ISSN1000-9825 Journal of Software 软件学报 2000,11(2):195~201

汉语最长名词短语的自动识别*

周强 孙茂松 黄昌宁

(智能技术与系统国家重点实验室 北京 100084)

(清华大学计算机科学与技术系 北京 100084)

E-mail: zhouq@s1000e.cs.tsinghua.edu.cn

摘要 通过对包含5 573个汉语句子的语料文本中的最长名词短语的分布特点的统计分析,提出了两种有效的汉语最长名词短语自动识别算法:基于边界分布概率的识别算法和基于内部结构组合的识别算法.实验结果显示,后者的识别正确率和召回率分别达到了85.4%和82.3%,取得了较好的自动识别效果.

关键词 最长名词短语,边界识别,句法分析.

中图法分类号 TP18

在自然语言句子的理解过程中,能否准确地识别其中的名词短语(np)起着很重要的作用.按照认知科学的观点,人类必须首先识别、学习和理解文本中的实体(entity)或者概念(具体的或抽象的),才能很好地理解自然语言文本.而这些实体和概念大都是由文本句子中的名词短语所描述的.因此,如果我们掌握了文本中的名词短语,就可以在很大程度上把握文本所表达的主要意思.

从组成结构上看,句子中的名词短语可分为以下3类:(1) 最短名词短语(minimal noun phrase,简称mNP):不包含其他任何名词短语的名词短语;(2) 最长名词短语(maximal noun phrase,简称MNP):不被其他任何名词短语所包含的名词短语;(3) 一般名词短语(general noun phrase,简称GNP):所有不是mNP和MNP的名词短语.从mNP到GNP再到MNP,自动识别的难度是在不断增加的.而在自然语言处理领域中,MNP的自动识别具有更为重要的意义.因为我们一旦很好地识别出了句子中所有的MNP,就可以很方便地把握句子的整体结构框架,从而很快构建出句子的完整句法树(森林).

正是认识到了这一点,近几年来,许多研究人员在MNP的自动识别方面进行了许多有益的探索,提出了一些行之有效的识别方法.在英语方面的工作主要有:(1) Church的简单名词短语抽取器[1],利用概率矩阵信息来确定句子中np的起始和终止位置.(2) Bourigault的术语抽取器LEXTER[2],通过构造两个阶段的自动分析器发现文本中的术语(即部分MNP).(3) Voutilainen的MNP获取工具:NPTool[3],利用两种有限状态分析机制(NP-否定机制和NP-肯定机制)来发现文本中可能的MNP.(4) Kuang-hua Chen等人的工作[4],利用统计分块(chunking)和有限状态分析相结合的方法来发现句子中的名词短语.

英语MNP自动识别的难点在于解决各个成分之间的联结(attachment)关系歧义.相比之下,汉语MNP的识别则更为困难,这是由汉语句法成分特有的套叠现象[5]所决定的.与英语不同的是,汉语中的任何句法成分都可以不经过任何形态变化,只需加上一个结构助词"的",就可以充当一个np的定语(当然,前提是两者之间可以存在修饰和被修饰的关系)而形成一个更长的np.这就大大增加了汉语MNP自动识别的难度.

从这几年来的一些研究实践来看,自动处理效果并不是很理想.主要的研究工作包括:(1) 李文捷等人[6]利用边界分布信息构造概率模型而进行的MNP自动识别实验,其开放测试的识别正确率达到了71.3%(在30篇

* 本文研究得到国家自然科学基金(No.69705005)和中国博士后科学基金(No.97005)资助.作者周强,1967年生,博士,副研究员,主要研究领域为计算语言学.孙茂松,1962年生,副教授,主要研究领域为中文信息处理,计算语言学.黄昌宁,1937年生,教授,博士生导师,主要研究领域为计算语言学,中文信息处理.

本文通讯联系人:周强,北京 100084,清华大学智能技术与系统国家重点实验室

本文1998-11-10收到原稿,1999-03-09收到修改稿

新闻报道语料中).(2) Angel S. T. Tse 等人[7]利用统计和规则相结合的方法,构造了"的"字名词短语自动分析器.实验结果为:识别正确率为 75%,召回率为 90%(在 15 篇汉语文本中).

