基于机器学习的分词不一致自动识别研究

卢俊之

（南京师范大学文学院，江苏 南京 210097）

摘 要： 分词不一致的处理是建设一个高质量的语料库所无法回避的问题，识别出分词不一致的不同成因是处理的前提和关键。本文提出了一种基于机器学习的分词不一致自动识别方案，通过两遍识别，以特征词法识别结果为基础，让机器从中学习到规则后辅以人工规则再处理第一遍未识别的不一致字串。我们对200万字语料库中的分词不一致字串进行了实验，封闭测试与开放测试的正确率分别达到85.22%和83.13%。

关键词： 分词；一致性；自动识别；机器学习

Automatic Identification of Inconsistency of Segment Based on Machine Learning

LU Jun-zhi

(School of Chinese Language and Literature, Nanjing Normal University, Nanjing, Jiangsu 210097)

Abstract: The treatment with inconsistency of segment is an inevitable problem in building a high-quality corpus. It is prerequisite and key to identify different causes with inconsistency of segment. This paper proposed an automatic identification scheme based on machine learning through identifying twice: using characteristic words law at first, secondly letting machine learn the rules from the result of first step and then using them cooperated with auxiliary rules to deal with first unidentified strings. We employed our scheme to process the inconsistency of segments in the corpus for 2 million Chinese characters. The precision rates are 85.22% and 83.13% respectively in close test and open test.

key words: Segment; Consistency; Automatic Identification; Machine Learning

1 引言

一个高质量、大规模的分词语料库是中文信息处理的根基。目前机器自动分词的正确率已达到97%左右。人工校对时，由于校对者依据的GB/T 13715-92《信息处理用现代汉语分词规范》对某些分词原则规定得比较含糊，加之校对者受语境干扰和自身语感的差异，时常会出现一个字串的意义、功能都是确定的但给出了不同的切分形式，我们称之为分词变异。由于意义、功能是否确定在分词层面机器难以判断，因此分词变异的外在表现和组合型歧义相同，同时，在一些专名和非专名串之间也存在着类似的问题。多种成因共同导致了复杂的分词不一致现象：一个相同的字串（不考虑它的意义、功能）在语料库中存在着不同的切分形式。

作者简介：卢俊之（1980-），男，江苏扬州人，硕士。 E-mail：lujunzhi@gmail.com

孙茂松（1999）[1]认为：衡量一个语料库质量的重要标准之一是分词后的语料库是否具有比较高的一致性。因此，建设一个高质量的语料库，分词不一致的处理是一个无法回避的问题，而处理的前提和关键就是识别出分词不一致的不同成因。

目前针对分词不一致问题的研究并不多，主要有孙茂松[1]总结了导致分词不一致的主要结构类型。杜永萍等（2001）[2]提出基于规则库的校对方法，通过人机交互完成一致性校对。刘江等（2005）[3]运用基于支持向量机的方法对分词不一致进行校对。苗玺等（2006）[4]从熟语料中人工归纳出判定和校对分词不一致的规则，并通过封闭性测试验证了规则的有效性。他们的研究成果表明，通过一系列的规则（以词性规则为主）处理分词不一致是可行的。

本文提出了一种基于机器学习的分词不一致自动识别方案，通过两遍识别，以特征词法识别结果为基础，让机器从中学习到规则后辅以人工规则再处理第一遍未识别的不一致字串。既解决了单一使用特征词法召回率过低和单一使用规则法容易忽视小规则的问题，也克服了原先方法前期需要投入大量人力进行人工校对和规则总结的缺陷。

我们从 1998 年 1 月《人民日报》200 万字的语料库中抽取到 40 926 个分词不一致字串，将字串相同的归为 1 组，共 1 797 组。从中抽样 10 131 条 473 组做人工识别，其中 5 065 条 230 组（样本 1）作为训练集用于观察和提取阈值，其余 5 066 条 243 组（样本 2）留作开放测试。

