DOI:10.19352/j.cnki.issn1671-4679.2002.03.017

文章编号 :1671-4679 (2002)03-0048-05

面向机器翻译系统 句法分析器的研究与实现

孟 遥[1], 赵铁军[2], 李 生[2]

(1.黑龙江工程学院 计算机科学与技术系,黑龙江 哈尔滨 150008;
2.哈尔滨工业大学计算机科学与技术学院,黑龙江 哈尔滨 150001)

摘要 :自然语言句法分析是机器翻译不可缺少的前期处理过程,文中总结了几代机器翻译系统中句法分析的经验,在最新研制的英汉双向机器翻译系统 MTS2000 中设计并实现了一个模块化的统计与规则相结合的句法分析模型。整个句法分析采用综合的策略,分别使用了隐马尔可夫方法、统计决策树方法、基于历史的句法分析等多种方法,并注意语义知识在句法分析中的应用。实验结果表明,模块化的句法分析器的设计方法,不论是对英语句法分析还是对汉语句法分析都是一种可取的方法。

关键词 :句法分析,统计与规则,机器翻译
中图分类号 :TP391.2 文献标识码 :A

Research on Parsing Technology in Machine Translation System

MENG Yao[1], ZHAO Tie-jun[2], LI Sheng[2]

(1. Heilongjiang Institute of Technology, Harbin 150008, China;
2. School of Computer Science & Technology, Harbin Institute of Technology, Harbin, 150001, China)

Abstract: Parsing is a central problem for many tasks in natural language processing, especially for machine translation. This paper presents the experiences in parsing for several machine translation systems, and proposes a modular parser using combination of statistic and rule methods. It integrates several strategies including Hide Markov method, statistic decision tree, history-based methods and employs semantic information in the algorithm. Experimental results indicate that this approach is successful in both English and Chinese parsing.

Key words: parsing; statistic rule; machine translation,

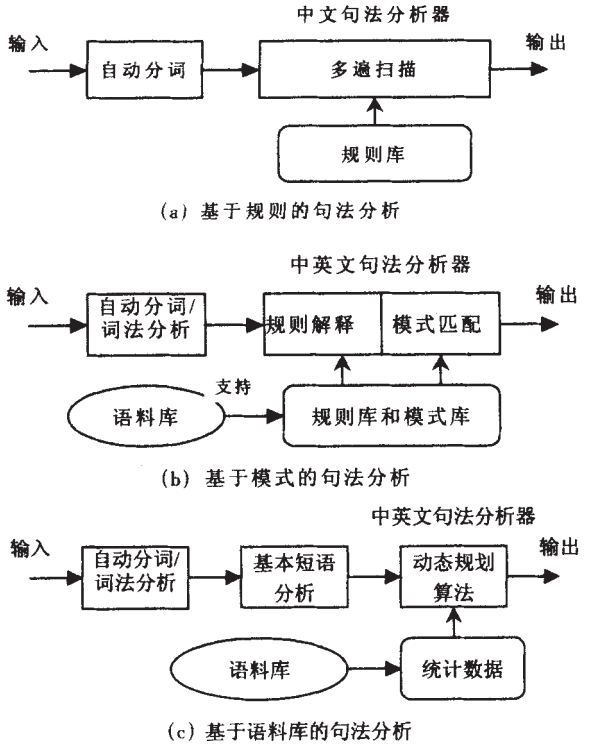
1 双向机器翻译与句法分析

机器翻译系统的研究与开发离不开对源语言的分析和理解，句法分析是源语言分析的关键步骤。哈尔滨工业大学机器翻译研究室多年来在汉语和英语的句法分析方面进行了一系列探索,先后实现了采用复杂特征集操作的确定性汉语句法分析器[赵铁军等,1992]、基于模式匹配的中英文句法分析器[赵铁军等,1995]、采用概率 GLR 算法的英语句法分析器[赵铁军等,1998]以及采用渐近式处理策略的中英文句法分析器。在研发过程中,我们的总体方法也实现了从基于规则方法到基于语料库方法的过渡。这一过程可以用图 1 表示。