本文提出了一种自动识别汉语 MNP 的新方法.它在对输入文本进行组块分析预处理的基础上,通过充分利用 np 边界分布信息和 np 内部结构组成知识,构造形成了高效的汉语 MNP 自动识别器.在约 7 万词的汉语语料上进行的 MNP 识别实验显示,正确率达到了 85.4%,召回率为 82.3%,取得了较为令人满意的识别效果.

1 最长名词短语的自动识别

图 1 给出的是我们进行汉语最长名词短语自动识别的基本流程图.以经过正确切分和词性标注处理的汉语句子作为分析器输入,MNP 的自动识别主要分两个阶段进行.首先,对输入句子进行分层次的预处理,包括自动发现一些特殊的成分组,如标点分隔结构、并列结构、固定搭配结构等,并在各个成分组内部及成分组之间进行词语块的成分边界预测,即确定每个词语是处于成分的左边界、右边界还是中间位置.然后,通过构造不同的 MNP 识别算法,准确地确定其中哪些成分边界是 MNP 的左右边界.下面给出了一个具体的分析实例.

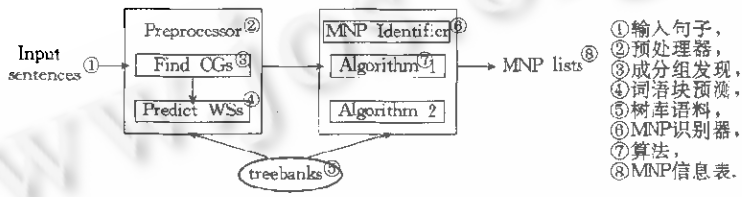

Fig. 1　Overview of maximal noun phrase identifier
图1　MNP自动识别的基本流程图

(a) 输入句子:我/r 爸爸/n 会/v 做/v 书架/n,/w 这个/r 书架/n 是/v 爸爸/n 去年/t 做/v 的/u。/w.

(b) 预处理结果*:{{我/r 爸爸/n}[会/v[做/v 书架/n},/w{这个/r 书架/n}[是/v[爸爸/n[去年/t 做/v]的/u}。/w}.

(c) MNP 识别结果(应识别出 3 个 MNP):{$_{MNP}$我/r 爸爸/n} 会/v 做/v 书架/n,/w{$_{MNP}$这个/r 书架/n}是/v{$_{MNP}$爸爸/n 去年/t 做/v 的/u}。/w.

在下面的几节中,我们将对有关的内容进行详细的说明.

1.1 MNP 识别的预处理

对输入句子进行预处理的目标,是将输入的〈词语,词类〉对的线形序列转化为包含成分边界信息的组块序列,从而为进一步的句法分析,如识别句子中的 MNP,构建句子的句法结构树提供有力的支持.

假设输入句子 $S=(W,T)$,其中 $W=w_1,w_2,\ldots,w_n$ 为句子的词语串,$T=t_1,t_2,\ldots,t_n$ 为各词语相应的词类标记串,则预处理过程应得到这样的组块描述序列:$CH=\{CG,WS\}$,其中 $CG=\{cg_{ij}\}$,表示若干从词语位置 i 开始到词语位置 j 结束的成分组,$WS=ws_1,ws_2,\ldots,ws_n$,表示标注了成分边界预测信息的词语块序列,其中 $ws_i=\langle w_i,t_i,bp_i\rangle$,$bp_i$ 可取值 0,1,2,分别表示此词语处于成分的中间位置、左边界和右边界.有关成分组和词语块的详细定义可参阅文献[8]中的汉语组块分析标注体系描述.

在我们目前的组块分析体系中,主要定义了以下几个成分组:

(1) 标点分隔结构(punctuation seperated structure,简称 PS),如{《鲁迅 全集 》}.

(2) 固定搭配结构(collocation structure,简称 COS),如{当老师 进来 的 时候}.

(3) 并列结构(CS)及并列成分(conjunction constituent 简称 CC),如{{发现 人才}、{培养 人才}和{使用 人才}}.

