Advances of Computational Linguistics in China

中国计算语言学研究前沿进展

(2009—2011)

孙茂松 陈群秀 主编



清华大学出版社

需要全本请访问 www.ertongbook.com

Advances of Computational Linguistics in China

中国计算语言学研究前沿进展

(2009-2011)

孙茂松　陈群秀　主编

清华大学出版社
北京

第十一届全国计算语言学学术会议组织情况

日　　期：　2011 年 8 月 20 日—22 日
地　　点：　洛阳中国人民解放军外国语学院

发起单位：　中国中文信息学会

赞助单位：　教育部语言文字信息管理司
富士施乐有限公司
百度公司
富士通研究开发中心有限公司
TRS 信息技术有限公司
腾讯科技（深圳）有限公司
中国中文信息学会

组织主办单位：　中国人民解放军外国语学院
清华大学智能技术与系统国家重点实验室

大会主席：　李宇明
大会副主席：　董振东
程　工

程序委员会主席：　孙茂松
程序委员会副主席：　张　普　易绵竹
程序委员会委员：（以姓氏拼音字母次序为序）
蔡东风　常宝宝　陈群秀　陈小荷　程学旗　黄河燕　黄萱菁
何婷婷　李　茹　刘　群　刘绍明　刘　挺　史晓东　宋　柔
吐尔根·依不拉音　姚天昉　俞士汶　张　全　赵　军
周国栋　周　明　朱靖波　亢世勇

组织委员会主席：　毕玉德
组织委员会副主席：　陈群秀　曹右琦　张克亮
组织委员会委员：　濮建忠　唐慧丰　张　兴　李治国　李　鹏　王慧兰　林　丽
张　飞　李　峰　李伟刚　崔晓菊　李　琳　阎丹辉　童毅见
孙士杰

前　言

由中国中文信息学会主办的“第十一届全国计算语言学学术会议”（The 11th Chinese National Conference on Computational Linguistics，CNCCL-2011，简称 CCL-2011），于 2011 年 8 月 20～22 日在河南省洛阳市中国人民解放军外国语学院举行。

“全国计算语言学学术会议”是国内历史最悠久、影响最大的计算语言学领域的系列学术会议。本届会议共征集到论文 167 篇，经大会程序委员会严肃认真地评审，最终录用 93 篇，录取率为 55.7%。其中词法、句法、语义和篇章分析方面的论文 35 篇，语言资源建设及相关技术方面 12 篇，机器翻译技术与系统方面 10 篇，智能检索方面 31 篇，其他 5 篇。程序委员会将从录用的论文中精选出一批论文，组成专辑，推荐给《中文信息学报》。经作者认真修改后，这些论文会以最快的速度于《中文信息学报》上发表。

本届会议论文集的内容非常丰富。与上届相比，无论是在研究与应用的深度还是在广度上都有了新的拓展，如词法、句法、语义和篇章分析方面的论文开始讨论异种语料融合、文本蕴含识别、隐喻及明喻计算等问题，语言资源建设及相关技术方面的论文涉及到了百科知识工程和大规模汉语事件知识库构建，智能检索方面的论文进一步围绕维基百科和博客等展开工作。总的感觉是，学者们近年来研究工作的“剑锋”所指，逐步更多地指向了较深层次的语义处理，这无疑是一个值得鼓励的正确方向。相信读者从这本论文集中一定会深受启发。

在此，我们诚挚地感谢大会邀请报告演讲者和全体作者对大会的热衷支持；感谢大会主席的有力指导，感谢程序委员会的辛勤劳动；感谢大会组织委员会的出色工作；感谢清华大学出版社在出版方面的积极配合；感谢赞助单位的慷慨解囊；感谢中国中文信息学会计算语言学专业委员会为会议牵头所做的贡献。这些共同的努力，确保了本届会议必将一如既往地取得圆满成功。

最后，祝与会代表们在洛阳工作顺利、生活愉快！建议大家在充分享受学术饕餮盛宴的同时，也能忙中偷闲，领略一下十三朝古都洛阳深厚的文化底蕴和生活意趣（古人曾感言“若问古今兴废事,请君只看洛阳城”、“谁家玉笛暗飞声，散入春风满洛城”）：邙山古墓群、龙门石窟、白马寺、关林、白云山……都值得大家玩味。洛阳、洛阳，洛水之阳。这里谨摘录三国时期著名文学家曹植所写的浪漫主义名篇《洛神赋》作为本段文字的结束，以表达我们对这座历史文化名城的尊崇之情，并借古人吉言，为我们这个以计算机处理老祖宗创造的无与伦比的宝贵财富——中文为己任的学术会议平添一点人文色彩：“……其形也，翩若惊鸿，婉若游龙，荣曜秋菊，华茂春松。髣髴兮若轻云之蔽月，飘飖兮若流风之回雪。远而望之，皎若太阳升朝霞。迫而察之，灼若芙蕖出渌波……柔情绰态，媚于语言。奇服旷世，骨象应图。……微幽兰之芳蔼兮，步踟躇于山隅。……含辞未吐，气若幽兰。华容婀娜，令我忘餐……浮长川而忘反，思绵绵而增慕……”

