标注语料机器校对的研究与实践

曲维光　　陈小荷

（南京师范大学文学院，南京，210097）

E-mail:wgqu@pine.njnu.edu.cn

摘要：本文讨论了标注语料校对的质量评价准则，并依此准则对经过机器标注和人工校对后语料的机器自动校对进行研究。利用预处理、基于统计和基于规则的校对过程来提高语料质量。通过实验证明，该方法不仅可以提高机器标注和人工校对后语料的质量，而且对标注语料的机器自动校对也有很好的效果。

关键词：语料校对；标注语料；统计方法；规则方法

Research and Experiment on Annotated Corpus Correction

Qu Weiguang　　Chen Xiaohe

(School of Chinese Language and Literature, Nanjing Normal Univ., Nanjing　210097, China)

E-mail:wgqu@pine.njnu.edu.cn

Abstract: In this paper, we discussed the rules to the evaluation of corpus' quality, and studied the correcting of corpus, which had been automatically segmented and tagged by computer and then corrected manually. Three steps: pre-processing, statistics-based methods and rule-based methods, were used to improve the corpus' quality.

Keywords: corpus correcting; tagged corpus; statistics-based method; rule-based method

一. 引言

带标注语料库的建设是当前语料库语言学和计算语言学研究的热点之一[1]，我国在此项目上投入大量人力和资金。经过多年的不懈努力，目前已经建立起几个规模较大、质量较高的具有分词和词性标注的汉语语料库，例如清华大学的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，以及北京语言大学的留学生中介语语料库。

我们目前正在进行100万字的留学生语料的分词和词性标注。常规的语料加工的基本方法是：首先利用软件进行分词和词性的自动标注，再由语言工作者进行手工校对，以进一步提高语料质量。

在实际工作中，利用软件进行分词和词性标注，具有标注速度快、一致性好等优点，但由于算法的局限性，仍然存在分词错误以及词性标注错误。利用语言工作者对经机器标注过的语料进行校对，可以充分发挥语言学专家知识的作用，尽量将机器标注语料库中存在的错误查找出来并加以纠正。但由于语料规模大，语料校对工作单调枯燥以及语言问题本身的复

曲维光，男，博士研究生，主要研究方向计算语言学和人工智能；陈小荷，男，教授，博士生导师，主要研究方向计算语言学和语料库语言学。

杂性等诸多原因，往往会出现下述几方面的情况：

1. 漏掉一些错误没有发现出来；
2. 一些错误，改了一部分，疏漏了一部分；
3. 由于误操作，会使语料中出现标记集以外的标注符号、将语料内容增加或减少，或者没有按照语料标注的规范进行处理等；
4. 对于一些语言现象，往往会出现不同的人校对成不同的结果；
5. 少数一些地方，会出现将正确的标注改错的情况。

对于以上情况，如何在经过机器分词和词性标注的语料（简称标注语料）以及人工校对语料的基础上利用机器自动校对来进一步提高语料的质量，便成为本课题的主要目标。

一个标注语料质量的好坏，主要由以下几个方面的标准来评估：

1. 正确性　　主要体现在：

a)分词的正确性

分词正确是词性标注正确的前提条件，没有分词的正确性，词性标注便无法正确。但由于歧义切分和未登录词的两大问题[2]，使得分词处理成为中文信息处理领域中富于挑战性的课题。

b)词性标注的正确性

只有经过正确分词的语料才有可能标注正确。但由于兼类词在常用词中的大量存在，往往会使词类排歧非常困难。

2. 一致性

在语料校对的实际中，经常会出现语义及语法功能相同的同一个标注语料序列，被校对成多种不同的修正序列。其中修改错误的应该校正回来，而有些修正序列则都可以说得过去。为了便于语料的正确检索，我们应该从这些候选修改序列中选取唯一的候选序列，以使相同的语料序列校对结果趋同。

3. 普遍认同性

为了保证语料的一致性，我们希望功能相同的标注语料序列校对结果趋同。在诸多候选修正序列都可行的情况下如何从中选择一个作为最终的修正序列，来保证校对结果符合大部分校对者的意图，便称为普遍认同性。

我们的语料自动校对工作，就是在以上的准则指导下进行的。

本文首先对语料库建设以及语料质量评价标准进行了讨论，在第二部分介绍了语料自动校对的一系列方法，并通过第三部分的实验数据证明这些方法的有效性，在最后对系统性能进行总结，指出系统中存在的一些问题、进一步改进的思路，以及我们目前正在开展的工作。

