一个人机互助的汉语语料库多级加工处理系统CCMP*

周强　俞士汶

北京大学计算语言学研究所
北京，100871

摘要：本文简要介绍了一个人机互助的汉语语料库多级加工处理系统的概况，包括它的设计思想和总体结构框架；并给出了短语自动划分和标注处理子系统的算法和实验结果。最后提出了进一步增强、改进系统的一些设想。

A man-machine mutually-dependent multistage processing system of Chinese language corpus

Zhou Qiang, Yu Shiwen

Institute of Computational Linguistics, Peking University
Beijing, 100871

ABSTRACT: In this paper, we introduce the survey of a man-machine mutually-dependent multilevel processing system of Chinese language corpus, including its designing idea and overall framework. Then, we describe the main algorithms and experimental results of an important sub-systems: the phrase bracketing and tagging sub-system. At last, we also propose some tentative ideas for improving the system in future.

1. 引　言

对原始语料进行多级加工处理，是语料库语言学研究的基础。为此，国外在这方面花费了大量的财力和物力，比较大的研究项目包括：英国 Lancaster 大学 UCREL 的 Lancaster Treebank 项目，从1986年到1994年的九年间，陆续开发了CLAWS1, CLAWS2, CLAWS3, CLAWS4等数个功能不同的词类自动标注工具([MI83],[GLS87], [LGB94])，并在语料库句法分析和标注方面积累了许多有益的经验([LG91])。美国 Pennsayvania 大学的 Penn Treebank 项目([MSM93])，通过吸收和改造一些现有的语料处理工具，如：Church 的词类标注工具([CW88])和 Hindle 的 Fidditch 句法分析器([HD89])，形成了一个完整的语料库加工处理系统。另外，它的一大特点是开发了功能强大、操作简单的语料校对工具，大大提高了人工校对的效率。

近几年来，对汉语语料库加工处理的研究也逐渐开展了起来，并在自动切词([LNY87],[XHS91])、词性标注([BXH92],[BXH92])和依存关系标注([ZH94], [LZH93])方面取得了可喜的成果。但由于各方面条件的限制，还没能形成一个完整的汉语语料库多级加工处理系统。

* 本课题受国家自然科学基金资助

从1992年初开始，北大计算语言学研究所开始进行汉语语料库的多级加工处理研究，经过几年的努力，提出了一些新的处理方法，开发和积累了许多有用的处理工具，逐渐形成了一个较为完整的汉语语料库多级加工处理系统。

本文将对这一系统的设计思想、总体结构和基本功能作一简要介绍。其中，第 2 节通过一个典型的语料加工模型阐述了系统的基本设计思想；第 3 节给出了系统总体框架；第4节则简要介绍了系统的一个主要子系统(即短语自动划分_标注)的基本算法，并给出了目前的一些实验结果。最后，在结语中，我们提出了一些改进设想。

2. 人机互助的语料加工处理模型

在语料库的加工处理过程中，随着人力物力的不断投入，经过校对的正确标注语料的数量也在不断增加。这是一笔巨大的财富，因为其中包含了丰富的语言学知识，并隐含了人进行排歧处理所用的各种知识。如何最大限度地发挥这个语言知识库的作用，是提高语料库处理系统整体性能的关键。基于这种认识，我们构造了图 1 所示的语料加工模型.

