一种切词和词性标注相融合的汉语语料库多级加工方法*

周强　　俞士汶
北京大学计算语言学研究所　北京 100871

[摘要] 本文通过深入研究汉语切词和词性标注处理的内在联系，提出了一种将两者结合起来处理的方法，并详细介绍了词性标注的基本设计思想，讨论了规则方法和统计方法相结合的排歧策略。用此方法对约 40 万字的语料进行了实际的切分和标注处理，其准确度分别达到了 96%和 94%。

A Chinese Language Corpus Processing Method combining Segmentation with Tagging

Zhou Qiang, Yu Shiwen
Institue of Computational Linguistics, Peking University, Beijing, 100871

ABSTRACT : In this paper, we describe a method that combine Chinese language segmentation with category tagging process. We introduce the basic idea of category tagging , discuss a grammatic category disambiguation scheme that integrates rule-based technique and statistics-based technique. Experiment results proved that the overall accuracies are 96% (for segmentation) and 94% (for tagging).

一　引言

在汉语中，短语(与词组同义)类型，是一种很特殊的语法形式。它具有与汉语句子结构相类似的主谓、述宾、联合等种种不同的构造方法，这和英语中的短语有很大差别。“如果我们把各类词组的结构和功能都足够详细地描写清楚了，那么句子的结构实际上也就描写清楚了，因为句子不过是独立的词组而已。”[1]从这个意义上看，现代汉语短语结构研究的重要性是不言而喻的。

要进行汉语短语结构的研究，就必须对大量的汉语语言事实进行调查研究，从中总结出有关短语组成的规律。在这方面，一个带多种标记的大规模汉语语料库就可以提供大量有用的信息。从本质上看，汉语语料库的多级加工，即从文本语料库('生'语料库)，经过自动切词，词性标注及短语结构标注而得到带不同标记的'熟'语料库，和汉语短语分析处理中的切词，词性标注，短语结构分析等各个阶段有着内在联系，完全可以把两者结合起来处理。一方面，利用短语分析中的各种技术对语料库进行多级自动标注，形成准确度较高的带多标记的语料库；另一方面，以新的语料库为基础，利用不同的统计工具，从中提取出大量的汉语语言事实，通过总结提炼，再结合到自动分析程序中，可以大大提高自动分析的效率和准确性。正是基于这个基本思路，笔者实现了一个汉语短语分析和语料库多级标注处理相结合的系统。

从结构上看，此系统分为两大处理系统：切词及词性标注处理和短语结构分析。本文主要介绍一下切词和词性标注处理的基本思想和相关的处理技术。有关短语分析的处理将另行撰文介绍。

在下面的几节中，第二节介绍了切词和标注相结合的处理策略，第三、四节讨论了词类标

* 本课题受国家自然科学基金委员会资助，项目号为 69073339.

注的基本设计思想及规则和统计相结合的排歧策略，最后给出了系统的一些实验结果。

二　切词和标注相结合的处理

汉语自动切词的基本原理是字符串匹配。为了提高切分精度，往往需要采用大量的知识以解决切词过程中的歧义切分问题。对此，许多研究人员进行了大量的工作，提出了一些有用的算法，如：采用词尾词构词检错/纠错的技术[2]，联想——回溯算法[3]，“生成——测试”方法[4]，及专家系统控制方法[5]。另外还有利用语料库统计技术和建立概率模型来校正切分歧义的方法[6][7]。

从处理效果看，这些系统的切分精度一般都达到了95%左右，具有很强的实用性。但由于目前切词系统的处理基本上是在词的一级上进行的，没有考虑大规模的语料库标注问题，缺乏一种从词类标注处理层次对切词系统切分精度的客观评价，因此给切词系统功能和性能的进一步提高带来了一定的局限性。