二．标注语料机器校对的方法和步骤

为了更好的描述，我们进行如下定义：

定义 1　词汇集：$WORD=\{word_i \mid i=1,2,\ldots,n\}$，

词性标记集：$POS=\{pos_i \mid i=1,2,\ldots,m\}$，

其中，$word_i$ 和 pos_i 分别为某个汉语的词和词性标记。

我们使用的的词性标记集，与北京语言大学的词性标记集相同，共有 49 种词性标注。

定义 2 标记词项：PAIR={(word, pos) |.pos 是词 word 在语料中某次应用中的词性标记}。

词典：DICT={(word,pos)|pos 是词 word 的词性}。一个正确的标记词项一定是在词典中的。

这里使用北京大学提供的一个月的《人民日报》标注语料进行统计，抽取一个带词性标记的词典，经过适当的词性标记转换，用作校对的参考标准。

定义 3　标注语句：SENT={(e_0,e_1,e_2,…,e_K)，其中对于 $1 \leqslant i \leqslant K$ ，有 $e_i \in PAIR$，e_0 是为叙述方便而在句首增加的空结点}。

定义 4　修正模式：机器标注语句 sent1=(e_{10},e_{11},e_{12},…,e_{1K})，与之对应的人工校对语句 sent2=(E_{20},E_{21},E_{22},…,E_{2L})，若 $\sum_{i=0}^{1I} e_{1i}.word = \sum_{j=0}^{2I} E_{2i}.word$，但 $e_{1I+1} \neq e_{2I+1}$，则分别求得语句 sent1 和 sent2 在 1J 及 2J 取最小值时的子序列：comp1=（e_{1I+1},e_{1I+2},…,e_{1I+1J}）、comp2=（E_{2I+1},E_{2I+2},…,E_{2I+2J}）使得 $\sum_{i=1}^{1J} e_{1I+i}.word = \sum_{j=1}^{2J} E_{2I+j}.word$。

子序列（e_{1I+1},e_{1I+2},…,e_{1I+1J}）、（E_{2I+1},E_{2I+2},…,E_{2I+1J}）称为语料的一个修正模式，记以<comp1,comp2>。comp1 称为原序列，comp2 成为修正序列。这里用Σ代表字符串的拼接。修正模式的集合用 PTN 表示。

例如，以下两句分别为机器标注和人工校对后的语句：

你/nr　知道/vt　不知/vt　道/ng　以前/ff　的/uj　同学/ng　秋/nt　贤/a　跳/vt　班/ng

你/nr　知道/vt　不/dz　知道/vt　以前/nt　的/uj　同学/ng　秋贤/nm　跳/vt　班/ng

从中，我们可以得到 3 个修正模式：

ptn1=<不知/vt　道/ng，不/dz　知道/vt>；　ptn2=<以前/ff,以前/nt>；ptn3=<秋/nt　贤/a,秋贤/nm >

通过上述分析，结合留学生语料的特点，将语料机器校对分成如下三个步骤：

步骤一：预处理　在此阶段，主要解决以下方面的问题：

1. 从经过机器标注和人工校对后的语料中找出标记词项 p∈PAIR，但 p∉DICT。若 p.word 在词典 DICT 中只有唯一的词性 pos，则 p.pos<-pos，否则找出来，由人工修改；
2. 查校对后的语料是否有内容的增减，若有，找出该句子，人工重新校对；
3. 查标注语料是否符合标注格式的要求。语料采用如下格式：

$word_1/pos_1$　$word_2/pos_2$　…，$word_i$ 与 pos_i 之间以符号/分隔，且三者之间没有空格；两个 PAIR 之间由两个空格分开。如果语料中有不符合标注格式的要求，全部改正过来。

经过上述步骤 1，语料成为符合标注格式，没有内容增删，没有非法 pos 标注的语料。后续处理均在预处理后的语料中进行。

步骤 2：基于统计的语料校对　在基于统计校对之前，首先达成如下共识：

A) 语料的校对者是语言问题的专家；

B) 语料的校对者可以将语料中存在的大部分问题校对出来；

C) 语料的校对者可能在工作中出现偶然错误；

我们的基本想法是通过对修正模式进行统计学习[3]，寻找出机器标注语料存在的问题以及人工校对的一些规律，利用学习得到的规律来指导机器自动校对。我们取约 97 万字(大约 67 万词)的经过机器标注的语料，分给 10 个语言学研究生进行人工校对。通过人工校对语料与机器标注语料的比对统计学习，共找出 38957 次个修正模式，13023 个修正模式词型。