编 者

2011 年 6 月 15 日

目 录

Ⅰ 词法、句法、语义和篇章分析

异种语料融合方法：基于统计的中文词法分析应用……孟凡东 徐金安 姜文斌 刘 群(3)
汉语词法分析中上文和下文孰重孰轻……于江德 王希杰 樊孝忠(9)
基于边界熵和卡方统计量的多领域适应性中文分词方法……韩冬煦 常宝宝(16)
基于“大词”实例的中文分词研究……修 驰 宋 柔(22)
统计与词典相结合的领域自适应中文分词……张梅山 邓知龙 车万翔 刘 挺(28)
一种利用注疏的《左传》分词新方法……徐润华 陈小荷(34)
基于属性信息的中文人名消歧……李 丽 孙甲申 王小捷 李 江 宋占江(40)
基于规则与统计的维吾尔族人名识别研究……赛依旦·阿不力米提 吐尔根·依布拉音(47)
基于信息库的新词词汇共性分析与教学策略研究……刘善涛 李 敏(52)
面向自动分词的三音节新词语构词法研究……徐艳华(58)
基于机器学习方法与搜索引擎验证的缩略语预测……焦 妍 王厚峰(64)
哈萨克语通用词汇自动提取方法研究与实现……王雅莉 古丽拉·阿东别克(70)
论蒙古语词素切分的实现……通拉嘎 赵小兵(77)
自动词性标注中语法因素和词汇因素对英汉语的不同影响……邢富坤 宋 柔(83)
藏语不规则动词的信息标注方法……江 荻(90)
汉语复合名词短语特征结构的标注研究……万菁 姬东鸿 任函 冯文贺(94)
树库中的歧义组合考察……李艳娇 杨尔弘(100)
基于树结构模式挖掘的非监督中文短语结构句法分析……张晓甜 赵 海(106)
基于主动学习的中文依存句法分析……陈 鑫 车万翔 刘 挺(112)
最大生成树算法和 Nivre 算法相结合的中文依存关系解析
……周惠巍 黄德根 高 洁 杨元生(118)
面向移进-归约句法分析器的单模型系统融合算法……马 骥 朱慕华 肖 桐 朱靖波(124)
一种基于 MIRA 和遗传算法的句法分析模型构造方法……王 丹 姬东鸿 黄 玮(131)
基于统计方法的蒙古语依存句法分析模型……斯·劳格劳 华沙宝 萨如拉(137)
基于依存树距离的语义角色识别方法……王 鑫 穗志方 李 芸(143)
服务于内容侧面发现的框架识别……王 荀 李素建 宋 涛 姜伯平(149)
越南语给予类动词的语义结构和层级分类研究……林 丽 毕玉德(155)
基于特征结构的汉语主谓谓语句语义标注研究……陈 波 姬东鸿 孙 程 吕 晨(161)
基于隐最大熵原理的汉语词义消歧方法……张仰森 黄改娟 苏文杰(167)
全词消歧的序列标注方法……周 云 王 挺 易绵竹 王之元(173)
基于词典的半指导学习古汉语全文词义标注……张颖杰 李 斌 陈家骏 陈小荷(180)
基于粗糙集方法的共指消解……贾修一 张亚兵 陈家骏 商 琳(186)
中文文本蕴含的推理模型……徐 幸 王厚峰(192)
基于 FrameNet 框架关系的文本蕴含识别……张 鹏 李国臣 李 茹 刘海静 石向荣(200)
面向自动理解的汉语明喻句的可计算性考察……宋 纯 李 斌 曲维光 陈小荷(206)
汉英词汇隐喻属性的对比分析与互增益技术……匡海波 李 斌 王嘉灵 王 帅 陈小荷(212)

Ⅱ 语言资源建设及相关技术

中文 CCG 树库的构建……宋 彦 黄昌宁 揭春雨 (221)
中文语义依存树库构建及自动分析技术……邵艳秋 邱立坤 梁春霞 毛 宁 (228)
语义角色句法实现的词汇语义制约信息库的建设及其应用……周明海 亢世勇 (234)
基于 HNC 的汉语词语知识库改进……王青海 马海慧 池毓焕 李 颖 董凌冲 (240)
《现代维吾尔语语法信息词典》数据库建设的研究
……加米拉·吾守尔 瓦依提·阿布力孜 吐尔根·依布拉音 (246)
构建大规模的汉语事件知识库……周 强 王俊俊 陈丽欧 (249)
百科知识工程……田 野 王渝丽 刘 康 赵 军 (256)
基于多特征表示的本体概念挂载
……徐立恒 刘 洋 来斯惟 刘 康 田 野 王渝丽 赵 军 (262)
相似词获取的集成方法……石 静 邱立坤 王 菲 吴云芳 (269)
领域本体构建中关系辅助判断技术研究……张晓莹 张桂平 王裴岩 (276)
蒙古文 WordNet 名词同义词集合构建算法……哈 斯 那顺乌日图 (283)
面向自然语言处理的韩国语隐喻知识库构建研究……徐 超 毕玉德 (288)

Ⅲ 机器翻译技术与系统

面向小领域的可信机器翻译技术研究……李贤华 于 淼 吕雅娟 (297)
规则和统计相结合的中文地址翻译方法……于 淼 吕雅娟 苏劲松 李贤华 (303)
粘贴模型在依存语法统计机器翻译中的应用……张 育 李良友 贡正仙 周国栋 (309)
基于反向转录语法的机器翻译混合解码策略……张 浩 肖 桐 朱靖波 (316)
统计机器翻译中的源语言重排序方法研究……梁芳丽 李 淼 李 文 陈 雷 乌达巴拉 (323)
一种适用于机器翻译的汉语分词方法……李博渊 奚 宁 黄书剑 张建兵 陈家骏 (329)
一种基于句法的用于汉英翻译的预调序方法……吴秋锋 黄书剑 戴新宇 陈家骏 (336)
基于双语平行语料的中文缩略语提取方法……刘友强 李 斌 奚 宁 陈家骏 (342)
基于关联度的汉藏多词单元等价对抽取方法……诺明花 刘汇丹 吴 健 丁治明 (348)
越汉双语句子自动对齐研究初步……陈坚忠 李 鹏 孙茂松 (354)

Ⅳ 智能检索（信息检索、信息抽取、文本挖掘、文本分类、文本聚类、自动文摘、问答系统、话题跟踪、情感分析、文本倾向性分析等）

基于 CRFs 的评价对象抽取特征研究……王荣洋 鞠久朋 李寿山 周国栋 (365)
基于相似度线性加权方法的检索结果聚类研究……刘海波 郑德权 赵铁军 (371)
基于用户点击信息检索评价方法综述……肖冬青 杨沐昀 李 生 齐浩亮 赵铁军 (377)
基于并列结构的概念实例和属性的同步提取方法……李文杰 穗志方 (382)
事件词驱动的文本事件信息结构初探……曾青青 杨尔弘 (388)
中文维基百科的结构化信息抽取及词语相关度计算……张红春 何婷婷 涂新辉 周琨峰 (394)
基于网页中深度并列结构的实例提取算法……张星星 穗志方 (400)
基于电子商务用户行为的同义词识别……张书娟 董喜双 关 毅 (406)
基于维基百科和模式聚类的实体关系抽取方法……张苇如 孙 乐 韩先培 (413)

