汉语分词及词性标注自动校验方法研究

钱揖丽　张虎

（山西大学计算机科学系, 太原 030006）

E-mail: qyl@sxu.edu.cn

摘要： 大规模的标注语料库是语料库语言学发展的重要基础。随着许多科学研究的进一步开展，我们对语料的加工质量提出了更高的要求。本文采用基于上下文搭配的规则和统计相结合的自动校验方法，对机器切分标注语料进行处理，并把自动校验过程中获取的信息，应用于语料库的构建，即采用滚动式的方法，建立大规模的、具有更高加工质量的标注语料库。

关键词: 自动分词; 词性标注; 自动校验; 语料库; 质量保证

引言

大规模的标注语料库是语料库语言学发展的重要基础。中文信息处理领域的许多科学研究，都是在大规模标注语料库的基础上进行的。计算机需要对大量的文本进行处理，对语料加工质量的要求也越来越高。而分词和词性标注是语料库加工的重要组成部分，是中文信息处理不可逾越的重要基础之一。目前，机器自动分词和词性标注的方法很多，但是分词和词性标注的质量仍然不能很好地满足实际应用的需要。

目前，国内已经建立了几个较大规模的切分和词性标注汉语语料库，包括清华大学的200万字的平衡语料库和北京大学与富士通合作开发的人民日报语料库。而在提高分词和词性标注质量方面的研究也在广泛进行着。但是，从对现有语料库进行自动校验的角度考虑，提高分词和词性标注的正确率，确保和提高语料的加工质量，这方面的研究报导还很少。

我们采用基于上下文搭配的规则和统计相结合的方法，提高分词和词性标注的正确率；采用滚动式的方法，实现语料库的构建。首先，我们对一部分机器自动切分标注过的语料进行人工校对，通过人工校对以后的语料和机器自动切分标注语料的对比，从中获取有用的信息，生成知识库。其次，对于下一部分新的机器自动切分标注过的语料，我们先根据已经获取的知识对其进行机器自动校验，进一步提高机器切分标注的水平，然后再对机器校验过的语料进行人工校验。最后，我们再将这部分人工校验前后的语料进行对比，把从中获得的有用信息补充到已有的知识库中。采用这种滚动式的构建方法，随着语料库规模的不断扩大，知识库中的可用信息越来越丰富，机器自动校验的能力越来越强，而所需的人工操作却越来越少。

1 知识库的构建

1.1 知识获取

我们将相同语料的人工切分标注结果同机器切分标注的结果进行对比，获取知识，获得两者切分、标注结果不同的所有情况，并从切分、标注不同的词的前后各抽取一个词根及其词性，初步建立知识库。

1） 词及其切分（wword、word）的表示

表示方式：词根 1 词根 2……词根 n＋切分点 1，切分点 2，……，切分点 n

例　　如：“俄国 ns 化学 n 家 k ”词的部分表示为：“俄国化学家 2，4”

(切分点 2，4 表示在第二个字（即“国”字）及第四个字（即“学”字）之后切分)

2） 词性（wcx、cx）的表示

表示方式：词根 1 的词性/词根 2 的词性/……/词根 n 的词性

例　　如：“俄国 ns 化学 n 家 k ”的词性表示为：“ns/n/k”

3） 知识的表示格式

每一个知识由 wword、wcx、wbefore、wcxbefore、wafter、wcxafter、word、cx、rightcc 和 zongcc 这几部分构成，分别表示错误的切分、错误切分部分的词性、wword 之前一个词、wword 前一词的词性、wword 之后一个词、wword 后一词的词性、正确的切分、正确切分部分的词性、知识正确使用的次数和知识的使用总次数。

例 1： 机器切分标注语料例句：

俄国 ns 化学 n 家 k 门捷列夫 nh 对 p 不 d 同性 f 质的 n

元素 n 进行 v 分类 v 整理 v

人工切分标注语料例句：

俄国 ns 化学 n 家 k 门捷列夫 nh 对 p 不同 a 性质 n 的 u

元素 n 进行 v 分类 v 整理 v

获得知识：

不同性质的 1,3＃d/f/n＃对＃p＃元素＃n＃不同性质的 2,4＃a/n/u＃6＃6

（注：＃表示不同字段之间的分隔）

1.2 知识的初筛选

初建知识库中，wword 的前后词性结构完全相同的所有知识，都被保留下来，产生的知识库规模较大。这样，不仅导致所占的计算机的存储空间过大，而且还使检索知识所用的时间过长，影响搜索的效率。所以我们对知识库中的知识进行比较和分析，对于 wword、wcx、word 及 cx 均相同的知识，如果它们的 wcxbefore、wcxafter 也相同，那么我们只保留这些知识中的一个。通过对初建知识库中的知识进行筛选，把具有相同结构的知识归并为一个知识，这样既减小了知识库的规模，减小了存储空间，又缩短了搜索时间。

