栋, 等 (2013) 基于清华汉语树库的复句关系词识别与分类研究. *北京大学学报(自然科学版)*, **12**, 118-124.

[3] 胡金柱, 陈江曼, 杨进才, 等 (2012) 基于规则的连用关系标记的自动标识研究. *计算机科学*, **7**, 190-194.

[4] 王慧兰 (2013) 汉语句类依存树库的构建研究. *北京大学学报*(*自然科学版*), **1**, 25-30.

[5] 李晓琪 (1991) 现代汉语复句中关联词的位置. *语言教学与研究*, **2**, 79-91.

[6] 张仕仁 (1993) 汉语复句的结构分析. *中文信息学报*, **4**, 43-54.

[7] 胡金柱, 吴锋文, 李琼, 等 (2010) 汉语复句关系词库的建设及其利用. *语言科学*, **3**, 133-142.

[8] 向磊 (2014) 基于决策树的汉语复句关系词自动识别中规则挖掘方法研究. 华中师范大学, 武汉.

[9] 姚双云 (2008) 复句关系标记的搭配研究. 华中师范大学出版社, 武汉.

[10] 舒江波 (2011) 面向中文信息处理的复句关系词自动标识研究. 华中师范大学, 武汉.