基于航空领域本体知识库的语义检索研究 ..... 李伟刚　张克亮　王慧兰 (419)
基于依存关系的旅游景点评论的特征-观点对抽取 ..... 吴苏红　王素格 (425)
基于依存句法和短语结构句法结合的金融领域事件元素抽取 ..... 孟 雷　丁 效　秦 兵　刘挺 (432)
生物医学文献蛋白质关系抽取——从实体识别到网络构建
..... 杨志豪　赵哲焕　李彦鹏　胡运翠　谢 冬　林鸿飞 (438)
基于权重标准化 SimRank 与半监督学习的产品属性归类 ..... 杨 源　马云龙　林鸿飞 (444)
商品品牌名称挖掘 ..... 何正焱　王厚峰 (450)
基于 LDA 模型的文本聚类研究 ..... 董婧灵　李 芳　何婷婷　涂新辉　万 剑 (455)
基于跨语言广义向量空间模型的跨语言文档聚类方法 ..... 唐国瑜　夏云庆　张 民　郑 方 (462)
文本摘要中的句子抽取方法研究 ..... 张龙凯　王厚峰 (468)
面向冗余度控制的中文多文档自动文摘 ..... 王红玲　黄超超　张明慧　周国栋 (474)
面向开放的限定领域的交互式问答语料分析
..... 张耀允　王晓龙　王 轩　徐睿峰　侯永帅　范士喜 (480)
基于层次聚类的网络新闻热点发现 ..... 彭楠赟　王厚峰　凌晨添 (487)
基于话题模型的科技文献话题发现和趋势分析 ..... 贺 亮　李 芳 (493)
基于不平衡数据的中文情感分类 ..... 王中卿　李寿山　朱巧明　李培峰　周国栋 (500)
情感分类中不同主动学习策略比较研究 ..... 居胜峰　王中卿　李寿山　周国栋 (506)
部分监督的音乐情感分类 ..... 王 静　朱慕华　胡明涵 (512)
基于语境歧义词的句子情感倾向性分析 ..... 宋艳雪　张绍武　林鸿飞 (519)
基于语义块的事件倾向性分析研究 ..... 韦向峰　张 全　缪建明　池毓焕 (527)
基于贝叶斯及多模式串模糊匹配算法的不良短消息甄别混合模型
..... 张文波　蒋春华　姚天昉 (533)
博客中重复评论发现 ..... 刁宇峰　王 昊　林鸿飞　杨 亮 (540)
基于维基百科层次分类框架的主题推荐系统的研究
..... 谢 科　刘奕群　岑荣伟　马少平　茹立云　杨 磊 (547)
面向对话语料的标签推荐 ..... 房冠南　袁彩霞　王小捷　李 江　宋占江 (553)

V　其他

一种表征字符信息量的法则：Character's Law ..... 李国华　咎红英 (563)
一个支持人工校对的中文简繁体转换工具 ..... 张小衡 (569)
基于 Unicode 编码的藏文转写拉丁文本的算法 ..... 康才畯　江 荻 (576)
面向语音识别错误恢复的澄清式疑问句生成 ..... 于 东　贾 磊　徐 波 (581)
基于词汇评分的汉语作文自动评分 ..... 彭星源　柯登峰　赵 知　陈振标　徐 波 (588)

Ⅰ 词法、句法、语义和篇章分析

异种语料融合方法：基于统计的中文词法分析应用*

孟凡东[1,2]，徐金安[1]，姜文斌[2]，刘 群[2]
[1]北京交通大学 计算机与信息技术学院，北京 100044
[2]中国科学院 计算技术研究所 智能信息处理重点实验室，北京 100190
E-mail: {mengfandong, jiangwenbin, liuqun}@ict.ac.cn; jaxu@bjtu.edu.cn

摘 要：基于统计的中文词法分析往往依赖大规模标注语料，语料的规模和质量直接影响词法分析系统的性能。高覆盖率、高质量的语料资源非常有限，而且适用于不同领域的语料往往具有不同的分词和词性标注标准，难以直接混合使用，从而导致既有资源未能充分利用，分词精度下降等问题。针对该问题，本论文提出了简单有效的异种语料的自动融合方法，并通过实验验证了提案方法的有效性、较强的实用性以及对多种语料融合的可扩展性。

关键词：词法分析；语料融合；领域适应

A Method of Merging Corpora in Different Annotation Standards: An Application on Statistics Chinese Lexical Analysis

Meng Fandong[1,2], Xu Jin'an[1], Jiang Wenbin[2], Liu Qun[2]
[1] School of Computer and Information Technology, Beijing Jiaotong University, Beijing 100044
[2] Key Laboratory of Intelligent Information Processing, Institute of Computing Technology,
Chinese Academy of Sciences, Beijing 100190
E-mail: {mengfandong, jiangwenbin, liuqun}@ict.ac.cn; jaxu@bjtu.edu.cn

Abstract: Large scale manually annotated corpora are usually used on research of statistical Chinese lexical analysis. The scale and quality of corpora will affect the performance of statistical lexical analysis directly. Corpora in high quality and high rate of coverage are very valuable but limited, and it is very difficult to combine corpora of different domains directly since they are different in segmentation and part of speech (POS) tagging standards. These problems make it difficult to utilize existing resources and result in performance decline of Chinese lexical analysis. To address this domain adaptation issue, this paper presents a simple but effective strategy to optimize the performance and domain adaptability of Chinese lexical analysis by merging different domains' corpora automatically. Our experiments verified the validity, stronger practicability, and extensibility to multiple corpora of our method.