它们一般具有这样的句法特征:(1) 成分组内部的成分只与成分组内部的其他成分发生句法作用.(2) 整个

* 其中的大括号对"{}"表示发现的一个特殊成分组,而中括号"["和"]"则分别表示此词语处于某个成分的左边界和右边界.

成分组作为一个整体与句子中的其他成分发生句法作用.通过利用一些特殊词语项,如上面例子中加下划线的词项,或一些特殊的结构特征,如并列结构的并列成分之间存在着较强的内容相似性等,可以构造出简单而有效的识别算法来自动发现句子中所有可能的成分组,有关的详细算法可参阅文献[9].

而对于词语块中的成分边界信息,则可以通过构造如下的统计模型来进行自动预测:考虑词语序列(W_{ij},T_{ij}),选择合适的成分边界标记序列$BP_{ij}=bp_i,bp_{i+1},\ldots,bp_j$,使得$P(BP_{ij}|W_{ij},T_{ij})$达到最大.文献[10]给出了具体的预测算法.

1.2 MNP 自动识别器

MNP 自动识别器的处理目标是在成分边界预测信息的基础上,准确地确定句子中所有 MNP 的左右边界.利用不同的树库统计知识,我们构造了两种识别算法:(1) 基于 np 边界分布概率的识别算法(算法 1).(2) 基于 np 内部结构组合的识别算法(算法 2).

1.2.1 基于 np 边界分布概率的识别算法

首先,对于经预处理得到所有成分左右边界,进行以下处理:

- 如果$P(np_L|t_{i-1},bp_i=1,t_i)>\alpha$,则设置此边界为可能的 MNP 左边界.
- 如果$P(np_R|t_i,bp_i=2,t_{i+1})>\beta$,则设置此边界为可能的 MNP 右边界.

其中两个概率阈值α和β的设置需要兼顾以下两个目标:(1) 尽可能排除统计噪声的影响,即不至于发现许多无意义的 MNP 边界;(2) 尽可能保留真正的 MNP 边界.参照实验语料中的 MNP 左右边界的概率分布信息,我们选择了以下两个概率阈值:$\alpha=0.05,\beta=0.15$,基本上达到了以上目标.

为了确定以上发现的哪个左右边界对可以形成真正的 MNP,我们设计了一个 NP 栈结构.在此基础上实现了以下的 MNP 组合算法,即顺序处理所有可能的 MNP 边界,若是左边界,则压入 NP 栈中,否则,检查它是否能与栈顶元素组合,若是,则弹出栈顶元素,并组合形成一个可能的 MNP,此过程不断进行,直至不能组合为止;将形成的 MNP 压入 NP 栈中,继续处理下一个 MNP 边界直至句子结束.最后,通过检索 NP 栈,就可以输出所识别出的所有 MNP.

1.2.2 基于 np 内部结构组合的识别算法

从实验结果(见下一节)来看,仅仅依据 np 边界分布概率,MNP 自动识别效果并不是很理想.为此,我们设计了一个新的 MNP 识别算法(算法 2),其基本思路是利用 np 内部结构组合知识构造自底向上的 MNP 部分分析器.

在文献[11]中,我们曾提出一种基于括号匹配原理的汉语句法分析方法,它在经过成分边界预处理的输入句子上,通过括号匹配操作发现句子中所有可能的句法成分,以形成输入句子的完整分析树(森林).通过对这一算法的适当简化,并充分考虑 MNP 的结构分布特点,我们构造了一个新的 MNP 识别算法,其基本内容如下.

背景知识:句法结构归约规则:{结构组合→{句法标记}}

基本操作:

① 组合基本成分(bc)

若词语块序列$ws_i,\ldots,ws_j$满足以下条件,则它们可形成一个基本成分:

(a) $bp_i=1,bp_j=2$

(b) $\forall\ ws_k,k\in(i,j)$,有$bp_k=0$

② 发现可能的 MNP 右边界成分

按照 MNP 的结构分布规律,具有以下特征的成分可能成为 MNP 的右边界:

(a) 词语块$ws_i=$(“的”,$uJDE$,2)

(b) 词语块$ws_i=$(*,n,2)

(c) 被归约为 np 的基本成分

(d) 其他可能成为 MNP 中心成分的词语块或基本成分

③ 向左扩展组合 MNP

从可能的 MNP 右边界成分出发,不断与其左相邻成分组合形成新的 MNP,直至不能组合为止.