下面简要回顾一下我们实现的各种句法分析器的特点。

采用复杂特征集操作的确定性汉语句法分析

收稿日期 :2002-04-17
基金项目 :黑龙江省教育厅科研基金资助项目 (0511179)。
作者简介 :孟遥(1970~),女,黑龙江工程学院计算机科学与技术系讲师,研究方向 :计算机应用。



(a) 基于规则的句法分析

(b) 基于模式的句法分析

(c) 基于语料库的句法分析

图 1　句法分析从规则方法到语料库方法的过渡

器应用于我们研制的 CEMT-III 汉英机器翻译系统中。该分析器是基于规则的,分析过程是多遍扫描的确定性分析。分析算法根据规则自底向上地完成结点的合并，分析成功体现为森林能够合并为一棵句法树。规则由手工获取,规则按照词类和短语符号来进行索引,以提高匹配速度。伴随着结点的合并,运用了大量的复杂特征集运算,从而为词性消兼和节点的正确合并提供支持。复杂特征集运算和多遍扫描策略成为以后不断继承的分析策略。

基于模式匹配的中英文句法分析器应用于我们研制的 BT863-I 汉英双向机器翻译系统中。该分析器采取了面向实例的知识获取，以期获得细粒度的翻译知识。知识表示采用了规则和模式结合的方式,其中各个组块内部使用规则表示,组块之间使用模式表示。相应地,句法分析的实现也分为规则解释和模式匹配两个部分。模式匹配算法采用最大匹配的二分查找算法，保证了翻译速度和匹配的正确性。基于模式的策略的基础是基于语料库方法。以语料库为基础,我们把实例扩展到模式，具有更强的普遍性，可以覆盖一类语言现象。可以把这些模式看作是语言学中句型的机器表示。通过该系统的实践,我们开始认识到语料库对句法分析和机器翻译的支持，同时也体会到大规模的知识获取的困难。

采用概率 GLR 算法的英语句法分析器应用于 BT863-II 英汉机器翻译系统中,其初衷是解决源语言分析中存在的各种歧义。因为当时我们尚没有采用词性标注技术,采取了部分规则来消除词性兼类,考虑到 GLR 算法具有一定的词性消兼的能力[Tomita, 1986],因而使用了概率 GLR 算法。其中概率数据来自我们利用该分析器 初始不带概率而由人工选择正确的句法树）分析得到的约 2000 个英语句子的句法树。在规则消兼之后,分析器还进行了基于规则的部分 NP 识别。尽管该分析器分为几个阶段，但是由于整个句法分析基本在 GLR 算法阶段实现，使得句法树的全解输出极其庞大。并且,因为 GLR 使用的 CFG 规则覆盖面有限,就很难面向真实句子进行分析。实践使我们体会到：对于不同问题进行个别处理,分模块、分步骤地解决句法分析中的各种歧义，应该是一种可取的策略。因此,我们在新的 MTS2000 系统中采用了渐近式处理策略。

2　双向机器翻译系统句法分析技术

在我们实现的 MTS2000 双向英汉机器翻译系统中,综合了句法分析中的多种方法,采用规则与统计相结合的总体思想,具体实现时,使用渐近式的问题解决策略,将句法分析分为若干个子问题分而治之,取得了较好的效果。

我们认为在开发一个面向实用的句法分析器时应整体考虑基于规则和基于统计两种方法各自的优缺点。采用综合的解决策略,尽可能发挥每种方法各自的优点,最大可能地减少句法分析过程中出现的错误[C.Manning 1999]。MTS2000 句法分析器的开发就是从这种思想出发,采用统计与规则结合的方法,解决句法分析中所遇到的问题。

