基于复句语料库的分词系统的研究

杜超华[1]，沈威[1]，姚双云[2]

（1 华中师范大学计算机科学系；武汉；430079 2 华中师范大学语言与语言教育研究中心；武汉；430079）

摘　要：复句在书面语中具有举足轻重的地位，如何让计算机正确理解复句是中文信息处理中一个值得重视的问题。现有的分词系统对复句关系词的正确切分与标注上不足以满足对复句进行层次分析和语义分析的需要。本文建立的分词系统在复句中关系词的切分和标注上做出了必要的改进。

关键词：汉语复句语料库；关系词；分词

Research on a Chinese Word Segmentation System Based on the Corpus of Chinese Compound

DU Chaohua，Shen wei，Yao Shuangyun

（Department of Computer Science; Central China Normal University; Wuhan Hubei; China）

Abstract: Compound sentences occupy very important status in writing language. How to make computers understand compound sentences correctly is a problem for us to think much of in Chinese information procession. The compound relatives' segmentation and tagging of the word segmentation systems in existence can't satisfy the demands of compound sentences hierarchical and semantic analysis. The paper founds a word segmentation system that made some necessary improvement in the segmentation and tagging of compound relatives.

Keywords: Chinese Compound Sentences Corpus; Relative Words;Word Segmentation

1、引言

复句在书面语中占有举足轻重的地位，而且英语中的重句翻译与汉语中的复句有紧密的关系，因而复句的研究是十分有必要的。但针对复句研究的具体要求，现有的自动分词系统却略显不足。为此，本文在现有分词技术的基础上，对复句关系词的切分与标注上进行改进，以适应课题研究的需要。本文首先阐述了有关复句和自动分词方面的知识，最后给出了笔者开发自动分词系统的研究过程。

2、复句相关知识

复句是包含两个或两个以上分句的句子，是自然语言中最普遍的语言现象。要想让计算机真正理解汉语，首先得让计算机理解汉语的句法规律。要真正掌握汉语的句法规律，必须重视复句的句法规律。我们以《人民日报》连续语料制作了一个语料库（该语料主要包括《人民日报》1999 年和 2000 年的全部内容以及 2001 年的大部分内容，均由连续文本组成。），命名为“RMRBLX 语料库”，该语料库包含词数 16082581 个。在 RMRBLX 语料库

基金支持：本课题得到国家重点实验室开放研究基金（SKLSE04-018）和湖北省科技攻关项目（2005AA101C43）资助

作者简介：杜超华：男，(1981-)硕士研究生，主要研究方向：中文信息处理和软件工程 dch@mail.ccnu.edu.cn

中，总共有 846973 个句子，其中单句个数 263792，占总数的 29.1%。复句个数 584181，占总数的 70.9%。[3] 绝大多数的复句，或者分句与分句之间用了特定关系词语，或者分句与分句之间可以用上特定关系词语。特定的复句关系，由特定的复句关系词语标示出来。因此，复句关系词语成了复句在语表形式上的关系标志。大体来说复句关系词有以下四种[1]：

第一，句间连词。它们通常连接分句，不充当句子成分。如“因为、所以、虽然、但是、不但、而且”等等。

第二，关系副词。它们一般既起关联作用，又在句子里充当状语。如“就、又、也、还”等等。

第三，助词“的话”。这是一个表示假设语气的助词，总是用在假设分句末尾，标明分句与分句之间具有假设和结果的关系。

第四，超词形式。它们本身已不是一个词。如“如果说、若不是、不但不、总而言之”等等。

要想正确的理解复句，必须分清复句的类型并划分好复句的层次关系。而这的基础就是正确的切分和标注了复句中的关系词语。

3、现有分词系统不适合复句研究的地方

复句的类别自动划分和层次的标注是复句研究的必要任务。而关系词的正确切分和标注是基础。如复句：你只有意识到这一点，才能更深刻了解我们战士在朝鲜奋不顾身的原因。(《谁是最可爱的人》)

经过免费版中科院自动分词系统 ICTCLAS 切分后结果为：

你/r 只/d 有/v 意识/n 到/v 这/r 一点/t ，/w 才/d 能/v 更/d 深刻/ad 了解/v 我们/r 战士/n 在/p 朝鲜/ns 奋不顾身/i 的/u 原因/n 。/w