为每个校对者建立以下的数据结构：

```
struct{
PTN     aPatten;          //修正模式；
long    numOfCor;         //修正次数，即校对者将语料中原序列校对成修正序列的次数；
long    total;            //原机器标注语料中原序列出现的次数；
};
```

定义第 i 个校对者校对率：　　$crate_i=numOfCor_i/total_i$；

校对率反映了人工校对语料中原序列被校正的比率。表 1 列出两个修正模式及其相关的数据。

考虑到应该避免由于某个校对者及其相应语料可能存在的偏差，例如对于某个修正模式，校对者 1 中 total=2, numOfCor=2; crate=2/2=100%; 校对者 2 中，total=300,numOfCor=3;crate=3/300=1%;两者对于同一个修正模式的校对率截然不同，这里定义了总校对率：　$rate=\sum_{i=1}^{n} numOfCor_i / \sum_{i=1}^{n} total_i$，较好地解决了上述问题。

修正模式	不知/vt 道/vt 不/dz 知道/vt			人和/a 人/ng 和/cc		
修改人数	10			10		
修改人编号	numOfCor	total	crate	numOfCor	total	crate
1	3	4	0.75	1	7	0.14
2	3	3	1	1	1	1
3	8	8	1	1	4	0.25
4	5	5	1	2	6	0.33
5	3	4	0.75	2	4	0.5
6	9	9	1	3	4	0.75
7	4	6	0.67	1	3	0.33
8	5	5	1	6	6	1
9	1	1	1	2	2	1
10	8	8	1	1	1	1

表 1. 修正模式及相关数据

上例子中，rate=(2+3)/(2+300)=1.66%，而不是（1%+100%）/2=50.5%。

由于该语料是留学生语料，语料中大量存在语法、词汇等不规范甚至错误的现象，所以，

我们没有考虑修正模式与其前趋和后继的共现情况，而只考虑对修正模式本身的统计学习。这里我们没有讨论词性兼类问题，即 comp1 相同而 comp2 不同的修正模式不记入后续统计。

首先建立修正模式表。修正模式表是从修正模式中抽取的，用于对语料进行自动校对的修正模式集合，语料中具有的原序列都被修改成修正序列。

该表主要成员由对修正模式进行统计学习而得到。按修改人数 count(有 count 个校对者语料中统计出该修正模式)及总校对率 rate 指标计算出修正模式词型数，如表 2。

count / rate	≥1	≥2	≥3	≥4	≥5	≥6	≥7	≥8	≥9	=10
≥0.0	13023	2047	814	429	268	167	114	74	45	18
≥0.5	10503	1249	381	173	103	68	52	33	22	12
≥0.8	9129	878	216	74	42	25	21	16	11	7
≥0.95	8934	750	141	30	12	3	2	1	1	1

表 2. 参数 rate、count 对修正模式数的关系表

在参数选择的过程中，我们首先排除了 rate 较小(≤ 0.5)的情况，因为这些修正模式的可信程度太低。为了在高正确性的前提下尽可能多地校正错误，count 不宜过大，否则修正模式会很少。

第二部分，我们选取这样一类修正模式进入修正模式表：修正序列中有 ei∈PAIR, ei.word∈DICT;且在 DICT 中只有一个词性 pos，但 pos≠ei.pos。我们将这类修正模式改正后，送入修正模式表。这实际是将错误的词性修正过来。

例如：一个修正模式<[/bb,[/w>，我们知道[是 wh 词性，于是将修正模式改成<[/bb,[/wh>，又增加一个修正模式<[/w,[/wh>用于将原来改错的语料纠正过来。

第三部分，我们从多映射修正模式中，选取在词典中只有一个词性，但修正模式有多个词性的模式，将修正模式进行修改，使之与正确修正模式相一致。

例如在统计到的修正模式中，有<不可/dz 以/p, 不/dz 可以/vm >,<不可/dz 以/p, 不/dz 可以/p >两个多映射修正模式，但在 DICT 中“可以”只有一个词性 vm，于是将第一个修正模式送入修正模式表，而将第二个修正模式修改成<不/dz 可以/p, 不/dz 可以/vm >,从而将原来改错的语料纠正过来。

步骤三：基于规则的修改

在对修正模式的统计中，我们发现对数词和时间名词的修改比较混乱，有人从合，有人从分，倘若仍以统计的方法，往往会使结果难以体现出语料的一致性。为了解决以上问题，我们首先请语言学家根据语料的具体情况，并参考我国现有语料库的标注方案，提出专家级的修改规则，将这些规则应用于语料校对中，保证了语料的一致性。