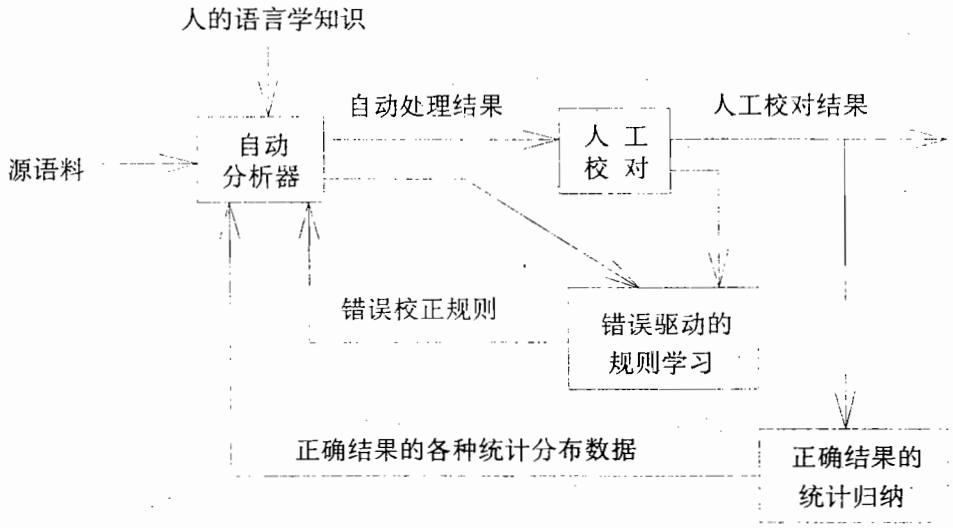

图 1　CCMP系统的语料加工模型

此模型具有以下几个特点:

①. **普遍性知识和特殊性知识相结合**　当正确标注的语料达到一定规模以后，从中统计得到的分布数据近似地反映了语言中的一些普遍规律，将这些数据运用于自动标注处理，就可以期望获得较高的处理正确率。但语言是千变万化的，具有许多特例。系统配备了一个由错误驱动的规则学习过程，通过将自动处理结果和人工校对结果相比较，发现错误所在，从中总结归纳出若干特殊情况的处理规则。如此，将统计得到的普遍性知识和学习得到的特殊性知识相结合，可以大大提高自动标注处理的性能。

②. **人机处理相结合**　机器处理的优势在于它有强大的计算能力，可以大规模地处理语料。而人工标注的优势则是精确度，因为人能利用上下文信息和知识来排歧。这两方面的优势在图 1 所示的模型中都得到了充分的发挥：一是利用统计数据，构造适当的统计模

型进行自动标注处理；二是通过人工校对，保证最终处理语料的正确性。而对于错误校正规则的学习，则要经历一个 手工->半自动->全自动 的发展过程。最初是人工总结，随着研究的深入，可以逐步利用一些统计工具降低人工处理的工作量，当技术上达到成熟时，就可以利用机器学习技术自动习得有用的规则。

③. **具有整体性能的增量提高性** 随着正确标注语料规模的不断扩大，将使统计数据反映的信息更加全面，错误校正规则的条件约束更为精确，从而提高了自动标注处理的正确率，降低了人工校对的工作量，使系统的整体性能得到增强。

3. 系统总体框架

3.1 总体结构图

综合不同阶段的语料加工模型，并加上各种辅助处理工具，就形成了如图 2 所示的汉语语料库多级处理系统总体结构图。

3.2 资源数据库说明

3.2.1 语料库

根据语料处理层次和内容的不同，将所有语料组织成如下的三级结构:

(1). 加工深度级: 按照加工深度的不同，将语料分成六大部分: 原始文本、切分结果、词类标注结果、切分和词类标注的人工校对结果、短语划分和标注结果、短语分析的人工校对结果。

(2). 内容分类级: 在不同的加工深度下，按语料内容和研究重点的不同进行分类组织。

(3). 文件组织级: 将同一内容的语料，组织成一定长度的语料文件 。

3.2.2 电子词典

保存了进行语料加工处理所需要的有关词语的各种语言学知识，目前主要使用了句法特征信息，它们来源于北大的“现代汉语语法电子词典”([YZG92])以及目前开始建造的“现代汉语短语属性信息库”。

为提高处理效率，目前把电子词典分成两大部分:

1).切分和词类标注词典: 包含词语和词类信息，收录了约有 4.5 万个词条;