主控结构:NP 栈(类似算法 1)

基本流程:从左向右扫描整个句子,顺序执行基本操作①、②和③,直至句子结束.

2　实验结果分析

我们的 MNP 识别实验采用了约 9 万字规模的树库语料作为测试样本.利用其中正确标注的 MNP 作为评价标准,可以方便地对自动识别结果进行评估,从而可以对 MNP 自动识别器的处理性能有一个客观而全面的认识.

2.1　预处理结果分析

对于两种不同层次的预处理结果,我们定义了以下几个评价指标:

(1) 成分组边界正确率(constituent group precision,简称 CGP)=具有正确的边界位置的成分组总数(Cort_CG)/识别出的成分组总数(sum of identified constituent group,简称 SCG).

(2) 成分组边界交叉率(constituent group crossing ration,简称 CGR)=与树库成分交叉的成分组总数(Crossed_CG)/识别出的成分组总数(SCG).

(3) 词语块界定预测正确率(word stem precision,简称 WSP)=具有正确成分边界预测的词语块总数(Cort_BP)/语料中的词项总数(WSSum).

表 1 和表 2 列出了有关的实验结果.从中可以看出,除了并列结构(CS)边界外,大多数的成分边界预测都达到了很高的准确率,从而为进一步进行 MNP 的自动识别提供了较为可靠的基础.

Table 1　Preprocessing results of WS prediction

表 1　预处理阶段的词语块边界预测结果

WSSum	Cort_BP	WSP(%)
64 426	62 402	96.86

Table 2　Preprocessing results of CG identication

表 2　预处理阶段的成分组边界识别结果

CG Types①	SCG	Cort_CG	Crossed_CG	CGP(%)	CGR(%)
PS	10 845	10 591	228	97.7	2.1
COS	1 259	1 201	36	95.4	2.9
CS	763	606	142	79.4	18.6
CC	1 096	1 018	70	92.9	6.4

①成分组类别.

2.2　最长名词短语识别实验

对于 MNP 自动识别器的处理性能,我们主要考察了以下两个指标:

(1) MNP 正确率(precision of the maximal noun phrase,简称 MNPP)=正确识别的 MNP 总数(Cort_MNP)/自动识别出的 MNP 总数(EMNP);

(2) MNP 召回率(recall of the maximal noun phrase,简称 MNPR)=正确识别的 MNP 总数(Cort MNP)/树库中的 MNP 总数(CMNP).

表 3 列出了目前的实验结果,其中简单 MNP 的词长<5,复杂 MNP 的词长≥5.从中可以看出,对于只利用边界分布概率知识的算法 1,MNP 识别的正确率和召回率分别为 69.2%和 70.9%,处理效果并不是很理想.当我们利用更为丰富的语言学知识,如 np 的内部结构组合规则等来进行 MNP 自动识别时(算法 2),正确率和召回率都有较大的提高,分别达到了 85.4%和 82.3%,显示出较为令人满意的自动识别效果.

另外,实验结果还显示出,目前的两个算法对复杂 MNP 的自动识别效果比较差,较长的复杂 MNP 的识别正确率一般要比简单 MNP 低约 16 个百分点,召回率则更低.这表明目前的自动识别器在对复杂 MNP 的识别机制上还存在着较大的缺陷,有待于今后进一步加以改进.

Table 3 Maximal noun phrase identifying results

表 3 MNP 自动识别结果

		CMNP	EMNP	Cort_MNP	MNPP(%)	MNPR(%)
Algorithm① 1	Simple MNP②	4 359	4 724	3 352	71.0	76.9
	Complex MNP③	865	626	350	55.9	40.5
	Total④	5 224	5 350	3 702	69.2	70.9
Algorithm 2	Simple MNP	4 359	4 279	3 763	87.9	86.3
	Complex MNP	865	754	534	70.8	61.7
	Total	5 224	5 033	4 297	85.4	82.3

①算法,②简单 MNP,③复杂 MNP,④合计.