三. 实验及结果

为了验证上述方法的有效性，我们设计了两个实验。第一个实验是封闭测试，将校对过

程应用于前面用于统计训练学习的语料，看能否提高手工校对后的语料质量；第二个实验进行开放测试，直接对另一批只经过机器标注的语料进行机器校对，看利用该方法直接进行语料自动校对的效果。

引入绝对改正数与绝对修正率[4]，用于对校对效果的评价。

设总修改次数为 zongcc，正确修改次数为 rightcc；定义：

绝对改正数 acn=正确修改次数数-错误修改次数=rightcc-(zongcc-rightcc)=2*rightcc-zongcc；

绝对改正率 acr=绝对改正数/总修改次数=2*rightcc/zongcc-1。

首先进行封闭性实验。分别取 count>=3,rate>=0.8 和 count>=2,rate>=0.8 两组参数进行修正模式表的选取，选取的修正模式表分别命名为修正模式表 1 和修正模式表 2。对经过 10 人校对的全部语料分别进行上述步骤的校对，结果如表 3，修改了 2582 处和 4220 处，分别占手工校对数量的 6.63% (2582/38957)和 10.83%(4220/38957)；经检查，分别有 14 处修改错误和 122 处，占整个修改次数的 0.5%（14/2582）和 1.94%（82/4220）。

	预处理	统计方法	规则方法	zongcc	改错数	rightcc	acn	acr(%)
修正模式表 1	342	1961	279	2582	14	2568	2554	98.9
修正模式表 2	342	3375	503	4220	122	4098	3976	94.2

表 3. 封闭测试结果

进行开放测试，选取若干篇机器标注后的文本，词数为 52409 个，经语言学专家仔细检查后确认有 841 处校对错误，因此机器校对之前错误率为 841 / 52409=1.6%。

仍然使用上面形成的两个模式修正表进行自动校对，结果如表 4，利用修正模式表 1，自动校对数量分别占手工校对数量的 56.00%(471/841)和 87.6%（737/841）；经检查，分别有 3 处和 34 处修改错误，占整个修改次数的 0.64%(3/471)和 4.61%(34/737)。

	预处理	统计方法	规则方法	zongcc	改错数	rightcc	acn	acr(%)
修正模式表 1	66	326	79	471	3	468	465	98.7
修正模式表 2	66	614	57	737	34	703	669	90.8

表 4.开放测试结果

利用修正模式表 2 进行机器校对之后，共修改 737 处。其中改正了 703 处错误（包括 52 处使得语料一致性增强的修正），增加了 34 处错误。因此，机器校对使得错误率降低为 (841－703＋34) / 52409=0.33%，使语料质量有了明显提高，错误率降低了 4.8 倍，一致性有了很大提高。

从两个实验的acr结果我们看到，封闭语料的acr值随着修正模式选择参数的放松而降低。我们应该更仔细地选择参数，保证将错误修正的同时，不将原本正确的改错。

四．结语及讨论

本文讨论了语料校对的质量评价准则：正确性、一致性和普遍认同性，并依此准则对经过机器标注和人工校对的语料进行机器校对。通过统计学习建立修正模式，通过合理选择修正模式表来实现语料的自动校对，经过预处理、基于统计和基于规则三个步骤来提高语料质

量。通过实验证明了该方法的有效性。

该系统尚待改进的地方主要有 1）只考虑了单类词的机器校对，没有考虑兼类词的机器校对；2）基于规则的校对方法中，规则的数量较少，质量有待提高。3）对于修正模式表条目的选择方法还有待进一步提高。应从修正人数、总校对率等参数入手，建立数学模型，找到参数的选择与校对效果的数学模型，用以指导修正模式表条目的选择，进一步提高校对水平。目前我们正在对兼类词建立统计模型，试图利用机器学习的方法自动抽取校对规则，以提高语料校对的质量。

参考文献

1. 黄昌宁, 李涓子. 语料库语言学[M]. 北京: 商务印书馆, 2002.4
2. 刘开瑛. 中文文本自动分词和标注[M], 北京: 商务印书馆, 2000
3. 史忠植. 知识发现[M]. 北京: 清华大学出版社, 2002.1
4. 钱揖丽, 张虎. 汉语分词及词性标注自动校验方法研究[C]. 北京：第一届学生计算语言学研讨会论文集. 2002.8