2). 短语分析词典: 在词条中包含了丰富的句法特征信息，可根据语料的不同适当调整词典的规模。

以后，随着系统集成度的提高及电子词典信息的不断扩大，可考虑将两者合二为一。

3.2.3 规则库

在分析大规模的真实语言文本过程中，会遇到各种各样的歧义现象。而要消除这些歧义，就必须依靠大量的语言学知识。将这些知识形式化，我们就得到了大量不同类型的消歧规则。它们主要保存在规则库中。

3.2.4 统计信息库

包含了对语料库信息的各种统计结果。如，带词性标记的词频统计表，两个词性的共现频率矩阵，短语结构分布信息等。它们为基于统计的语料库处理技术提供了客观的语言分布数据。

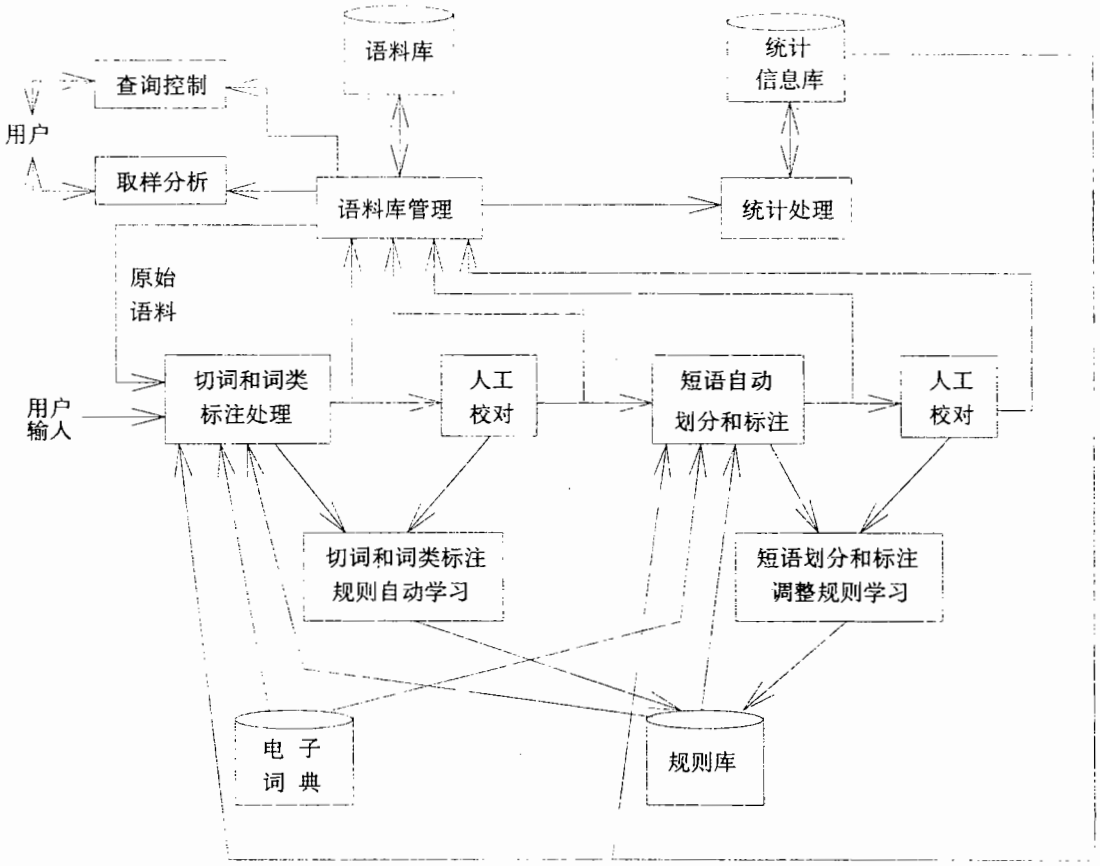

图 2 . 汉语语料库多级加工处理系统CCMP总体结构图

3.3 各基本处理模块功能简介

i). **语料库管理模块** 通过设置不同的文件目录和自动标识不同的结果文件后缀为各级语料库处理结果提供一个统一的信息存取机制。

ii). **查询控制模块** 为用户检索语料库信息提供各种方便的查询工具。其中一项重要的信息是文中关键词索引 KWIC(Key Word In Context)。