笔者在对约四十万字的语料进行的切分和标注实践中，发现在切词过程中使用词类信息会带来许多好处，概括起来，主要有以下几点：

1). 利用歧义切分字段中的不同切分词的词类组合关系及上下文词类信息，可以解决绝大部分切分歧义现象。

我们知道，在汉语自动切词过程中，会碰到两种不同类型的歧义切分现象，即交集型歧义和组合型歧义。

考虑交集型字段 S＝ABC，它有两种可能的切分结果：AB＋C 和 A＋BC，从而形成两种词类组合：$C_{AB}+C_C$ 和 C_A+C_{BC}，而它们在特定的语言环境下出现的可能性是不一样的。这样我们就可以依据词类共现频度的高低，并参照前后词的相关信息，从中选择出正确的切分结果。

而对组合型字段 S＝AB，由于它本身不提供正确切分的特征信息，因此只有通过考察它与其前趋或后继词之间的关系，才能确定正确的切分。在这方面，前后词的词类信息和组合型歧义字段中两个词的词类组合关系起了很重要的作用。

2). 有助于利用汉语构词法构造新词，解决一部分未登录词的处理问题。

对于汉语自动切词，尽管“分词规范”中给出了切词单位的概念，但它对机器如何辩识这样的单位并没提供实际的帮助，因此研究和借鉴汉语构词法的研究成果，对于我们确定合适的切分单位，特别是将一些未登录词正确地切分出来，还是很有意义的。陆志韦先生在[8]中总结了一些汉语词的常见构词格，如，名词的常见偏正式构词格有：

i). 单音节的名词＋单音节的名词：　铁路、马车、牛肉

ii). 单音节的名词＋双音节的名词：　手指甲、牛鼻子

iii). 双音节名词＋单音节名词：　电流表、热带鱼

iv). 双音节动词＋单音节名词：　证明信、消防队

从中我们可以总结出许多有用的基于词类组合的构词规则，把它们应用于切词处理，可以达到以较小的切词词典取得较好的切分效果的目的。因为词典过于庞大，也会产生歧义过多的负面效果，这一点也已为研究者所认识，因此我们希望将词典控制在一个适当的规模上。

3). 有助于发现切词错误。

在汉语里，某些词类组合出现的频度是很低的，如：d＋n＋句尾标记，v＋u＋d＋句尾标记。因此，如果在切词过程中切分出了这样的词类组合，我们就有理由认为切词处理可能出错了。在下面的切分实例中，就存在着这样的错误：

i). 买/v　了/u　一头/d　牛/n　。/w

ii). 他/r　球/n　打/v　得/u　最好/d　。/w

由此可见，利用词类信息，可以为机器自动检测切词错误提供一种有力的手段。

正是基于以上的种种认识，笔者在对大规模的汉语语料的加工处理过程中，采用了一种切词和标注相结合的方法，其基本处理流程为：

i). 利用一个带词类标记的切词词典，完成自动切词，并给每个切分单位标上所有可能的词类标记。

ii). 对切词结果进行基本的构词法处理，如：词缀合并，重叠形式组合等。

iii). 通过词类排歧，完成词类的自动标注。

iv). 利用构词规则，通过合并，发现一些符合汉语构词规律的未登录词并确定其词类。

v). 检测切分结果的词类组合关系，发现一些可能的切分错误，返回切词阶段进行回溯处理。

另外，为了保证工作的规范性，笔者正在着手制定一个切词和标注同时进行的工作规范，以"信息处理用现代汉语分词规范"和北大的现代汉语词语分类体系[9]为基础，描述了 26 个大类词的切分和标注标准，并对其中的名词、动词、形容词等词类的构词规则进行了详细的讨论，列出了一些基本的构词情况，以期在以后的研究过程中保持处理的一致性。

三　词类标注的基本设计思想

在对语料库信息的加工处理过程中，词类标注是一项很重要的工作。它的任务就是给语料库中的每个词赋一个合适的词类标记。由于自然语言中存在着大量的词的兼类现象，因此给语料库的自动词类标注带来了很大困难。