2 分词不一致现象分析

2.1 成因分析

分词不一致现象的成因主要归为以下 3 类，在识别时分别标注对应的字母：

（A）、分词变异。比如：

四/m　个/q　大字/n　“/w　福如东海/i　”/w

恭贺/v　新春/t　”/w　的/u　金色/n　大/a　字/n　分外/d　醒目/a　。/w

（B）、组合型歧义。比如：

多方/d　筹集/v　资金/n　6 亿/m　元/q　，/w

运/v　来/v　山基七/n　5 6 0 0/m　多/m　方/q　，/w

（C）、专名和非专名串。比如：

《/w　屯力/n　报/n　》/w　记者/n　张/nr　大和/nr

密度/n　之/u　大/a　和/c　持续/vn　时间/n　之/u　长/Ng　，/w

2.2 组合情况分析

每一组的情况并不一定是单纯的，会有 A、B、C3 类组合出现的可能，在识别时需要分别给予对应的标识。

常见的组合有一合对多分和一分对多合两种，前者比如：

等到/v　波罗的海/ns　三/m　国/n　加入/v　北约/j　(1)

因为/c　等/v　到/v　片子/n　剪/v　出来/v　，/w　(2)

事故/n　证明/n　及/c　医院/n　证明/n　等/u　到/v　保险/n　公司/n　索赔/v　，/w　(3)

其中，(1)和(2)的关系是 A 类，而(1)和(3)的关系是 B 类。

后者比如：

屋子/n　分/v　东/f　西/f　两/m　个/q　厢房/n　(4)

广场/n　东西/f　两侧/f　百/m　米/q　灯/n　廊/Ng　(5)

武侠小说/n　本身/r　是/v　娱乐性/n　的/u　东西/n　(6)

其中，(4)和(5)的关系是 A 类，而(4)和(6)的关系是 B 类。

3 算法设计

算法要点：

1、 抽取语料库中所有分词不一致的字串、词性标记和上下文等信息；

2、 前期处理全部 C 类；

3、 第 1 遍识别，使用特征词法标记部分 A 类，并让机器从中学习规则；

4、 第 2 遍识别，使用学习到的规则加上人工辅助规则标记剩下的 A 类和 B 类。

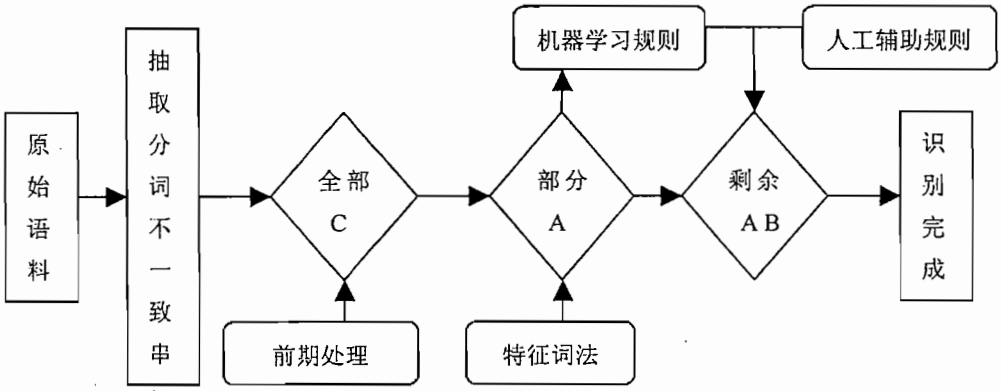

图 1 算法流程图

Fig.1 Flow chart of algorithm

3.1 分词不一致字串的抽取

表 1 中列出了每一个分词不一致字串抽取的信息，其中上下文的观察窗口设置为[-5,+5]。

表 1 数据库的各字段

Tab.1 Each field of database

字段名	说明	举例
Keystr	无分词信息的字串	大堤
Keywords	分词后的词串	大 堤
Keytags	词串的词性标记	/a/n
Wordnum	词串中词的数目	2
Leftwords	上文内容	挖/v 砂/n ，/w 对/p 长江/ns
Rightwords	下文内容	产生/v 了/u 严重/a 的/u 危害/vn

3.2 前期处理

前期主要将有明显特征的 C 类先行识别出来，减少后期处理的负担，避免造成干扰。

由 C 类引起的分词不一致可以从词性标记判定，专名包括人名（nr）、地名（ns）、机构团体名（nt）、其他专名（nz）四种。剔除专名后，如果该组不再存在不一致现象，则可以断定其不一致完全是由专名造成的，不对该组再做处理；如果剔除后不一致依然存在，则该组中剩下的不一致字串进入下一环节处理。

3.3 特征词法

前期处理后，关键将集中在识别 A 类和 B 类。我们首先使用特征词法标注一部分。

A 类分、合两种形式由于意义、功能是确定的，因此语境中可能会存在着相同的特征词，而 B 类这种可能性将远小于 A 类。

首先，一组中以 Keytags 的不同分类，一般情况分为分(wordnum>1)、合(wordnum=1)两类，但也有两类以上的，比如 keystr＝大地的组就有三类：/a；/a/u；/a/d。