iii). **取样分析模块** 实现从不同处理层次的语料中进行随机取样的功能，为用户进行语言分析提供方便。

iv). **统计处理模块** 根据不同统计模型的数据要求，对不同处理层次的语料库信息进行多方面的统计处理。

v). **切词和词类标注处理模块** 实现了切词和词类标注相融合的处理([ZY93])。其中，我们采用了较小的词类标记集([BD94])，并提出了一种规则和统计相结合的汉语词类自动标注方法([ZQ94])，取得了较好的实验效果。

vi). **短语自动划分和标注模块** 实现短语的自动划分和标注。详见第4节。

vii). **人工校对模块** 监控人工校对过程，提高处理效率

viii). **规则学习模块** 通过对自动处理结果和人工校对结果的比较，总结和归纳错误校正规则。

4. 汉语短语的自动划分和标注

4.1 基本处理算法

对于一句已完成了正确切词和词性标注处理的句子，如何确定其中不同短语的边界位置，将它们用括号正确地划分出来，并标以合适的短语标记，是汉语短语自动划分和标注算法所要解决的主要问题。对此，我们提出了一套基于统计的自动处理算法。它分为预测划分点、括号匹配和分析树生成等三个处理阶段。其中，预测划分点阶段的主要任务，是依据句子中的词语、词类信息，预测短语结构的边界点，即在哪个位置应予以左分('['), 哪个位置应予以右分(']'), 而哪个位置不可能是短语与短语的分界点(' '); 而在括号匹配和分析树生成阶段，则要根据预测得到的划分点，进行左右括号的匹配，同时，逐步生成句子的分析树(或森林)，并进行统计排歧，最终得到一棵最佳的分析树，从而可以完成对句子的短语自动划分和标注。有关此算法的一些细节，可参看([ZQ95])。

如何充分利用树库(treebank)中包含的丰富的句法信息进行统计排歧处理，是这种算法成功的关键。

4.2 统计信息和排歧处理

根据算法不同阶段的处理要求，我们可以从树库中统计得到以下三组数据:

(1). 边界分布信息.

这组数据反映了不同的语境信息对边界划分点的影响能力的大小，包括:

①. 词语信息的作用: [W 和 W]

②. 词类信息的作用: [$t_i\ t_{i+1}$ 和 $t_{i-1}\ t_i$]

它们在划分点预测过程中发挥着重要作用。

(2). 边界标记信息

这组数据表明，在不同的语境下，左右划分点可能是哪些短语的边界。例如:

n [p → vp 0.531, pp 0.462, np 0.007

表示名词(n)和介词(p)之间的左分点作为动词性短语 vp 的左边界的概率为 0.531，作为介词短语 pp 的左边界的概率为 0.462，作为名词性短语 np 的左边界的概率为 0.007。这组数据是进行括号匹配的重要依据。

(3). 短语结构信息

这组数据反映了语料中短语成分组合成不同短语的可能性大小。如: vp → v+np , 0.132 , 表示在统计语料中 v+np 可以组合为 vp 的概率为 0.132。此组数据在分析树结构排歧中将起重要的作用。

([ZQ95])中详细介绍了如何利用这些统计数据，构造不同的统计模型进行自动排歧处理的方法。

4.3 实验结果

目前，利用此算法对 1400 多句汉语句子进行了自动短语划分和标注。为较好地了解系统的处理性能，对两项重要的技术指标: 括号交叉率和标记错误率进行了分析。其中，括号交叉率记录了自动分析得到的括号对与正确划分的括号对形成交叉(crossing)情况的比率; 而标记错误率则反映了短语标记被标错的情况所占的比率(详见[ZQ95])。有关的实验结果为:

括号交叉率为: 13.98%

标记错误率为: 8.65%

5. 结 语

本文简要介绍了一个人机互助的汉语语料库多级加工处理系统的设计思想、总体框架及基本处理工具的算法和实验结果。目前，此系统已初具规模，完成的处理工具包括:

(1). 一个比较完善的语料库组织和管理工具库。

(2). 一个较为成熟的切词和词性标注相融合处理子系统

(3). 一个初步可用的短语自动划分和标注处理工具

(4). 面向加工语料(切分、词性标注、短语划分和标注)的查询、统计和取样工具。

在我们的多级加工汉语语料库构建研究项目中，此系统正发挥着越来越重要的作用。

在今后的研究中，我们将在以下几个方面对系统进行改造和完善，使系统的整体性能得到进一步的提高:

(1). 改进短语自动划分和标注算法，使之能很好地处理更多的真实文本。

(2). 探索机器学习和统计处理的新方法，争取在未登录词的正确切分和短语歧义自动辨析等方面有所突破。

(3). 开发功能强大、操作简单的人工校对辅助工具，提高人工校对的处理效率。

(4). 加强各子系统和处理工具间的信息传递和反馈，提高系统的集成度。

(5). 扩大加工的范围(如增加词语的注音)和深度(如同形义项的标注)

参考文献

[BD94] 北京大学计算语言学研究所 (1994). “现代汉语文本切分与词性标注规范 (V1.0)”，内部资料

[BXH92] 白栓虎、夏莹、黄昌宁，(1992). “汉语语料库词性标注方法研究”，机器翻译研究进展，408—418

[CW88] Church, Kenneth W. (1988). " A stochastic parts program and noun phrase parser for unrestricted text." In *Proceedings, Second Conference on Applied Natural language Processing*. 136-143.

[GLS87] Garside,R.; Leech,G.; and Sampson,G. (1987). *The Computational Analysis of English*. Longman.

[HD89] Hindle, Donald (1989). "Acquiring disambiguation rules form text." In *Proc. of ACL-27*.

[LG91] Leech, G.; and Garside, R. (1991). "Running a grammar factory: The production of syntactically analysed corpora or 'treebanks' ". In Stig Johansson and Anna-Brita Stenstrom (eds.) *English Computer Corpora : Selected papers and Research Guide*. 1991. 15-32

[LGB94] Leech, G.; Garside, R. and Bryant M. (1994). "CLAWS4: The Tagging of the British National Corpus", In *Proc. of COLING-94*, 622-628

[LNY87] 梁南元，(1987). “书面汉语自动切词系统－ＣＤＷＳ”，中文信息学报，1(2)

[LZH93] 李京葵，周明，黄昌宁. (1993). “统计和规则相结合的汉语句法分析研究”，计算语言学研究和应用，北京语言学院出版社，176-182 .

[LZZ92] 刘开瑛、郑家恒、赵军，(1992). “语料库词类自动标注算法研究”，机器翻译研究进展，378—386

[MI83] Marshall, I. (1983). "Choice of grammatical word-class without global syntactic analysis: tagging words in the LOB Corpus", *Computers and Humanities*, 17, 139-150

[MSM87] Mitchell P.Marcus, Mary Ann Marcinkiewicz, and Beatrice Santorini (1993). "Building a Large Annotated Corpus of English: The Penn Treebank", *Computational Linguistics*, 19(2), 313-330

[XHS91] 徐辉、何克抗、孙波，(1991). “书面汉语自动切词专家系统的实现”，中文信息学报，5(3)

[YZG92] 俞士汶、朱学锋、郭锐，(1992). “现代汉语语法电子词典的概要与设计”，In *Proc. of ICCIP92*, 186-191

[ZH94] 周明，黄昌宁 (1994). “面向语料库标注的汉语依存体系的探讨”，中文信息学报，8(3)，35-52

[ZQ94] 周强 (1994). “规则和统计相结合的汉语词类标注方法”，中文信息学报(待发表)

[ZQ95] 周强 (1995). “汉语短语的自动划分和标注”(待发表)