基于词典和语素的交集型歧义消除模型

李春雨[1]，王 勇[2]
(1. 浙江机电职业技术学院，浙江 杭州 310053；
2. 杭州茂亨自控仪表有限公司，浙江 杭州 310053)

摘 要：提出了一种消除中文分词中交集型歧义的模型。首先通过正向最大匹配法和逆向最大匹配法对中文文本信息进行分词，然后使用不单独成词语素表对分词结果进行分析对比消歧，得到符合汉语语境的结果。整个过程分为歧义识别、歧义分析、歧义消除三个阶段。实验结果表明，该模型可以有效降低由交集型歧义引起的中文文本切分错误率。

关键词：自然语言处理；分词；交集型歧义
中图分类号：TP391.3 **文献标识码**：A **文章编号**：1674-7720(2013)04-0012-03

Elimination of Chinese overlapping ambiguity based on dictionary and morpheme

Li Chunyu[1], Wang Yong[2]
(1. Zhejiang Institute of Mechanical & Electrical Engineering, Hangzhou 310053, China;
2. Maoheng Auto Control Meter Co., Ltd., Hangzhou 310053, China)

Abstract: This paper puts forward an overlapping ambiguity elimination model of Chinese segmentation. Firstly, the model segments the Chinese text corpus by maximum matching method and reverse maximum matching method, and then analyzes the segmentation to eliminate the ambiguity based on non-single morpheme list to obtain correct result which fits the Chinese context. The totally processing includes three sections as following: ambiguity detection, ambiguity analysis and ambiguity elimination. The experiment data indicates that the model referred in the paper reduces Chinese segmentation mistakes caused by overlapping ambiguity effectively.

Key words: nature language processing; word segmentation; overlapping ambiguity

在英文和其他西方语言系统中，文本书写时通常是词与词之间用空格隔开，但中文的书写形式却是连续的字串，词与词之间没有任何标志。而对于中文来说，如果不进行词语的有意义切分，句子将没有任何的意义[1]。分词是中文信息处理的第一步，就目前来说，较为常用的中文分词方法主要分为两类：基于规则的方法和基于统计的方法[2]。基于规则的分词方法的核心在于建立一个完备的词典，然后通过该词典对句子中的切分片段进行匹配，以完成分词过程。较常用的基于词典的中文分词方法有正向最大匹配法、逆向最大匹配法和最佳匹配法；基于统计的分词法的基本原理是对语料库中相邻字的组合频度进行统计，根据一定的频度计算公式来决定字符串成为词的可能性进行分词。字词共现的频度高低体现了汉字之间结合关系的紧密程度。当紧密程度高于某一个阀值时，便可认为此字符串可能已经构成了一个词[3-5]。这些方法有效地促进了中文分词研究的进一步发展，但在实际应用中仍然有很多因素影响着分词的准确率，其中较常见的就是分词的歧义问题。

本文建立了一个中文分词的模型来减少中文分词中的歧义问题，以提高分词的准确率。该模型基于正向最大匹配法和逆向最大匹配法来完成分词过程，通过对两种分词方法产生的分词序列进行比较分析，最终通过基于罚分机制的歧义消除算法选出正确的序列来完成分词。

1 最大匹配法与交集型歧义

最大匹配法有正向最大匹配法 MM 法(Maximum Matching Method) 和逆向最大匹配法 RMM 法(Reverse Maximum Matching Method)两种基本方法。它们具有原理简单、时间复杂度低、易于实现等优点，但是不足之处在于往往不能识别出切分歧义而导致文本切分错误[6]。而中文语言环境中歧义的存在是一个很普遍的现象，据统计，MM 法对于文本的错误切分率为 1/169，RMM 法对于文本的错误切分率为 1/245[7]。

导致分词错误的切分歧义主要有组合型歧义和交

集型歧义两种。在所有的歧义现象中，普通的交集型歧义现象所占的比例为85%以上[8]，所以交集型歧义在中文文本中是极为常见的。以文本“他的确切地址在这儿”为例，通过MM法进行切分的结果为“他/的确/切/地址/在/这儿”，用RMM法得到的结果为“他/的/确切/地址/在/这儿”，可见两种方法得到了不一样的分词结果，而有差别的“的确切”部分存在的歧义就是交集型歧义。

2 交集型歧义消除模型

2.1 歧义分词

歧义消除的过程通常是与分词结合在一起的，对于中文文本来说，如果存在歧义，分别通过MM法和RMM法所得到分词结果是一样的，反之则不一样。对于存在交集型歧义的文本，交集型歧义消除模型首先需要将文本用MM法和RMM法分别进行切分以得到两个不同的切分结果。除此之外还可以通过其他的分词方法得到更多的切分结果，但实验证明MM法和RMM法的结合分词能够识别出绝大多数的交集型歧义，基于此点以及效率上的考虑，本文的模型中只保留使用MM法和RMM法两种切分方法来进行对比分析。

以文本“他明白天为什么下雨”为例，可以通过MM法和RMM法分别得到结果(1)和结果(2)：

结果(1)：他/明白/天/为什么/下雨

结果(2)：他/明/白天/为什么/下雨

2.2 不单独成词语素表

在本文所研究的交集型歧义消除模型中还需要用到一个不单独成词语素表。该表包含了一些在中文语境中单独出现通常没有意义的一些字，比如“第”，当“第”单独出现时基本上没有任何意义，但是“第”通过与其他字的组合却能具有很多不同的意义，例如“及第”，“第一”等。在交集型歧义消除模型中，不单独成词语素表所包含的不单独成词的语素完备性对分词的模型在实际应用当中的文本切分准确性是紧密联系在一起的，语素表完备性越高则文本切分越准确，反之则越不准确。