2.3 错误实例分析

对算法 2 的识别错误实例进行分析发现,其错误原因主要可归纳为以下两个:一是由于预处理结果的边界预测错误而引起的;二是由于识别算法处理能力限制而引起的.显然,我们更关心的是其中的第 2 类错误,因为从中可以总结出一些用于改进识别算法的建设性意见.为此,我们设计了这样一个特殊实验,通过对 MNP 自动识别器的输入信息进行以下处理:(1) 删除自动发现的与树库中的成分相交叉的成分组描述;(2) 用树库中正确的成分边界信息替换自动预测出错的词语块描述,可以为 MNP 自动识别器提供成分边界完全正确的输入句子.表 4 列出了在此条件下算法 2 的自动识别结果,与表 3 的结果相比较可以发现,原来的识别错误中差不多有一大半是由于预处理结果的边界分析错误所引起的.

Table 4 Maximal noun phrase identifying results (A2) by using correct preprocessor data

表 4 使用正确预处理数据的 MNP 自动识别结果(算法 2)

	CMNP	EMNP	Cort_MNP	MNPP(%)	MNPR(%)
Simple MNP①	4359	4301	4073	94.7	93.4
Complex MNP②	865	734	626	84.9	72.4
Total③	5 224	5 038	4 699	93.3	90.0

①简单 MNP,②复杂 MNP,③合计.

对于剩余的 525 个错误实例(包括 339 个识别错误和 186 个未召回错误),我们从以下几个不同的角度对它们进行了深入细致的分析.

首先,通过考虑自动识别出的 MNP 的左右边界中是否有一个是正确的,可以将这些错误实例分为以下几个类型:

(1) 左边界识别正确的错误实例,进一步可细分为:

(a) 错误实例的右边界>正确实例的右边界(类型 I),

(b) 错误实例的右边界<正确实例的右边界(类型 II).

(2) 右边界识别正确的错误实例,进一步可细分为:

(a) 错误实例的左边界<正确实例的右边界(类型 III),

(b) 错误实例的右边界>正确实例的右边界(类型 IV).

(3) 自动识别出的 MNP 的左右边界都不正确(类型 V)

表 5 列出了这些不同类型的 MNP 错误实例的分布情况.从中可以看出,右边界识别正确的错误实例数目(199)远远超过左边界(56),这从一个侧面显示出汉语 MNP 识别的难点在于确定其左边界.如何寻找更好的方法,准确地确定复杂定语的左边界位置,将是我们今后研究的一个重点.

Table 5 Distribution of maximal noun phrase identifying errors (total 339) with different error types

表 5 具有不同错误类型的 MNP 识别错误分布(总数 339)

Error types①	I	II	III	IV	V
Distribution frequency②	25	31	45	154	84
Distribution ratio③(%)	7.4	9.1	13.3	45.4	24.8

①错误类型,②分布频度,③分布率.

对错误实例的内部结构组合进行分析,可以发现其中汉语的一些常见轻歧义结构占了很大比例.考虑结构组合:"v np 的 n",对于其中不同的词语,可能有以下两种合理的分析结构:(1) [$_{np}$[$_{vp}$ v np]的 n],如:[$_{np}$[$_{vp}$参加学术讨论会]的 老师];(2) [$_{vp}$ v [$_{np}$ np 的 n]],如:[$_{vp}$看 [$_{np}$老古董 的 电影]].而目前我们的识别算法只能统一地识别为结构(2),这导致了对类似结构识别正确率和召回率的降低.如何加强对歧义结构的识别能力,将是我们今后研究的另一个重点.

3 结束语

作为一项重要的应用基础研究,MNP 的自动识别对于自然语言处理领域中的许多应用研究,包括句法分析、信息检索、信息抽取、机器翻译等,都具有重要的实践意义.本文在汉语 MNP 的自动识别方面进行了一些有益的探索,通过对语料文本中的最长名词短语的分布特点的统计分析,提出了两种有效的汉语最长名词短语自动识别算法:基于边界分布概率的识别算法(算法 1)和基于内部结构组合的识别算法(算法 2),取得了较好的自动识别效果.在今后的研究中,我们将在以下几个方面对这些算法进行改进和提高:

(1) 寻找确定复杂定语的左边界的更好方法;

(2) 加强对歧义结构的识别处理能力.