对英语的句法分析的研究一直是自然语言处理领域的研究热点，一些浅层句法分析问题,比如词性标注、BNP 识别、介词短语附着等也已经取得了很好的效果。对深层次的完全句法分析的研究也十分活跃，比较有代表性的有 Collins 的头驱动的统计句法分析模型[Collins, 1999]、Goodman[Goodman, 1997]的包含复杂特征集的统计模型、Magerman[Magerman, 1995]的基于统计决策树的模型和 Charniak[Charniak, 1997]的基于词汇依存关系的句法分析器。其中 Collins 在 1999 年实现的句法分析器为目前所看到的最好的句法分析器。

尽管上面提到的句法分析方法都取得了较好的效果，但这些模型采用的都是一体化的解决策略，它对句法分析的各个环节中出现的子问题并没有分别考虑，对不同的问题也难以采用不同的方法。Collins 对 Penn treebank 00 section (此为一个比较公认的测试语料，许多句法分析系统都有根据该语料所做的测试结果) 所做的句法分析结果显示，其词性标注和 BNP 识别的精确率都没有达到目前最好结果 (Collins 的词性标注正确率为 96.6%, BNP 识别精确率为 92.4%, 目前所见到最好结果为词性标注 97.4%, BNP 识别精确率为 93.4%)。因此句法分析精度的进一步提高可以从渐近式的句法分析方法考虑，充分利用浅层句法分析所取得的结果，用模块化的分治的方法进行句法分析。

从开发一个实用性的句法分析器的角度出发，模块化的句法分析方法可以体现了处理自然语言的层次性观点。语言是分层次的符号系统，层次化地解决问题，有利于在解决高层次问题时，发现问题和方便地寻求解决问题的方法，具有可控性好的优点，对于面向实用化的系统来说，这一点具有非常重要的意义。另外也体现了处理自然语言的系统工程观点。用系统论的观点来分析处理对象，用系统工程的观点来求其实用。

因而我们在开发 MTS2000 句法分析器的过程中，将英语的句法分析分为：词性标注、BNP 识别、介词短语附着确定、从句边界确定和最终的句子骨干分析，对于汉语的句法分析我们将其分为：分词和词性标注、关联词组的识别、汉语基本短语的识别和最终汉语句法分析。

3 MTS2000 英语句法分析器

3.1 MTS2000 英语句法分析器总体结构

整个英语句法分析过程分为句子的浅层句法分析和句子骨干部分深层句法分析两大部分。输入一个英语句子，经过词性标注后，首先进行句子的浅层句法分析，完成英语基本名词短语(BNP)识别和介词短语附着(PPA)确定以及从句边界的识别，随后浅层句法分析的结果作为句子骨干部分句法分析的输入，最终完成整个句子的句法分析。在这样的渐近式句法分析过程中，后面的处理模块需要用到前面的处理结果，为避免错误累加，我们在每一个模块处理结束后都增加了一个基于规则的错误校正机制，尽可能保证后面模块所接收到的结果是正确的。

3.2 英语句法分析各模块介绍

在整个英语句法分析过程中我们使用了多种统计学习方法，分别实现了基于隐马尔可夫模型的词性标注，决策树的 BNP 识别，决策树与语义结合的介词短语附着确定，基于语料库的从句识别和基于历史的句子骨干部分分析。各模型的训练数据主要来源于 Penn treebank。

哈尔滨工业大学机器翻译研究室从 1994 年开始，对于英语词性标注问题，先后对基于规则和基于 HMM 的方法进行了研究，词性标注正确率分别达到 86%和 95%。在总结前面经验的基础上，我们提出了将规则和 HMM 模型相结合的方法，将规则引入放在 HMM 计算之前，减少了计算时间复杂性。并将形态还原的部分分析结果与词性标注结合关联起来，许多形态分析的特征直接应用指导词性标注。取得比较好的标注结果，正确率达 97%。