Keywords: lexical analysis; merging corpora; domain adaptation

1 前言

词法分析是自然语言处理领域的基础性研究课题之一，词法分析的精度直接影响自然语言处理后续工作的效果。基于统计的词法分析很大程度上依赖于语料库，加大训练语料，可以直接提高词法分析的精度。但是，手工标注大规模语料代价昂贵。并且，不同领域的语料切分和标注的标准往往不同，难以直接混合使用。图 1 以人民日报语料和宾州中文树库语料为例，具有不同的切分和词性标注标准，在人民日报语料中“高新技术”为一个词，标注为名词（n），在宾州树库中，“高新技术”被分为“高”“新”“技术”，并分别标注为形容词（JJ）、形容词（JJ）和名词（NN）。并且，这两种语料的词性标注集也不同，名词的标注分别是 n 和 NN。



图 1 人民日报语料（上面）和宾州中文树库语料（下面）的分词和词性标注标准举例

* 本文承中央高校基本科研业务费专项资金项目(2009JBM027)和国家自然科学基金项目(60873167, 60736014)的资助。

针对上述问题，Jiang et al.(2009)提出了一种基于错误驱动的方法。利用源语料信息，将其分词和词性标注标准作为特征指导目标分析器，使其产生更好的效果。解码时，首先用源词法分析器对测试语料切分，再用目标词法分析器进行第二次切分，此时以第一次的切分结果为特征，即利用源语料指导目标词法分析器。该方法明显地提高了词法分析精度，是目前中文词法分析中效果最好的方法之一。但是该方法的解码过程略为复杂，不如一次解码的效率高。

本文在 Jiang et al.(2009)基础上提出了异种语料的自动融合策略，以此提高中文词法分析的精度。本方法的思想是先将源语料的分词和词性标注标准进行转化，使其与目标语料的一致，再将转化后的语料与目标语料融合，训练一个新词法分析器。利用这个新的词法分析器可以直接进行解码，不需要二次解码。实验结果表明，本方法可以明显提高中文词法分析精度。与 Jiang et al.(2009)的方法相比，本方法与其具有相当的词法分析性能，甚至比其略高。并且，本方法具有更快的词法分析速度，只进行一次解码，简化了解码步骤，更具有实用性。而且，本方法可用于进一步融合其他领域的语料，更好地提高词法分析性能。因此，本方法更具有可扩展性。

本文在第 2 节简要介绍采用的词法分析方法，第 3 节详细阐述语料自动融合思想，第 4 节是实验及结果分析，第 5 节是对本文的总结与展望。

2 中文词法分析方法

本文采用判别式的词法分析方法。将分词和词性标注问题转化为字符（汉字）分类问题。根据 Ng and Low (2004)的方法，分词采用四种位置标记，b 表示词首，m 表示词中，e 表示词尾，s 表示单个汉字独立成词。即一个词只可以被标记成 s（单字词）或 bm*e（多字词）。联合分词与词性标注就是对于每个字，有位置标记和词性标记，例如“e_v”，表示一个动词的词尾。

2.1 分词特征模板

根据 Ng and Low (2004)的方法，用 C_0 表示当前的汉字，C_{-i} 表示 C_0 左边第 i 个汉字，C_i 表示 C_0 右边第 i 个汉字。$Pu(C_i)$ 用于判断当前汉字 C_i 是否为分隔符(是就返回 1，否则返回 0)。$T(C_i)$ 用于判断当前汉字 C_i 的类别：数字，日期，英文字母，和其他(分别返回 1，2，3 和 4)。

表 1 特征模板

序号	特征模板
1	$C_i (i=-2...2)$
2	$C_i C_{i+1} (i=-2...1)$
3	$C_{-1} C_1$
4	$Pu(C_0)$
5	$T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)$

1: **Input:** Training examples (x_i, y_i)
2: $\vec{\alpha} \leftarrow 0$
3: **for** $t \leftarrow 1 \cdots T$ **do**
4: **for** $i \leftarrow 1 \cdots N$ **do**
5: $z_i \leftarrow \arg\max_{z \in GEN(x_i)} \Phi(x_i, z) \cdot \vec{\alpha}$
6: **if** $z_i \neq y_i$ **then** $\vec{\alpha} \leftarrow \vec{\alpha} + \Phi(x_i, y_i) - \Phi(x_i, z_i)$
7: **Output:** Parameters $\vec{\alpha}$

图 2 感知机训练算法的伪代码

表 1 描述了分词和词性标注的特征模板。假设当前分析的汉字是“450 公里”中的“0”，特征模板生成的特征：$C_{-2}=4$，$C_{-1}=5$，$C_0=0$，$C_1=$公，$C_2=$里；$C_{-2}C_{-1}=45$，$C_{-1}C_0=50$，$C_0C_1=$0公，$C_1C_2=$公里；$C_{-1}C_1=5$公；$Pu(C_0)=0$；$T(C_{-2})\,T(C_{-1})\,T(C_0)\,T(C_1)\,T(C_2)=11144$。

2.2 训练算法

本文采用了 Collins（2002）的平均感知机训练算法。训练的过程就是学习一个从输入 $x \in X$ 映射到输出 $y \in Y$ 的判别模型，X 是训练语料中的句子集合，Y 是相应的标记结果。Jiang et al.(2009)中使用了 $GEN(x)$ 函数列举输入 x 的所有候选结果，表示每个训练实例 $(x, y) \in X \times Y$ 映射到特征向量 $\Phi(x, y) \in R^d$，对于一个特征向量，$\vec{\alpha} \in R^d$ 是与其对应的参数向量。对于一个输入的汉字串 x，