国外在建立了 Brown 语料库和 LOB 语料库后，从六十年代起，一些学者对英语语料库的自动词类标注进行了研究，提出了一些不同的排歧技术。其中比较著名的是基于规则的 TAGGIT 系统和基于统计处理的 CLAWS 和 VOLSUNGA 算法。[10]

国内对汉语词类标注进行研究的主要有清华大学和山西大学。他们的基本处理思路是：首先对数万字的语料进行人工标注，然后通过统计，从中提取出带词类标记的词频统计表和词类共现频度矩阵，利用其中的信息，通过建立概率计算模型而完成词类自动标注[11,12]。

与以前的词类自动标注系统相比，我们现在的处理又有自己的特点。其基本设计思想，概括起来主要有以下几点：

1). 以带词类标记的词典为基础

我们现在的初始词类标注是在切词过程中，通过使用带词类标记的切词词典而完成的。其词类信息主要来源于北大的"现代汉语语法电子词典"[13]。由于这些信息是依据朱德熙先生的"按照词的语法功能进行分类"的标准，由语言学家亲自审定和填写的，因此具有很高的准确度。利用这些信息进行初始词类标注，就可以很好地保持标注结果的一致性，为以后进一步排歧处理打下很好的基础。

2). 使用了较小的标记集

在我们目前的研究中，对词的词类标注还只限于基本词语分类描述，计有 a～z 共 26 个词类标记(详见下表)：

名词	n	区别词	n	助词	u	简称略语	j
时间词	t	形容词	a	语气词	y	习用语	l
处所词	s	状态词	z	象声词	o	语素	g
方位词	f	动词	v	叹词	e	字	x
数词	m	副词	d	前接成分	h	标点符号	w
量词	q	介词	p	后接成分	k		
代词	r	连词	c	成语	i		

同时，为保存人工校对中发现的一些新信息，对一些词典中未登录的专有地名、专有人名等给出了子类标记ng(专有名词)，ngp(指人的专有名词)。再加上表明语素字性质的标记：Ng(名词性语素字)、Ag(形容词性语素字)，Vg(动词性语素字)，组成了由31个标记组成的词类标记集。

较小的标记集，可以降低系统处理的复杂度。同时，也可以使我们把精力主要集中于那些最有可能出现的歧义组合上，从而有助于提高系统处理的准确度。

3). 标注结果与词典信息相结合，形成一个立体知识库

尽管我们现在的标记集很小，但通过和语法电子词典相结合，可以很方便地扩大标记集的规模。这是因为在我们的电子词典中，对每类词的语法属性都由语言学家进行了相当充分的发掘，以期尽可能全面地揭示这类词的语法功能和分布情况。从广义上看，这些语法属性描述也可以看作是分类信息。如果我们以词及词类信息为关键字检索电子词典，就可以得到这个词的详细语法属性信息，这样，以标注结果为一个平面，而以各词的语法属性表作为纵深，形成了一个立体的知识库。根据不同的处理要求，我们就可以利用语法属性信息，给语料库中的词语标上不同的词性标记。另外，也可以利用这立体知识库，进一步进行短语及句子结构分析。实际上，本系统中的短语结构分析正是在此基础上进行的。

4). 规则排歧和统计排歧相结合

考虑到规则方法和统计方法各有优势，笔者对语料库的词类标注采用了规则方法和统计方法相结合的策略。首先，通过对大量语料的统计分析，找到出现频度最高的歧义现象，考察它们所在的不同语言环境，从中提取出一些规则模式，以排除那些最常见的、语言现象比较明显的歧义现象。然后，利用从人工校对过的语料中统计得到的信息，构造概率模型以解决出现频度较低的歧义组合及进行未登录词的词性推断。