因为我是中国人，所以我这么做。切分结果为：

因为/p 我/r 是/v 中国/ns 人/n ，/w 所以/c 我/r 这么/r 做/v 。/w

而事实上，“只有 p，才 q ”，“因为 p，所以 q”分别是复句中典型的条件、因果句式，复句研究中标为连词。为方便复句研究，我们在北京大学的标注系统之上建立关系词类别符号：共 12 类，因果关系词（cyg）、推断关系词(ctd)、假设关系词(cjs)、条件关系词(ctj)、目的关系词(cmd)、并列关系词(cbl)、连贯关系词(clg)、递进关系词(cdj)、选择关系词(cxz)、转折关系词(czz)、让步关系词(crb)、假转关系词(cjz)。并收入关系词 357 个。笔者的分词系统从关系词独立的用法信息和关系词语之间的匹配使用情况入手，在复句语料库 CCCS[1]中提取上下文特征属性，建立体现关系词用法的知识库，进而指导关系词语的识别和消歧。

4、本分词系统的实现过程

4．1 整理词表及相关辅助工作

以前汉语分词的一个困难在于没有一个通用的分词词表。2003 年底清华大学智能技术与系统国家重点实验室制定汉语通用词表。每个词条包括词，词的拼音和词的频率，笔者在此基础上做出以下补充：去掉每个词条的拼音，给词典中每个词条加上词性标注，并将总词频按照训练语料中各词性所占的比例进行词频分配。按照所用编程工具 Delphi 对词排序并以文本文件存储。单独建立一个包含所有关系词以及其标注的文本文件。本文采取的是正反相机械匹配的分词方法与统计的分词方法相结合，并辅以地名人名机构名以及关系词的相关知识和规则对结果进行求精。

4．2 定义

设句子 S 正向反向最大匹配的结果分别为 SEG1 和 SEG2，则

（1）SEG1，SEG2 中，词中长度为一的切分定义为碎片。（不包括标点）

（2）SEG1，SEG2 中，对应的一段连续的不同切分对应的源串定义为一个歧异段。

（3）正向反向机械匹配后得到的最优结果中连续的碎片词以及其前后的词语定义为一个未登陆段。

4．3 分词算法的预处理过程

[1]华中师范大学语言与语言教育研究中心开发的“汉语复句语料库（the Corpus of Chinese Compound Sentences，简称 CCCS）属于汉语专用语料库，CCCS 语料库的建设以小句中枢理论为背景，它将成为汉语本体领域和中文信息处理领域的一项重要资源。

加载词典并初始化相关的数据。对要处理的文本进行预处理。将要处理的文章根据标点分解成句子。从预处理的结果中取出一个句子。采取正反向结合的机械匹配办法进行初步切分。在切分的过程中，还要带上各个词所有可能的词性标注及其对应的频度。未登陆词在标为名词的同时还标记为碎片，以供下一步所用。

如果正反向切分结果一致，则作为正确的切分结果。如果正相切分结果不一致，则以两者中切分碎片词个数较少的作为正确结果。如果正反向切分结果不一致，而碎片词的个数又相同。则：对句中的歧异段进行最大概率法：根据词表把歧义段的所有可能词找出来，然后把所有可能的切分路径找出来，找出概率最大的路径作为输出结果。

4．4 未登陆词和关系词的切分处理

（1）根据规则进行未登陆词（包括人名，地名，机构名）的处理。

人名、地名、机构名相关规则：规则由肯定规则和否定规则两部分组成。绝大多数的未登陆词在初步切分后以分词碎片的形式存在。故在分词碎片中找到潜在的人名地名或者机构名，然后根据肯定规则或者否定规则进行确认。

（2）根据所制定得关系词规则库进行关系词的切分工作。

关系词的规则由切分规则和标注规则两个部分构成，其中每条规则又由两个成分组成：一个为改写规则(rewrite rule)，另一个为触发环境(triggering environment)。