1.3 知识库的评价和维护

通过对大量语料的不断学习，随着构建的语料库规模的日渐庞大，我们得到一个信息比较完备的知识库。但是知识库中知识的质量参差不齐，使用频度和正确使用频度各不相同，相差很大。不同质量的知识不仅影响机器自动校验修正的能力，而且还会影响机器自动校验修正的速度和效率。为了提高机器

的自动校验修正能力，我们需要人工对机器自动学习到的知识进行统计，对知识库中的知识进行评价，并根据评价结果对知识库进行维护。

1) 评价参数

知识的使用总次数 zongcc；

知识正确使用的次数 rightcc。

2) 评价函数

我们采用以下两个函数，来对知识库中的每一个知识进行评价。

- 绝对改正数＝改对数－改错数

$$Anet = \text{rightcc} - (zongcc - rightcc) = 2 * rightcc - zongcc \quad (1)$$

- 相对改正率＝绝对改正数 /（改对数＋改错数）×100%

$$Rnet = \frac{Anet}{zongcc} * 100\% = (2 * \frac{rightcc}{zongcc} - 1) * 100\% \quad (2)$$

2 自动校验修正

2.1 自动校验修正机器切分标注结果

即根据已经建立的知识库对机器切分标注的结果进行自动的校验修正，从而进一步提高语料分词及词性标注的正确率。

算法 1. 自动校验修正算法

1）取出机器切分标注语料中的一个词根、词性、前一个词根的词性以及后一个词根的词性，当前词根构成串β，转（2）；

2）搜索知识库，查找与当前情况的词及切分 wword、词性 wcx、前一个词的词性 wcxbefore、后一个词的词性 wcxafter 均相同的记录，即与其相匹配的知识。如果找到，则根据知识中正确的词切分 word 及词性标注 cx 修正当前的错误切分及词性标注，并将该知识的使用次数 zongcc＋1，然后转（4）；如果未找到，则转（3）；

3）在所有知识的 wword（串α）中进行检索，搜索当前词根（串β）。如果检索到，且串β在串α中的起始位置为 1，则将串β和其后面的一个词根链接，构成一个新的串β，转（2）；否则转（4）；

4）判断语料中所有的词是否处理完毕，如果没有，则转（1）；否则结束。

执行以上算法，那么语料处理完毕以后，将在当前语料所在的位置自动新建一个子目录，同名存放经过机器自动校验修正的所有语料文件。

2.2 机器学习

如果我们采用常见的“先学后验”的学习方法，即对机器切分标注结果先进行机器自动校验修正，然后在此基础上再进行人工校验修正的学习过程。这种学习校验过程存在以下问题：

1) 无法识别出机器改对的部分

对机器自动校验修正的结果进行人工校验修正时，我们无法判断哪些部分是机器初分初标

就正确的，哪些部分是经过机器自动校验修正以后才改正的。

2) 无法识别出机器改错的部分

对机器自动校验修正的结果进行人工校验修正时，我们无法判断哪些是机器初分初标错误但未被校验修正的部分，哪些是校验修正不当而导致错误的部分。

所以，“先学后验”的学习方法，只能获得一个最后的结果，而无法了解机器自动校验修正的过程和效果。基于以上问题，我们采用“学验同步”的学习方法。即人和机器同时对机器自动切分标注结果进行校验修正，针对改动部分，机器自动修改生成修改表（如图1所示），人工修改生成人工表（如图2所示）。然后将得到的人工表同修改表作对比（如图3所示），则可以很清楚地了解机器的自动校验修正过程，从而解决上述问题。

算法2. “学验同步”的学习算法

1） 人工表和修改表中都有的知识，表明是经过机器自动校验修正后改正的部分，则相应知识的正确使用次数 rightcc＋1；

2） 人工表有，而修改表没有的知识，表明应该修改，但是机器未能识别，应将其添加到知识库中；

3） 人工表没有，而修改表有的，表明是机器改错的情况。

这样，采用这种“学验同步”的学习方法，我们不仅仅得到了切分和标注正确率都有所提高的语料，而且我们还可以了解和掌握机器学习和校验的情况，从而对机器学习和校验的效果作出评价。