目的是找到一个满足下式的输出结果$F(x)$：

$$F(x) = \arg\max_{y \in GEN(x)} \in \Phi(x, y) \cdot \vec{\alpha} \quad (1)$$

其中$\Phi(x,y) \cdot \vec{\alpha}$表示特征向量$\Phi(x,y)$和参数向量的内积。本文沿用此方法。

图 2 描述了感知机训练算法。本文使用了“平均参数”技术（Collins, 2002）避免过拟合。

3 语料自动融合

本文采用自动融合语料的方法提高词法分析的精度。基本流程如下（流程如图 3 所示）：

1、将源语料（语料 1）转化为与目标语料切分和词性标注标准一致的语料（语料 3）。

2、将目标语料（语料 2）和转化后的语料（语料 3）合并，成为更大的语料（语料 4）

3、用语料 4 训练新的分词和词性标注模型。本方法的关键是第一步。

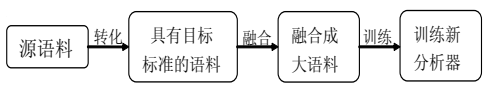

图 3　方法流程图

3.1 分词和词性标注标准的自动转化

为了方便说明，“源语料”表示其他领域的语料，“目标语料”表示当前训练词法分析器所需要的语料；“源标准”表示“源语料”的分词和词性标注标准，“目标标准”表示“目标语料”的分词和词性标注标准；“源分析器”表示用“源语料”训练的词法分析器，“目标分析器”表示用“目标语料”训练的词法分析器。

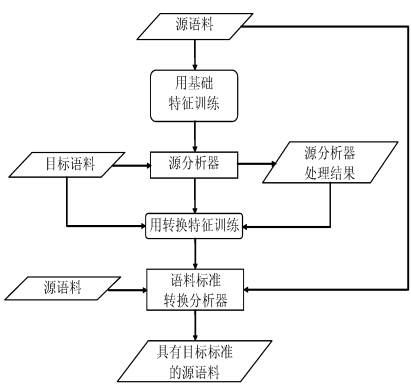

图 4　将源语料转化为具有目标标准的语料

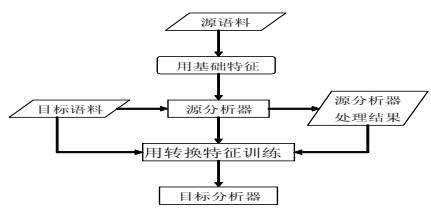

图 5　Jiang et al.(2009)的训练流程

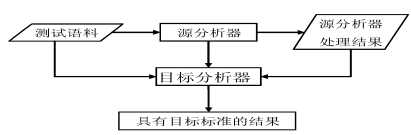

图 6　Jiang et al.(2009)的解码流程

首先，为了获取源标准，用源语料训练一个源分析器，该分析器是用来处理目标语料使其带有源标准的。然后利用这个带有源标准的语料（作为源转化特征）和目标语料训练一个从源标准到目标标准的转换分析器。最后，用这个转换分析器处理源语料（并将源语料作为源转换特征），使其具有目标标准。经过以上步骤，便成功地将源语料转化为具有目标标准的语料。图 4 描述了语料标准转化的过程。该方法是合理的，因为目标语料经源分析器处理后，分词和词性标注的格式与源语料很相似，当然也存在一定的噪声，因为源分析器的精度不是百分之百。但是再通过转化训练，将源标准转化为目标标准的同时，起到了修正源分析器错误结果的作用，使得模型具有一定的容错能力。最后，再用该模型处理源语料，便可将源语料转化为具有目标标准的语料。

表 2 描述了转换特征的一个例子。假设正在分析汉字串“美国副部长”中的“副”字，该汉字串经源分析器处理后被切分和标注为“美国/ns　副/b　部长/n”，而目标语料中切分和标注情况

为“美国/NR　副部长/NN”。以联合分词与词性标注为例，语料标准转化过程如下：经源分析器处理后，汉字串“美国副部长”中的“副”字被标记为“副/s_b”，表示“副”是单字副词，经过转换后“副”字被标记为“副/b_NN”，表示一个名词的词首。除了“$@=s$”和“$@=s_b$”以外，转换特征和基础特征基本一致，其中“$@=s$”表示源分析器标记当前汉字的位置信息为“s”，单子词；“$@=s_b$”表示源分析器标记当前汉字的位置和词性信息为“s_b”，单字副词。

表2　用于训练转化模型的转换特征

基础特征	$C_{-2}=$美　$C_{-1}=$国　$C_0=$副　$C_1=$部　$C_2=$长
	$C_{-2}C_{-1}=$美国　$C_{-1}C_0=$国副　$C_0C_1=$副部　$C_1C_2=$部长
	$C_{-1}C_1=$国部
	$Pu(C_0)=0$
	$T(C_{-2})T(C_{-1})T(C_0)T(C_1)T(C_2)=44444$
分词转换特征	$@=s$
分词与词性标注转换特征	$@=s_b$

3.2　训练与解码

将上面处理好的具有目标标准的源语料与目标语料合并，用这个合并后的大语料训练，便可得到一个新的词法分析器。训练新的词法分析器只用基础特征，不需要转化特征。

本方法与 Jiang et al.(2009)的方法有些类似，但也有很大的不同。图 5 和图 6 分别是 Jiang et al.(2009)方法的训练流程图和解码流程图。Jiang et al. (2009)的方法旨在利用源语料信息，将其分词和词性标注标准作为特征指导目标分析器。该方法在解码时分为两步，首先用源词法分析对测试语料进行切分，然后再用目标词法分析器切分一次，并以第一次的切分结果为特征指导第二次的切分。此方法取得了很好的效果，但是需要两次解码，增加了解码的复杂性。本论文方法旨在利用语料自动融合技术，训练出一个更好的词法分析器。其优点体现在词法分析精度高，只需一次解码，更具有实用性。并且，本方法还可以融合多领域语料，不限于两种，更具有扩展性。

4　实验与结果分析

4.1　实验数据、环境和评测方法

本文实验采用人民日报语料和宾州中文树库语料 5.0。这两种语料库具有不同的分词和词性标注标准，词性标注集也不同 (例如图 1 中的描述)。人民日报训练语料与测试语料的句子数分别为 100344 和 19007，宾州树库训练语料与测试语料的句子数分别为 18074 和 348。