然后，抽取每一类中所有不一致字串的上下文特征词构成一张左、右特征词表。特征词是由 leftwords 和 rightwords 去掉一些没有明显识别作用的词类（比如标点、连词）和词（比如“是”）组成。为避免对于高频特征词的过低估计，规定一个词出现几次就在特征词表中登记几次。

接着，将不同分类两两组合，计算它们之间特征词共现率，公式为：

$$R=\frac{\frac{M(table_{left1},table_{left2})}{N(table_{left1})}+\frac{M(table_{right1},table_{right2})}{N(table_{right1})}}{2} \quad (1)$$

其中，$table_{left1}$ 和 $table_{right1}$ 分别是组合中特征词数较多的表，M（table1,table2）函数对 table1 中的每个词考察其是否在 table2 中出现，如出现则计数加 1，N(table)函数计算 table 的特征词数。

最后，我们还需要从训练语料中提取阈值去判断组合内是否具有同一性。其中有 2 种情况：1、$wordnum_1 \neq wordnum_2$，则超过阈值可初步识别该组合为 A 类；2、$wordnum_1 = wordnum_2$，则超过阈值可将该组合合为一类处理。尽管两种情况结果不同但判断同一性这点是相同的，因此只需一个阈值就够了。我们使用出现频率较高的情况 1 作为阈值提取的依据。

表 2 中统计了不同阈值和 A 类的识别正确率、成功识别组合数之间的关系。

表 2 阈值与 A 类的识别正确率、成功识别组合数

Tab 2. Threshold value and precision rate/identification number of Class A

阈值	正确率（%）	成功识别组合数（个）
0.01	81.18	69
0.02	87.5	63
0.03	90.32	56
0.04	90.57	48
0.05	89.58	43
0.06	88.89	40

在第 1 遍识别时，我们希望得到较高的正确率。从表 2 中可以初步将阈值确定在 0.03－0.04 之间，经过进一步计算，当阈值为 0.032 时，正确率达到最高为 91.67%。对于情况 2 也使用 0.032 作为阈值。

在训练语料中统计，该轮共识别出 60 个组合为 A 类，召回率为 38.7%，其中正确 55 个，识别错误的 5 个词中，除 1 个是由于语料库本身的分词或词性标注错误造成误判外，其余 4 个均可使用 3.5 节中的人工规则 1 予以修正。

3.4 规则学习

使用特征词法的目的除了识别出部分 A 类以外，更重要的是我们希望将特征词法的识别结果提供给机器以学习到 A 类规则库。在 3.3 节中我们希望得到最理想的正确率，但对于机器学习来说，由于后期还有规则频率阈值的约束，就并不需要那么高的正确率而希望机器能多学习到一些规则备用。因此，我们将阈值从 0.032 降低到 0.02 作为规则提取的阈值，以放弃部分正确率的代价换来成功识别组合数的增加。

机器在对 40 926 个分词不一致字串进行第 1 遍识别后共学习到不同规则 186 条，舍弃低频规则，将出现频率 3 次以上的 44 条规则放入 A 类规则库。表 3 中列出了出现频率最高的 5 条规则。这些高频规则与孙茂松 [1] 所总结的导致分词不一致的主要结构类型是吻合的。

表 3 高频规则表

Tab.3 High frequency rules

	和 keytags	分 keytags	出现次数	可能的结构
1	/v	/v/v	84	动补结构
2	/v	/d/v	38	状中结构
3	/v	/v/n	38	动宾结构
4	/n	/n/n	27	定中结构
5	/n	/a/n	24	定中结构

3.5 人工辅助规则

为解决一些具有一定规律性和普遍性但规则复杂机器难以学习的情况，我们添加了 3 条人工辅助规则与 3.4 节的规则库配合使用。

人工规则 1：如果分、合两类情况中出现一类所有不一致字串的前邻接词与词性高度一致（对数词只要求词性一致，不考虑标点和助词等特征性不强的词类），而另一类中所有字串的前邻接词不含有该词，则将这种对类内具有高度相似性对类外具有严格排他性的组合认定为 B 类。在实际操作中，我们将高度一致的标准设置为：频率<10 的类相似度＝100%；频率≥10 的类相似度>90%，这样可以有效避免高频类中由于极个别字串分词错误造成的干扰。