参考文献

1 Church K. A stochastic parts program and noun phrase parser for unrestricted text. In: Proceedings of the 2nd Conference on Applied Natural Language Processing. Austin: Association for Computational Linguistics, 1988. 136~143

2 Bourigault D. Surface grammatical analysis for the extraction of terminological noun phrases. In: Boitet C ed. Proceedings of the 15th International Conference on Computational Linguistics (COLING'92). Nantes: Academic Press, 1992. 977~981

3 Voutilamen A. NPTool, a detector of English noun phrases. In: Church K ed. Proceedings of the Workshop on Very Large Corpora: Academic and Industrial Perspectives. Columbus: Association for Computational Linguistics, 1993. 48~57

4 Chen Kuang-hua, Chen Hsin-hsi. Extracting noun phrases from large scale texts: a hybrid approach and its automatic evaluation. In: Proceedings of the 32nd Annual Meeting of Association of Computational Linguistics. New York: Association for Computational Linguistics, 1994. 234~241

5 Lu Jian-ming. Special nesting phenomena of Chinese constituents. In: The Optional Papers of Lu Jian-ming. Zhengzhou: He'nan Education Press, 1993. 174~192
(陆俭明.汉语句法成分特有的套叠现象.见:陆俭明自选集.郑州:河南教育出版社,1993.174~192)

6 Li Wen-jie, Zhou Ming, Pan Hai-hua *et al*. Corpus-based maximal-length Chinese noun phrase extraction. In: Chen Li-wei, Yuan Qi eds. Advances and Applications on Computational Linguistics. Beijing: Tsinghua University Press, 1995. 119~124
(李文捷,周明,潘海华等.基于语料库的中文最长名词短语的自动提取.见:陈力为,袁琦主编.计算语言学进展与应用.北京:清华大学出版社,1995.119~124)

7 Tse A S Y, Wong Kam-fai *et al*. Effectiveness analysis of linguistics- and corpus-based noun phrase partial parsers. In: Choi Key-sun ed. Proceedings of the Natural Language Processing Pacific Rim Symposium (NLPRS'95). Taljon: Academic Press, 1995. 252~257

8 Zhou Qiang, Huang Chang-ning. Chunk parsing scheme for Chinese sentences. Technical Report, TR98002, Department of Computer Science and Technology, Tsinghua University, 1998
(周强,黄昌宁.汉语组块分析标注体系.科技报告,TR98002,清华大学计算机科学与技术系,1998)
9 Zhou Qiang. Phrase bracketing and annotating on Chinese language corpus [Ph. D. Thesis]. Beijing: Beijing University, 1996
(周强.汉语语料库的短语自动划分和标注研究[博士学位论文].北京:北京大学,1996)
10 Zhou Qiang. A model for automatic prediction of Chinese phrase boundary location. Journal of Software, 1996,7(supplement):315~322
(周强.一个汉语短语自动界定模型.软件学报,1996,7(增刊):315~322)
11 Zhou Qiang. Implementation of Chinese parsing algorithm based on bracket matching principle. In: Chen Li-wei, Yuan Qi eds. Language Engineer. Beijing: Tsinghua University Press, 1997. 194~200
(周强.汉语匹配分析算法的实现.见:陈力为,袁琦主编.语言工程.北京:清华大学出版社,1997.194~200)

Automatic Identification of Chinese Maximal Noun Phrases

ZHOU Qiang SUN Mao song HUANG Chang-ning

(*State Key Laboratory of Intelligent Technology and Systems Beijing* 100084)
(*Department of Computer Science and Technology Tsinghua University Beijing* 100084)

Abstract Based on the statistical characteristics of Chinese maximal noun phrases (MNPs) in a Chinese corpus with 5 573 sentences, two efficient identifying algorithms for Chinese MNPs: (1) To identify MNPs by using boundary distribution probabilities; (2) To identify MNPs by using internal structure rules, are proposed in this paper. Experimental results show better performances: precision 85.4% and recall 82.3%, by using identifying algorithm (2).

Key words Maximal noun phrase, boundary identification, syntax parsing.