训练和解码实验环境。操作系统：Red Hat Enterprise Linux AS，X64；处理器：Quad-Core AMD Opteron Processor 8347HE，1.9GHZ；内存：64G；编译环境：GCC4.1

本文采用 F-measure 来评价词法分析精度, $F_1=2PR/(P+R)$，其中 P 是准确率，R 是召回率。

4.2　结果与分析

表 3 的前三行是单独的在相应的语料上利用感知机算法训练的模型，即 Baseline 模型。表中 PD 表示人民日报语料，CTB 表示宾州中文树库语料，PD→CTB 表示将人民日报语料融入到宾州树库语料中，CTB→PD 表示将宾州树库语料融入到人民日报语料中，PD+CTB 表示将人民日报语料与宾州树库语料直接合并。“—”表示没有做该部分实验，因为 PD 与 CTB 词性标注集不同。

分别比较表 3 的第一行和第三行，可以看出联合分词与词性标注要比单独分词的精度高，因为词性标注信息相当于是额外的特征（Ng and Low, 2004）。同时可以看出，用 PD 训练模型，并且

在 CTB 上进行测试，无论是分词还是联合分词与词性标注，精度都会下降很多（F_1 值只有不到92%），比单独在 CTB 上训练的模型精度（97%以上）低很多。虽然 PD 比 CTB 大很多，仍然不会提高精度，因为不同领域的语料的分词和词性标注标准不同。然而，利用本方法将 PD 融入 CTB 后，在 CTB 上做测试，无论单独分词还是联合分词与词性标注，F_1 值都有很明显的提高，较单独 CTB 训练的模型提高 0.81 个百分点，联合分词与词性标注的 F_1 值分别提高了 0.38 个百分点（不算词性标注）和 0.96 个百分点（算词性标注）。将 CTB 融入 PD 后，在 PD 上测试，单独分词和联合分词与词性标注的 F_1 值也都有提高。因为 CTB 语料相对 PD 语料太少，只要不到其五分之一，因此 F_1 值的提高不明显。直接将 PD 与 CTB 合并训练，无论在 PD 还是 CTB 上测试，F_1 值都下降很多。尤其是在 PD 上测试，F_1 值急剧下降，可见不同标准语料直接合并产生的负面影响也很大。

表 3　单独分词、联合分词与词性标注的结果

训练	测试	分词结果（ F_1%）	联合分词和词性标注的结果（F_1%）	
			不算词性标注	算词性标注
PD	PD	97.27	97.57	94.54
PD	CTB	91.67	91.79	—
CTB	CTB	97.30	97.77	93.10
PD->CTB	CTB	98.11	98.15	94.06
CTB->PD	PD	97.29	97.58	94.54
PD + CTB	CTB/PD	95.26/83.58	—	—

表 4 中，源语料是 PD，目标语料是 CTB，测试集是 CTB 测试集。从表 4 可以看出，本方法与 Jiang et al.(2009)的方法相比，分词和联合分词与词性标注的性能基本与其相当，甚至略高一些，因为大语料具有更高的词语覆盖率，而如果遇到没有出现的词语，基于错误驱动的修正方法仍然无法很好的处理。而且本方法的解码速度快很多，其中分词速度提高了 34.15%，联合分词与词性标注的速度提高了 53.38%。并且，解码步骤简单，只有一步，实用性更强。

表 4　方法比较

项目 / 方法	分词（F_1%）	联合分词与词性标注（F_1%）/ 时间（秒）	
		不算词性标注	算词性标注
Jiang et al.(2009)	97.93	97.99/3.28	93.81/101.40
本方法	98.11	98.05/2.16	94.06/47.27

表 5 中，“错误融合法”指的是首先利用目标分析器处理源语料，使其具有目标标准，再将处理后的源语料合并到目标语料中，再由这混合后的大语料训练出新的目标分析器。该方法看似更简单，但源语料经目标分析器处理后，虽然接近目标标准，却有很多错误的切分结果，直接使用会产生负面影响。表 5 的结果表明，利用该方法得到的分词结果比融合语料前只提高 0.05 个百分点，不排除是融入大语料提高了词语覆盖率所起的作用。而且联合分词与词性标注的 F_1 值比融合语料前低很多，可见融合了含有错误信息的语料将导致词法分析精度的下降。

表 5　错误融合法

训练	测试	分词（F_1%）	联合分词与词性标注（F_1%）	
			不算词性标注	算词性标注
CTB	CTB	97.30	97.77	93.10
PD->CTB	CTB	97.35	96.24	91.34

综上所述，通过一系列实验，从正、反两面都说明了本方法的有效性和较强的实用性。

5 结语

本文提出了一种异种语料的自动融合方法，将该方法应用于中文词法分析，明显地提高了词法分析性能。我们用人民日报语料和宾州中文树库语料进行了实验，并且利用平均感知机算法，分别在人民日报语料、宾州中文树库语料以及融合后的语料上训练模型，对各个模型的分词以及联合分词与词性标注的效果进行了比较，实验结果表明，本方法确实可以提高词法分析精度。

本文还将本方法与 Jiang et al. (2009)的方法进行了比较，本方法在保证了与 Jiang et al. (2009)的方法具有相当性能的情况下，提高了分词以及联合分词与词性标注的解码效率。本方法具有更简单的解码步骤，实用性更强。而且本方法不局限于融合两个领域的语料，更具有扩展性。

接下来，我们将继续研究语料标准的转化方法，以及后续改进的语料合并方法，例如语料加权合并。并且，进一步融合其他领域的语料以提高词法分析精度。

参 考 文 献

[1] Wenbin Jiang, Liang Huang, and Qun Liu. 2009. Automatic Adaptation of Annotation Standards: Chinese Word Segmentation and POS Tagging – A Case Study. *In Proceedings of the 47th Annual Meeting of the Association for Computational Linguistics*.