人工规则 2：如果合时 keytags 为/i 或/l，或分时 keytags 前含/h 或后含/k，则认定该组合属于 A 类。孙茂松[1]、苗玺[4]均将成语、习用语、前后词缀作为构成 A 类的常见结构。这类词合时词性标记单一但分时由于本身结构复杂造成词性多样，机器难以学习到强势规则而造成漏检。

人工规则 3：对分时 wordnum 全部大于 2 的组标为 A 类。尽管这种多词不一致串在人工内省造句时仍然可能是 B 类，但数目极少且在真实语料中很难发现，我们可以直接识别为 A 类。

3.6 规则库识别

进行第 2 遍识别，将一组中按 Keytags 的不同分类，对 wordnum 不同的类两两组合。首先使用 3 条人工辅助规则对每个组合进行判断，如果符合则完成识别，否则如果该组合在第 1 遍识别时未能标识则从 A 类规则库中寻找匹配的规则，如果找到则标识为 A 类，否则标识为 B 类。

由于词性标注中标注为词（比如名词 n）和标注为语素（比如名语素 Ng）的分词单位除了能否成词这一区别外，它们具有较强的相似性，所以我们在使用规则库判定时，模糊了词和语素的区别，将它们等同看待，以扩大规则的适用范围。

4 实验结果与讨论

4.1 实验结果

我们使用此算法标注了抽取出的全部 40 926 个分词不一致字串，将样本 1 中 5 065 条 230 组人工标注结果与机器标注进行比对，一组中标识全部正确的算作正确，封闭测试正确率为 85.22%，对于 A 类，正确率 86.21%，召回率 88.65%；对样本 2 的 5 066 条 243 组进行同样的比对，开放测试正确率为 83.13%，对于 A 类，正确率 86.62%，召回率 86.08%。

4.2 错误分析

我们统计了样本 1、2 中所有的识别错误，原因主要分为 5 类：

1、机器学习的规则不够，无法识别出 A 类，比例为 42.7%。这主要是由于第 1 遍未被识别的一些组中的规则过于弱势而机器没有学习到或学习到但频率未超过阈值。

2、机器学习的规则失效，将 B 类误识为 A 类，比例为 38.7%。比如表 3 中列出的/a/n—/n 是识别 A 类的强势规则，但是下面这组却属于 B 类：

描绘/v　了/u　春/Tg　回/v　大地/n　，/w 万物/n 复苏/v ，/w

天/n　大/a　地/n　大/a　不/d　如/v　党/n　的/u

3、语料库本身分词或词性标注错误，比例为 8.0%。比如：

你们/r　要/v　嫌/v　麻烦/a　的/u 话/n　(7)

天寒地冻/i　，/w　大雪/t　封门/v ，　(8)

这些错误会在按 keytags 分类时将本是一类的字串分属两类反之将本不是一类的却合在一起，在第 1 遍识别时可能出现两类之间特征词共现率高估(比如 7)，而在第 2 遍识别时则可能出现规则不够（比如 8)。

4、特征词共现率高估，将 B 类误识为 A 类，比例为 6.7%。

主要出现在一些不一致字串数目较少的组，偶然出现的 1 个相同词就会造成共现率高估。

5、人工辅助规则失效，比例为 4.0%。

5 结语

如何提高识别的正确率是我们进一步研究的重点，我们觉得主要有这几方面的工作要做：(1)、深入观察语料，挖掘分词不一致现象不同成因之间的深层规律，拓展目前特征词加规则库的模式；(2)、让机器在自动学习规则的基础上能归纳、合并规则，同时增加规则的要素约束规则失效；(3)、提高算法的鲁棒性，能够抵抗语料库自身错误等干扰因素带来的影响。

参考文献：

[1] 孙茂松. 谈谈汉语分词语料库的一致性问题[J]. 语言文字应用，1999，02：88-91.

[2] 杜永萍，郑家恒. 分词及词性标注一致性校对系统的设计与实现[J]. 电脑开发与应用，2001，10：16-18.

[3] 刘江，郑家恒，张虎. 中文文本语料库分词一致性检验技术的初探[J]. 计算机应用研究，2005，09：52-54.

[4] 苗玺，郑家恒. 中文语料库分词不一致的分类处理研究[J]. 山西大学学报，2006，01：22-25.