对于英语基本名词短语 (BNP)的识别，我们也探索了多种方法，尝试了基于 HMM 的 BNP 识别、基于规则的 BNP 识别和基于决策树的 BNP 识别方法。目前的实验结果表明：使用决策树方法可以很好地考虑远距离的上下文限制，并且对一些难以识别的 BaseNP 可以方便灵活地加入语义和词汇信息以帮助识别。这种方法知识获取容易，算法速度很快 (与句子长度是线性关系)而且算法具有学习能力。是一种较好的 BNP 识别方法。

具体实现时我们把 BaseNP 识别工作分成两步，第一步使用决策树的分类方法，将一个英语句子中的每一个词分为是属于 Base NP 的和不属于 Base NP 的两类。第二步采用某些组合策略将属于 Base NP 的词进行一定的结合，构成 Base NP 短语。决策树实现时使用 ID3 算法。

介词短语附着歧义是英语句法分析中较难处理的问题，通常的解决策略是把问题进行简化处理，转化为考察与消歧问题密切的四个核心词之间关系，通过建立统计模型，从概率角度解决此问题。从存在短语附着歧义的英语句子中，抽取四个核心词，形成四元组 (v, n_1, p, n_2)，这里：v 表示动词短语的核心词、n_1 表示名词短语的核心词、p 表示介词短语的介词、n_2 表示介词短语的核心名词。消岐问题转化为 $(v, n_2) \rightarrow v$、$(v, n_2) \rightarrow n_1$ 的决策。这种方法存在着两个问题：一个是单纯使用词汇信息而引起的数据稀疏问题，另外就是难以解决多重附着的问题。针对这两个问题，我们提出了运用 ID3 决策树和 WordNet 语义信息相结合的方法来确定英语介

词短语附着对象。将介词的附着问题抽象成词汇和语义元组的组合,然后转变为对每个词汇和语义元组属性值的决策。并通过递归运用前面的决策结果解决多重介词短语分析[赵铁军,2001],介词附着的正确率为 85%。

为了降低整体句法分析的复杂度, 在词性标注、BNP 识别和介词短语附着判别之后进行了统计与规则相结合的从句识别[ZhangJing, 2001]。

首先我们将从句的识别问题转化为识别从句的左边界和右边界两个问题。以句子中的限定动词为中心,左右扩展,使用最大规则匹配方法先找出从句的左、右边界可能的范围,然后通过对训练语料进行统计而获取的从句边界词与其前一个词和后一个词的三元同现概率数据,在最大长度规则匹配的范围内寻找从句左边界和右边界的最可能位置。这种综合的解决策略使从句边界识别的精确率达到 90%以上。

输入的英语句子经过词性标注、BaseNP 识别、介词短语附着、从句分解等步骤后,得到了句子骨干结构。对句子骨干结构的进一步分析是本句法分析系统中最核心的部分。

这里我们使用了一种类似于基于历史的句法分析的方法。句法分析的整个过程是分层进行的,在每一层的输入是其下面一层的输出。每一层得到所有能够匹配句子子串的规则。随后对这些规则进行挑选, 挑选时使用一个组合型的概率评价函数,选取概率函数评分最大的规则序列,执行规则序列中的每一条规则,合并生成新结点。以此类推,自底向上完成整个句法分析。

在规则选择时,我们使用一个组合型的概率评分函数将规则本身的出现概率与规则所处的上下文环境一起考虑,以此选取概率最优规则链。

我们使用 Penn treebank 00 section 测试集对句法分析的各模块进行了测试,并与其它句法分析方法进行了对比,结果如下:

表 1　测试结果

MTS2000 句法分析模块	精确率	召回率	目前最好结果	
			精确率	召回率
词性标注	96.04%		97.4%	
BNP 识别	92.2%	92.8%	92.3%	93.2%
介词附着确定	84.1%		84.6%	
从句识别	89.5%	87.5%		
句法分析	84.1%	84.4%	88.1%	88.6%