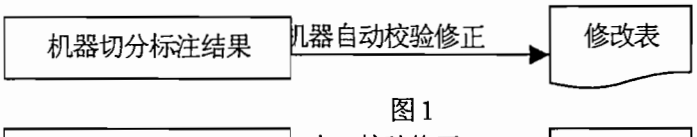

图1

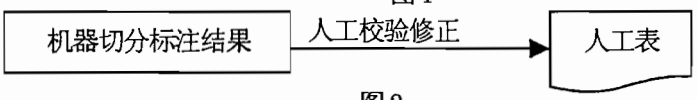

图2

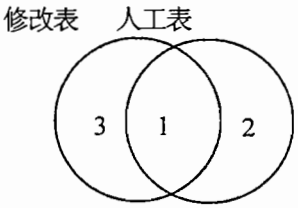

图3

3 实验结果及分析

我们以20万汉语平衡语料作为训练集，通过对机器切分标注结果和人工标注结果的比较建立初始知识库，然后再由机器对此20万汉语平衡语料的机器初分初标结果进行自动校验修正，作封闭测试。通过机器初分初标结果、机器自动校验修正结果以及人工标注结果三者的对比，得到的实验结果如下：

表1 封闭测试实验结果

机器初分初标后语料存在的错误数	3598		
自动校验改正数	3373		
机器自动校验修正后语料存在的错误数	225	机器改错数	52
		机器未识别数	173
自动校验改正率	93．75%		
自动校验改错率	1．44%		

自动校验未识别率	4．81%

我们再根据此知识库对20万语料外的另外4万汉语平衡语料的初加工结果进行自动校验修正，作开放测试，通过机器初分初标结果、机器自动校验修正结果以及人工标注结果的对比，得到的实验结果如下：

表2　开放测试实验结果

机器初分初标后语料存在的错误数	839		
自动校验改正数	680		
机器自动校验修正后语料存在的错误数	159	机器改错数	8
		机器未识别数	151
自动校验改正率	81．05%		
自动校验改错率	0．95%		
自动校验未识别率	18%		

对以上实验结果进行分析，我们可以得出以下几点：

1)　开放测试时，未能识别改正的错误中，大部分都是由于已有的知识库不包含与其匹配的知识所致。

2)　机器自动校验修正的正确率，很大程度上取决于所建的知识库是否全面。如果知识库具有足够的全面性，那么通过对于知识库的搜索匹配，机器自动校验的正确率就会很高。

3)　但是，另一方面，由于要保证全面性，可能会保留过多的知识，会导致知识库规模过大，从而导致存储空间过大和检索速度过慢等问题。所以我们要根据不同的情况，选取合适的阈值，对知识库中的知识进行筛选，从而建立一个既不失效率，又具有相对全面的知识库。这仍然是一个有待于继续研究的问题。

4　结束语

实验测试结果表明，我们通过采用这种基于上下文搭配的规则和统计相结合的自动校验策略，对机器自动切分标注语料进行处理，获得了较高的自动校验正确率，进一步提高了机器分词和词性标注的正确率，提高了语料加工质量；同时，利用已有的经过机器自动分词和词性标注的语料，通过采用滚动式的语料库构建方法，构建了切分标注质量都比较高的400万汉语平衡语料库，为基于初加工语料库的进一步研究奠定了基础。

参考文献:

[1] 刘开瑛. 中文文本自动分词和标注. 商务印书馆. 2000.
[2] 史忠植. 高级人工智能. 科学出版社. 1998.
[3] 刘健. 基于实例的词性标注方法研究. 太原: 山西大学[硕士学位论文]. 2001.
[4] 安世虎. 刘淑辉. 模式匹配问题的进一步研究. 计算机应用研究. 1998. 15(4).
[5] 周强. 詹卫东. 任海波. 构建大规模的汉语语块库. 自然语言理解与机器翻译. 清华大学出版社. 2001.
[6] 高山等. 基于三元统计模型的汉语分词及标注一体化研究. 自然语言理解与机器翻译. 清华大学出版社. 2001.

作者简介: 钱揖丽, 1977 年生, 女, 山西平遥人, 硕士生, 助教, 主要研究领域为中文信息处理。

Research of Verifying Method of Chinese Word Segmentation and Part-of-speech Tagging

QIAN Yili ZHANG Hu
[1](Shanxi University, Taiyuan 030006, China);
E-mail: qyl@sxu.edu.cn

Abstract: The large-scale tagged corpus is the important basis of the development of corpus linguistics. Many reaseaches request corpora with higher processing quality. This paper presents a verifying method based on rules and statistic. This method obtains informations from the corpora's verifying course, and then applies these informations to the building of corpus. We use rolling method, build large-scale Chinese corpus, and we have obtained higher processing quality.

Key words: Chinese word segmentation;Part-of-speech tagging;Verifying;Corpus;Quality ensuring