而在实际处理过程中，两种方法的使用又有不同的侧重点。最初由于没有一个大规模的、带准确词类标记的语料库，因此只能以规则方法先标注一小部分语料，然后通过人工校对，发现和改正其中的错误，调整规则库内容。再用新调整的规则和从中统计得到的数据，使用规则和统计方法标注一部分新语料，如此，"滚雪球"似的一步步扩大处理语料的数量。随着语料库规模的不断扩大，将使规则描述越来越准确，统计信息越来越全面，从而可以充分发挥统计处理的优势，减少人工干预，达到两种方法的最好融合。

四　自动词类标注的排歧策略

4.1 规则排歧处理

规则处理的基本思路是利用上下文信息以确定一个多类词在特定的语言环境下到底应标上什么词类标记，其基本模式是上下文框架规则。在实际应用过程中，为提高工作效率，又把规则处理分为以下三个不同的阶段：

i). 特征词排歧：

对于那些出现频度很高，词类标记又较多的词(如："一"，"着"，"了"，"过"，"把"，"来"，"好"，"就"，……)，设置特定的规则，判别这些词在句子中出现的上下文环境，以确定不同的标注情况。这相当于词定位的排歧方法。

ii). 特定多词类排歧：

统计语料中经常出现的一些多词类组合，如：v－q，p－v，v－n，q－n，v－d，a－v等，调查这些多词类组合在不同语言环境下选取某个词类的可能性大小，大量使用词类划分的语法功能特征，特别是某类词区别于其它类词的自己所特有的语法分布信息，构造相应的上下文规则描述，以选择正确度较高的词类标记。这相当于多词类定位的排歧方法。

iii). 上下文关系排歧：

设置一些上下文词类标记匹配模式，描述在一定的词类环境下可能出现的词类标记的集合。在实际处理过程中，将句子中出现的多词类串的前后词类描述与这些规则模式相匹配，将

多词类集与规则描述中可能出现的词类集合进行交运算。若所得结果只有一个词类标记，则排歧成功；否则继续进行，直至不能匹配为止。这相当于模式定位的排歧方法。

4.2 统计排歧处理

经过规则排歧，对语料中的大部分多类词都赋予了一个较为正确的词类标记。但由于规则的不完全性，对于某些多类词及未登录词还不能处理，这就需要采用一种新的排歧和推断技术——统计方法加以处理。

首先，构造如下的统计计算模型：

令 $W=W_1W_2\cdots W_n$ 为一多词类词串，$C=C_1C_2\cdots C_n$ 为可能的词类标注结果串。$P(C|W)$为给定 W 条件下 C 出现的概率。如果不考虑更大的上下文，我们可以认为使得$P(C|W)$的值取得最大时的 C 出现的可能性最大。这样就把词类标注问题转化为寻找一组标记串 C′，使得：

$$P(C'|W)=\max_{C'\in C}P(C|W)$$

利用二元语法 Bi－gram 对此公式进行简化，可得：

$$P(C|W)=\prod_{i=1}^{n-1}P(C_{i+1}|C_i)\cdot P(C_i|W_i)$$

然后，利用经过机器处理和人工校对的一部分正确标注语料进行统计，得到词类标记共现频率矩阵及每个多词类词出现特定词类标记的频度统计结果，从中可以计算得到近似的 $P(C_i|C_j)$及 $P(C_i|W_i)$。

在实际处理时，应先对未登录词的词性进行推断。这种推断主要利用了前后词的词类信息。通过选择它与前接词类及后接词类共现概率最大的各两个词类作为此未登录词的词类标记串。这样就把未登录词现象转化为对多类词的处理问题。

对于一串多类词（以单词类标记词为头和尾，若此串在句首或句尾出现，则加上句首标记 ^ 或句尾标记 $），如果我们把每个词的不同词类纵向排列，而把不同的词横向排列，以词与词类标记组成的记录为节点，该节点上标上 $P(C_i|W_i)$，相邻词节点间通过标有 $P(C_i|C_j)$的弧相连。这样就形成了一个有向图，利用类似 VOLSUNGA[17]的算法搜索有向图，得到一条最佳路径，其上的概率乘积达到最大。这样就可以完成对多词类词串的词类自动标注。