[2] Hwee Tou Ng and Jin Kiat Low. 2004. Chinese part-of-speech tagging: One-at-a-time or all-at-once? word-based or character-based? *In Proceedings of the Empirical Methods in Natural Language Processing Conference.*

[3] Wenbin Jiang, Liang Huang, Yajuan Lv, and Qun Liu. 2008. A cascaded linear model for joint Chinese word segmentation and part-of-speech tagging. *In Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics.*

[4] Wenbin Jiang, Haitao Mi and Qun Liu. 2008. Word Lattice Reranking for Chinese Word Segmentation and Part-of-Speech Tagging. *In Proceedings of the 22nd International Conference on Computational Linguistics.*

[5] Kun Wang, Chengqing Zong and Keh-Yih Su. 2010. A Character-Based Joint Model for Chinese Word Segmentation. *In Proceedings of the 24th International Conference on Computational Linguistics.*

[6] Zhongguo Li and Maosong Sun. 2009. Punctuation as Implicit Annotations for Chinese Word Segmentation. *In Proceedings of Computational Linguistics.*

[7] Yue Zhang and Stephen Clark. 2007. Chinese segmentation with a word-based perceptron algorithm. *In Proceedings of the 45th Annual Meeting of the Association for Computational Linguistics*.

汉语词法分析中上文和下文孰重孰轻*

于江德[1]，王希杰[1]，樊孝忠[2]
[1]安阳师范学院 计算机与信息工程学院，河南 安阳 455002
[2]北京理工大学 计算机科学技术学院，北京 100081
E-mail：jiangde_yu@tom.com

摘 要：汉语词法分析是中文信息处理的基础，现阶段汉语词法分析的主流技术是基于统计的方法，这类方法的本质都是把词法分析过程看作序列数据标注问题。上下文是统计语言学中获取语言知识和解决自然语言处理中多种实际应用问题必须依靠的资源和基础。汉语词法分析时需要从上下文获取相关的语言知识，但上文和下文是否同样重要呢？为克服仅凭主观经验给出猜测结果的不足，我们对汉语词法分析的分词、词性标注、命名实体识别这三项子任务进行了深入研究，对比了上文和下文对各个任务性能的影响，在国际汉语语言处理评测 Bakeoff 多种语料上进行了封闭测试，采用分别表征上文和下文的特征模板集进行了对比实验，结果表明，上文和下文对汉语分词和中文命名实体识别性能的影响差别较大，对汉语词性标注的性能影响差别较小。

关键词：汉语词法分析；上下文；分词；词性标注；命名实体识别；特征模板

Which Is More Effective for Chinese Lexical Analysis: Above-context Versus Below-context?

Yu Jiang-de[1], Wang Xi-jie[1], Fan Xiao-zhong[2]
[1] School of Computer and Information Engineering, Anyang Normal University, Anyang 455002
[2] School of Computer Science and Technology, Beijing Institute of Technology, Beijing 100081
E-mail：jiangde_yu@tom.com

Abstract: Chinese lexical analysis is a foundational task for Chinese information processing. At the current, the mainstream technology of Chinese lexical analysis is based on statistical methods. These methods treat the analysis process as a sequence data tagging problem. Context is the necessary resource not only for obtaining linguistic knowledge in statistical linguistics but also for solving the problem in natural language processing. Chinese lexical analysis needs the help of correlative context. However, is above and below the same important? To overcome the lack of giving the result by the subjective experience, we studied the contribution of above and below for Chinese lexical analysis via the large number of experiments about word segmentation, POS tagging and named entity recognition. Closed evaluations are performed on many kinds of corpus from the international Chinese language processing Bakeoff, and comparative experiments are performed on different feature templates which describe above-context and below-context. Experimental results show that the performance is very different by the below-context and by the above-context.

Keywords: Chinese lexical analysis; context; word segmentation; POS tagging; named entity recognition; feature template

1 引言

汉语词法分析主要包括汉语自动分词、词性标注和命名实体识别三项子任务，它是中文信息处理领域的一项基础性研究课题[1,2]。现阶段，基于统计的方法是汉语词法分析的主流技术。对于汉语分词任务而言，自从 2002 年 Xue 等[3]在第一届国际计算语言学会下属的汉语语言处理特别兴趣研究组（special interest group on Chinese language processing，SIGHAN）研讨会上提出基于字标注（character-based tagging）的汉语分词技术后，近年来，基于类似思想的汉语分词技术得到了广泛关注和研究[4-6]。对于汉语词性标注任务，由于基于规则的方法适应性较差，所以已有的研究中

* 基金项目：由河南省高等学校青年骨干教师项目（2009GGJS-108）和河南省教育厅自然科学研究项目（2011B520004）支持。

基于统计语言模型的方法居多，已采用的统计语言模型主要有 N 元语法模型（N-gram）[7]、隐马尔科夫模型（hidden Markov model, HMM）[2]、最大熵模型（maximum entropy, ME）[8]、条件随机场（conditional random fields，CRFs）[9]、支持向量机（support vector machine, SVM）[10]等。在中文命名实体识别任务上，基于 HMM 的方法[2,11]、最大熵模型的方法、CRFs 的方法[12]以及 SVM 的方法是目前比较常见的方法。综合分析这些已有的汉语词法分析研究，都是将汉语词法分析的本质看作是一个序列数据标注问题，借助于统计语言模型实现。

统计语言建模中，上下文扮演着解决问题所需语言知识和资源提供者的重要角色[13]。通常情况下，上下文的选取是基于当前位置左右一定范围进行的，这个固定的范围称为“窗口”。一般情况下，选取的这个“窗口”是对称的，即上文和下文有相同的宽度。但是，这种对称地选取上下文是否合适呢？这取决于在解决具体问题时从宽度相同的上文和下文中获取的语言知识的多少，或者说上文和下文对具体问题解决的贡献情况。为了克服当前仅凭主观经验对称地选择上下文的不足，本文首先对基于统计语言模型的汉语词法分析中上下文的建模进行了详细解析，然后深入研究了汉语词法分析的三个子任务中上文和下文对性能的影响情况，并在国际汉语语言处理评测 Bakeoff 的分词、词性标注、命名实体识别三种语料上进行了封闭测试，采用表征上文和下文的不同特征模板集进行了对比实验，结果表明，上文和下文对汉语词法分析的影响大不相同，而且在三个不同的子任务中上文和下文对性能的影响差别又各不相同。