实验显示,MTS2000 句法分析器的 BNP 识别和介词短语附着都取得了较好的效果,最终句法分析结果虽然没有达到最好水平,但因为我们采用渐近式的模块化的句法分析模型,单个模块精度的提高都会给最终的结果带来明显的提高,随着单个模块的逐渐完善,系统的性能还会继续提高。另外针对模块化的方法带来的错误累积的问题,我们采用了基于规则的后校正机制,有系统地增加校正规则也可以进一步完善系统性能。

4　MTS2000 汉语句法分析器

4.1　汉语句法分析器的实现过程

汉语缺乏严格意义上的形态标志和形态变化。同英语相比, 汉语的句法分析面临更多的困难,主要表现在:1.汉语缺少形态标志,汉语中的同一个句法成份可以由属于不同词类的词来充任,同一个词在句法结构中又可以作不同的句子成份,形式上却没有任何不同的标志。因此,汉语的兼类问题更为突出、更难解决。2.知识获取问题是汉语句法分析实现大规模开放应用的瓶颈之一。同英语相比,汉语缺少经过句法标注的大规模汉语树库,这在一定程度上阻碍了汉语句法分析的发展。

MTS2000 的汉语句法分析器的实现, 经历了比英语句法分析更复杂的过程。整个句法分析器的实现分为两个阶段,首先是树库资源的建设,在标注树库的初始,我们在对比分析了目前 3 种典型的汉语句法标注体系的基础上, 提出了一种"自底向上"的汉语句法标注体系的设计方法, 完成了一套比较完整的汉语句法标注集[杨沐昀,2001]。随后在标注体系的指导下标注了 3 万句的汉语词性标注语料和 7000 句的汉语树库标注语料。句法分析的第二个阶段是句法分析模型的设计和实现过程。英语句法分析器的开发实践使我们感到模块化的统计与规则相结合的方法,对于一个实用的系统来说非常适用。因此在开发汉语句法分析器时,我们也采用了模块化的方法,整个句法分析分为词性标注、关联词组的识别、汉语基本短语识别和最终句法分析。

4.2　汉语句法分析各模块介绍

汉语句法分析我们使用分词和词性标注一体化的方法,分词获得的多个结果,由基于 HMM 的二元词性标注模型进行评价,将概率最大的词性标注链作为最终结果。针对汉语的特点,在分词时我们使用了多步处理的策略, 整个处理过程包括:消除歧义、句子的全切分、部分确定性切分、数词串处理、重叠词处理和基于统计的未登录识别。部分开放测试结果表明分词精确率可达 98%, 对 52 类词

性集的词性标注精确率为 93%以上[赵铁军,2001a]。

为了解决汉语中存在的远程搭配问题,我们建立了汉语关联词词表,分析中加入了关联词组的识别,并且在识别的过程中考虑了搭配中存在的嵌套问题。

受英语 BNP 识别的启发,国内出现了对汉语的基本短语的研究 [赵军,1999][周强,1996][周强,1999]。我们定义的汉语基本短语包含:名词短语(NP)、动词短语(VP)、形容词短语(AP)、副词短语(DP)、数量短语(MP)、时间短语(NT)和处所短语(NS)。汉语基本短语识别主要使用统计方法,其训练数据来源于我们标注的 7000 句树库。在识别时采用的基于评价的规则选择方法,对于从语料库中推导出的基本短语的规则,在规则选取过程中根据规则本身的概率及它的上下文关系,采用概率评价函数确定规则,用确定的规则合并基本短语。 组成 BNP 的规则为包含复杂特征的规则,分析结束时给出多个候选结果,通过不确定机制减少基本短语识别错误对整体句法分析产生的影响。针对 7000 句的封闭测试精确率和召回率分别为 92.1%, 91.5% ,对于开放语料汉语基本短语的识别精确率和召回率也达到 80%以上。

由于受语料库规模的限制,汉语的整体句法分析采用的是基于规则的方法,规则为包含复杂特征集的上下文相关规则。我们通过理论精化的方法完善句法知识库,保证新旧知识的一致性和系统性能的增量式提高[赵铁军,2001b]。