关系词的切分处理模型如图 1 所示。

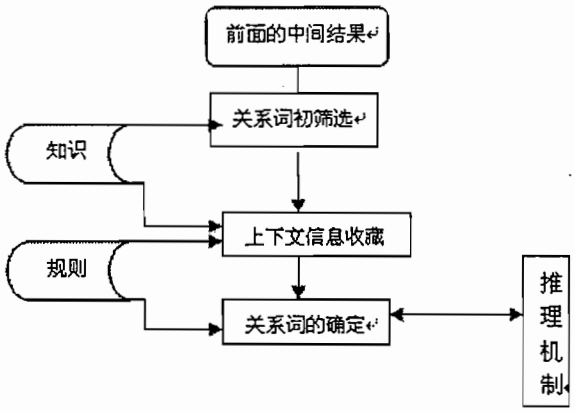

图 1

对于关系词“所以”：切分的改写规则为将“所以”切分为 2 个词。触发环境为：前面的分句没有“因为”且不在分句之首的位置。

例如：语言所以高昂的士气投入到这个项目中来，取得了良好的成绩。

根据北京大学计算语言学研究所开发的一个汉语切分与标注软件的网上测试版。其结果为：语言 所以 高昂 的 士气 投入 到 这个 项目 中 来 ， 取得 了 良好 的 成绩 。

采取规则后正确的切分结果为：语言 所 以 高昂 的 士气 投入 到 这个 项目 中 来 ， 取得 了 良好 的 成绩 。

4．5 词性标注

词性转移矩阵为：从某一词性标记转换另一词性标记的概率矩阵 $P=(P_{ij})$（其中 i,j 分别为 2 个标记在标记系统中的索引。根据训练语料库得到词性转移矩阵：

$$P=\begin{bmatrix} P_{00} & P_{01} & P_{02} & \cdots \\ P_{10} & P_{11} & P_{12} & \cdots \\ \cdots & \cdots & \cdots & \end{bmatrix}$$

$$P_{ij}=\frac{\text{标记 i 与标记 j 同现次数}}{\text{标记 i 的出现词数}}\times 100\%$$

①根据 VOLSUNGA 算法选取词性标记路径，得到词性标记结果。

此处因为词性转移概率是一个很小的正数，如果词数比较多，最后得到得各种标记串的概率都接近于零，无法在机器上表示出来，当然也无法比较大小。解决的办法是求转移概率的对数之和，把乘法变成加法。转移概率

的对数都是负数，对其结果取反则变成正数。

第 i 个词的词性 j 转换到第 i+1 个词的词性 k 的费用记为：

$Fee(i,j,i+1,k) = -\log_{10}(P_{jk}) - \log_{10}(P_k)$ 其中 P_k 为第 i+1 个词词性为 k 的概率。

根据动态规划的算法可在线性时间(Linear Time)求出句子的总费用最小的标注串即为最佳标记串。[7]

②根据关系词的标记规则对标记串进行扫描修正，并根据结果对规则库进行补充和修正。

5．结束语

本分词系统主要是针对复句研究的需要，在现有的分词方法的基础上，通过规则对复句关系词的识别和标注上求精。经统计，关系词的正确切分和标注准确率达到 85%左右。复句信息工程是一个长期而艰巨的任务，而对复句中关系词的正确切分和标注是复句分类，复句分层等后续工作的基础。只有基础打的扎实，才能给后面的处理工作减少麻烦。对关系词的切分与标注的求精还不够彻底。我们将对此进行进一步的研究。

参考文献

[1]邢福义．汉语复句研究．北京：商务印书馆，2001

[2]陈小荷．现代汉语自动分析—Visual C++实现．北京语言大学出版社，2000

[3]姚双云．小句中枢理论的应用与复句信息工程．《汉语学报》第 4 期．2005

[4]刘颖．计算语言学．北京：清华大学出版社．2002

[5]刘开瑛．中文文本自动分词和标注．北京：商务印书馆，2000

[6]赵伟; 戴新宇; 尹存燕; 陈家骏． 一种规则与统计相结合的汉语分词方法． 计算机应用研究 2004 年 03 期

[7]邓宏涛．中文自动分词系统的设计模型．计算机与数字工程 2005 年 04 期：138-140