五　切词和标注实验结果分析

利用上面介绍的方法，笔者对约 40 万字的动词用例语料库进行了自动切词和自动词类标注的处理，并对所有的机器处理结果认真地进行了一次人工校对。假定人工校对的结果是百分之百正确的，以此为基础，通过将人工校对结果与机器自动处理结果相比较，记录下所作修改的次数，就可以对系统的自动切词及自动词类标注的准确度有个粗略的估计。

表 5.1 列出了有关的比较结果及不同错误率计算结果，从中初步估计，系统的切分精度约为 96% ，标注准确度约为 94% 。

a). 比较结果记录：		b). 语料库词总数：	295522
1. 词组合情况：	3672	c). 不同错误率：	
2. 词交错情况：	492	1. 输入错误率：	0.108%
3. 词分解情况：	2107	2. 切分错误率：	2.122%
4. 词输入错误：	318	3. 标注错误率：	5.446%
5. 词类改变情况：	10132	*. 错误修正率：	7.676%
6. 词类增加情况：	1730		
7. 词类选择情况：	4232		
*. 总计修改次数：	22683		

表 5.1　两种语料结果的比较

六 结束语

本文介绍了一种切词和标注相结合的处理方法。概括起来，它主要有以下特点：

1). 通过应该带词类标记的切词词典将切词和标注联系起来。

2). 切词过程完成初始标注，而标注过程又对切词结果进行检测，通过信息的不断反馈，提高了最终处理结果的准确性。

其中，切词过程中通过采用汉语构词法的研究成果，达到了以较小的切词词典达到较好的切分效果的目的。而在标注过程中，则采用了规则方法和统计方法相结合的排歧策略，取得了很好的处理效果。

从对四十多万字的语料的实际处理结果来看，本方法的应用基本上还是成功的。在以后的研究中，还需进一步完善此方法，同时，通过对大量机器处理语料的分析，发现并改进系统设计中的一些缺陷。在此基础上，扩大语料的规模，争取近期内达到一百万词次的量级。

支持和参加此项研究的有陆俭明、朱学锋、王正秦、姚剑、姜新、郭锐、张芸芸、王惠等同志，在此致以诚挚的谢意。

参 考 文 献

1. 朱德熙，《语法答问》，商务印书馆，1985
2. 梁南元，“书面汉语自动切词系统－CDWS”，中文信息学报，87.2
3. 李国臣、刘开瑛、张永奎，“汉语自动切词及歧义结构的处理”，中文信息学报，88.3
4. 黄祥喜，“书面汉语自动切词的‘生成—测试’方法”，中文信息学报，89.4
5. 徐辉、何克抗、孙波，“书面汉语自动切词专家系统的实现”，中文信息学报，91.3
6. 黎邦洋等，“一种主要使用语料库标记进行歧义校正的，最大匹配汉语自动切词算法设计”，Proceedings of ROCLING Ⅳ，P135－146
7. 张俊盛等，“限制式满足及机率最佳化的中文断词方法”，Proceedings of ROCLING Ⅳ，P147－165
8. 陆志韦等，《汉语的构词法》，科学出版社，1964年
9. 俞士汶，“信息处理用现代汉语词语分类体系”
10. Steven J. DeRose, "Grammatical category disambiguation by statistical Optimization", CL, Vol 14, P31－39
11. 刘开瑛、郑家恒、赵军，“语料库词类自动标注算法研究”，机器翻译研究进展，1992，P378－386
12. 白栓虎、夏莹、黄昌宁，“汉语语料库词性标注方法研究”，机器翻译研究进展，P408－418
13. 俞士汶、朱学锋、郭锐，“现代汉语语法电子词典的概要与设计”，ICCIP92，P186－191