2 基于统计的汉语词法分析中上下文的建模

基于统计方法的汉语词法分析的实质是将汉语词法分析转化为序列数据标注问题，该问题可使用统计语言模型之一来实现。例如，N-Gram、HMM、ME、CRFs、SVM 等。在纷繁复杂的自然语言中，蕴含着一些内在的语言规律（某种程度上也可以称为语言知识），对一个具体的自然语言处理任务，这些语言规律是不可或缺的。对汉语词法分析这一具体任务而言，这些语言规律外在的表现形式就是句子中字之间、词之间、字与词之间、词与词性之间、字与命名实体之间、词与命名实体之间所存在的内在关联。统计语言模型是采用概率论与数理统计的方法从训练语料中统计出字、词、命名实体等不同粒度语言单位（linguistic unit）及其相应标记之间的内在联系，即在训练语料中这些不同粒度语言单位之间存在的概率分布规律，也就是训练语料中存在的语言规律。由此可见，利用统计语言模型对训练语料上下文进行建模的本质就是要准确描述和刻画上下文中存在的语言规律，其中，上下文“窗口”的设定和窗口中上下文特征的刻画是两个关键问题。

2.1 上下文“窗口”的设定

通常情况下，上下文的选取是基于当前位置前后一定范围进行的，这个固定的范围称为“窗口”。由于语言规律要从“窗口”的上下文中统计获取，所以上下文“窗口”开设的大小要依据具体的任务来设定。图 1 示意了汉语词法分析中可能的上下文“窗口”。图中方框中示意的上下文窗口是“3 个语言单位窗口”，即设定的上下文窗口包含当前语言单位及其前后各一个语言单位共 3 个。具体到基于字标注的汉语分词而言，此时的语言单位是字，上下文窗口所表现出来的上下文

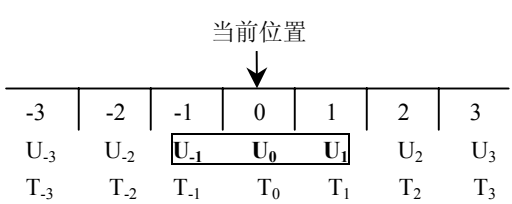

图 1 可能的上下文“窗口”

特征是指由当前字本身以及当前字前后各一个字所构成的特征。当然，上下文窗口也可以是“5语言单位窗口”、“7 语言单位窗口”等。一般情况下所设定的上下文窗口都是关于当前位置对称的，即取当前位置前后相同宽度的上文和下文。可是，在汉语词法分析中这样对称地选取上下文窗口是否合适？这将是本文要回答的问题。

2.2 上下文特征的刻画

上下文“窗口”确定之后，又会面临另一个问题：自然语言中的规律是通过“窗口”中一个个“上下文”样本表现出来的，如何将窗口中的上下文特征描述和刻画出来？在统计语言建模中，通常是采用特征模板来描述和刻画上下文中的语言特征。特征模板的主要功能是定义上下文中某些特定位置的语言成分与某类待预测事件的关联情况。习惯上，特征模板可以看作是对一组上下文特征按照共同的属性进行的抽象。在统计语言模型的训练学习中，每个特征都对应了一组特征函数，这些特征函数对该模型的训练学习至关重要。而这些特征又是通过特征模板扩展来的，所以，设定合适的特征模板就至关重要。

对于汉语词法分析而言，综合分析已有汉语分词、词性标注和命名实体识别三个子任务中特征模板的设定，可以发现，特征模板所刻画出的上下文特征主要是指上下文“窗口”中字、词等语言单位之间所组合成的特征。特征模板可以根据语言单位和当前位置的距离属性组合设定，可以是单个语言单位构成的特征模板。例如，U_{-1}，U_0，U_1 等，其中 U_{-1} 表示当前位置前一个语言单位，U_0 表示当前语言单位，U_1 表示当前位置的后一个语言单位。也可以是窗口内的两个语言单位联合组成的特征模板。例如，$U_{-1}U_0$ 表示当前位置前一个语言单位和当前位置的语言单位构成的特征模板。依此类推还可以有三个语言单位联合组成的特征模板等。

为了回答汉语词法分析中对称地选取上下文是否合适？就需要研究上文和下文对汉语词法分析性能的影响是否大体相等，或者看谁对汉语词法分析的影响更大。为此，我们需要设定仅仅刻画上文（包括当前位置）和仅仅刻画下文（包括当前位置）的特征模板集，以便进行对比实验。在本文后面的实验中后缀-Above 和-Below 示意的特征模板集分别表示相应模板集的上文部分和下文部分。

3 汉语词法分析中上文和下文孰重孰轻的对比实验及其结果分析

3.1 实验设计

本文分别针对汉语词法分析的三个子任务：汉语分词、词性标注、命名实体识别进行了上文和下文孰重孰轻的对比实验。其中汉语分词中上文和下文的对比实验采用基于字标注的汉语分词技术，使用四词位标注集（B、M、E、S）和条件随机场来具体实现。汉语词性标注中上文和下文的对比实验采用最大熵模型具体实现。中文命名实体识别中上文和下文的对比实验采用基于字标注的命名实体识别技术，采用条件随机场具体实现。

3.2 汉语分词中上文和下文孰重孰轻

3.2.1 实验数据集和性能评估

本组实验采用的训练语料和测试语料是 SIGHAN 举办的第二届国际汉语分词评测 Bakeoff2005 所提供的语料中的简体中文语料，分别是北京大学（PKU）和微软亚洲研究院（MSRA）提供的训练语料和测试语料。

在对汉语分词性能进行评估时，采用了常用的 5 个评测指标：准确率（*P*）、召回率（*R*）、综合指标 *F* 值（*F*）、未登录词召回率（*OOV RR*）、词表词召回率（*IV RR*）。准确率表示在切分的全