5 结束语

自然语言句法分析是一个集人工智能和语言学知识为一体的系统,它涉及到了人工智能、语言学、计算机等技术,在句法分析的过程应注重多策略的结合。在 MTS2000 英语和汉语的句法分析研究过程中, 我们尝试使用渐进式的问题解决方法,结合了基于规则的和基于统计的句法分析手段,解决句法分析所面临的问题。分解使得句法分析的每个任务更加具体,使句法分析更有层次性,便于采取更有针对性的解决方案。而统计与规则的结合可以更好地利用各自优点,为句法分析服务。初步的实验结果显示,这种句法分析器的设计方法,不论是对英语句法分析还是对汉语句法分析都是可行的。

参考文献

[1] E. Charniak. Statistical parsing with a context-free grammar and word statistics[C]. Proceedings of the Fourteenth National Conference on Artificial Intelligence. AAAI Press/MIT Press, Menlo Park, (1997).

[2] M. Collins. Head-Driven Statistical Models for Natural Language Parsing[D]. Ph. D. Thesis, The University of Pennsylvania, 1999.

[3] J. Goodman. Parsing Inside-Out[D]. Ph.D. Thesis, Harvard University, 1998.

[4] C. Manning, H. Schutze. Foundations of Statistical Natural Language Processing[M]. The MIT Press, 1999.

[5] Masaru Tomita. Efficient Parsing for Natural Language - A Fast Algorithm for Practical Systems[J]. Kluwer Academic Publishers, 1986.

[6] ZhangJing, et al. A Corpus-based Approach to English Subordinate Clause Identification[J]. High Technology Letters,Vol.7,No.1. 10-12.

[7] 张国煊,郁梅,王小华. 基于语料库的汉语短语边界划分的研究[A]. 第三届全国计算语言学学术会议论文集 《计算语言学进展与应用》[C]. 上海,1995. 94-99.

[8] 赵军,黄昌宁. 汉语基本名词短语结构分析模型[J]. 计算机学报,1999. 22 (2):136-141.

[9] 周强. 汉语语料库的短语自动划分和标注研究[D]. 北京大学博士论文,1996.

[10] 周强,黄昌宁. 基于局部优化的汉语句法分析方法[J]. 软件学报,1999. 10 (): 4-6.

[11] 赵铁军,李生,周明. 一种生成复杂特征集句法树的汉语句法分析方法与系统实现[J]. 中文信息学报,Vol. 6, No. 4. p11-23.

[12] 赵铁军. 面向实力、基于模式的机器翻译方法与实现[A].吴泉源,高文. ′95 智能计算机接口与应用进展[C]. 北京 :清华大学出版社,1995, 507-512.

[13] 赵铁军. 英汉机器翻译系统 BT863-II 的消歧策略研究[A]. 郑守淇,钱德沛,曾明. ′98 人工智能进展[C]. 西安 :西安交通大学出版社,1998, 207-212.

[14] 赵铁军,方高林. 英语介词短语附着决策的研究[J]. 高技术通讯,Vol.11,No.3. 36-40.

[15] 赵铁军,吕雅娟. 提高汉语自动分词精度的多步处理策略[J]. 中文信息学报,Vol.15,No.1. 13-18.

[16] 赵铁军、刘芳. 基于学习的句法知识库进化,智能计算机研究进展[A]. 怀进鹏. 863 计划智能计算机主题学术会议论文集[C]. 北京 :清华大学出版社,2000. 349-354.

[17] 孟遥,黄玉. 一个包含复杂特征的统计英语句法分析模型[A]. 黄昌宁,张普. 自然语言理解与机器翻译[C]. 北京 :清华大学出版社,2001. 167-172.

[18] 杨沐昀,赵铁军. 自底向上的汉语句法标注体系设计与实践[A]. 黄昌宁,张普. 自然语言理解与机器翻译[C]. 北京 :清华大学出版社,2001. 160-166.

[责任编校 :王黎]