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	$C_iC_{i+1}(i=-2...1)$
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。

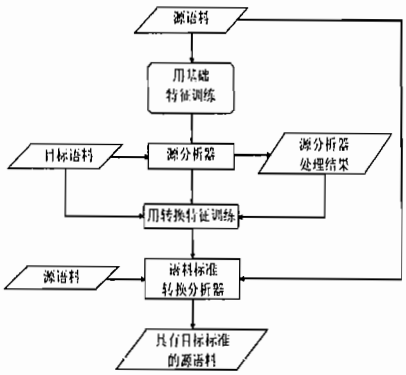

图 4　将源语料转化为具有目标标准的语料

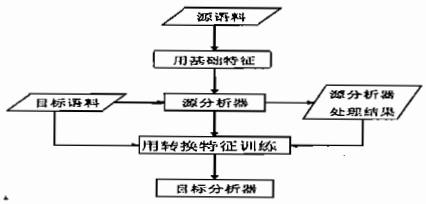

图 5　Jiang et al.(2009)的训练流程

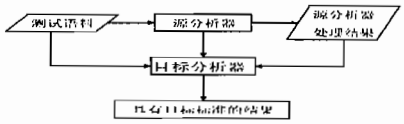

图 6　Jiang et al.(2009)的解码流程

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.*