2.3 消歧算法

交集型歧义消除模型中所使用的用来确保能够消除歧义的算法主要原理是通过引入针对切分结果赋予权值，然后对权值进行统计的方法来进行歧义消除的。

定义：ABC为文本，A、B、C均为切分单元，即ABC可被切为A/B/C，A、B、C分别被赋予初始权值R(A)=R(B)=R(C)=1。

现假设切分结果“A/B/C”中只有切分单元B属于2.2节所构建的不单独成词语素表，则切分单元B的权值会增加，即R(B)=2。

然后对切分结果“A/B/C”的权值进行统计，R(A)+R(B)+R(C)=1+2+1=4，通过不同的方法可以得到不同的切分结果，不同的切分结果的权值统计也会有区别。交集型歧义消除模型会将各个结果的权值统计进行比较分析，选出统计值较小的一个为消除歧义后的切分结果。

对于文中2.1节分别通过MM法和RMM法获得的结果(1)和结果(2)，分别对切分单元赋予初始权值：

结果(1)：R(他)=R(明白)=R(天)=R(为什么)=R(下雨)=1；

结果(2)：R(他)=R(明)=R(白天)=R(为什么)=R(下雨)=1；

通过将结果(1)和结果(2)与不单独成词语素表进行匹配，可以判断结果(2)中的“明”字属于不单独成词语素，即R(明)=2，通过结果权值统计：

结果(1)：R(他)+R(明白)+R(天)+R(为什么)+R(下雨)=1+1+1+1+1=5；

结果(2)：R(他)+R(明)+R(白天)+R(为什么)+R(下雨)=1+2+1+1+1=6；

然后通过对结果进行比较，交集型歧义消除模型选取权值统计较小的结果(1)为消歧后的正确结果，同时该结果也完全符合中文语境下的正确的表达意义。

2.4 模型示意图

通过以上的分析描述，交集型歧义消除模型消歧的过程主要分为三个步骤：发现歧义、分析歧义、消除歧义。发现歧义是通过MM法和RMM法对文本进行切分对比来识别歧义的存在；分析歧义的过程是以不单独成词语素表为基础，通过对文本切分单元进行权值赋予与统计来完成的；最后的消除歧义步骤则是对分析歧义的结果进行对比，剔除切分错误文本来消除歧义。图1是交集型歧义消除的示意图。

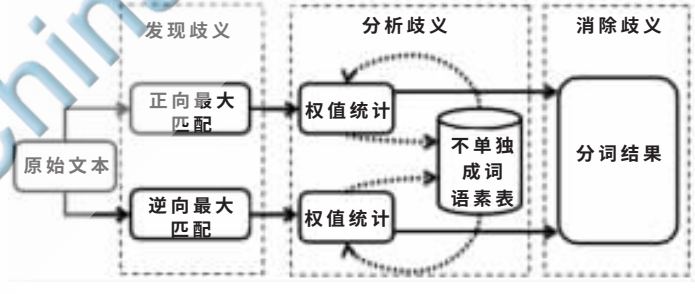

图1 交集型歧义消除示意图

3 实验及结果

对于中文分词来说，分词的高效性和准确性是极其重要的。在相同的条件下，更准确、更高效的分词方法就意味着更好的分词性能以及更快的分词速度。

(1) 效率分析

根据本文中的分词策略，对于一个中文句子来说，分别用正向最大匹配法和逆向最大匹配法得到两个分词结果序列，然后通过不单独成词语素表来对两个结果序列进行分析，整个分析过程不涉及到其他的分词方法。而正向最大匹配法和逆向最大匹配法基于其实现原理分词效果是非常理想的，在所有的中文分词系统中基本上都可以找到这两种方法的身影，所以本文中的分词

过程基于正向最大匹配法和逆向最大匹配法这两种基本方法，然后再结合不单独成词语素表，使分词的效率得到了保证。

(2) 准确性分析

在中文自然语言处理领域，正向最大匹配法和逆向最大匹配法是两个最基本的分词方法，不幸的是这两种方法都不能很好地解决中文语言环境中的分词歧义问题。因此，针对于这一系列因素，本文中提到的交集型歧义消除模型利用对切分结果进行基于不单独成词语素表的权值统计来选出相对权值较小的切分结果，进而保证中文分词中的交集型歧义的发现与消除。

(3) 实验结果分析

基于以上的规则，本文中开发了一个交集型歧义消除系统，其中不单独成词语素表包含了 4 871 个不单独成词语素，同时从 2012 年的人民日报中选取了 6 篇文章作为实验的原始语料库。通过用交集型歧义消除模型获得的消歧结果与单独使用正向最大匹配法和逆向最大匹配法所得到的结果进行对比来分析系统的效率和准确度。

表 1 和表 2 分别为单独使用 MM 法和 RMM 法进行文本切分时的切分准确率。表 3 为采用交集型歧义消除模型进行切分的准确率，从中可以看到交集型歧义消除模型针对于同一语料库的文本切分准确率最高。

表 1 单独使用 MM 法进行切分的准确率

	1	2	3	4	5	6
正确切分数	105	118	128	130	151	163
总字数	120	132	140	156	164	190
准确率	0.875	0.894	0.914	0.833	0.921	0.858

表 2 单独使用 RMM 法进行切分的准确率

	1	2	3	4	5	6
正确切分数	110	121	128	144	152	174
总字数	120	132	140	156	164	190
准确率	0.917	0.924	0.914	0.923	0.927	0.916

表 3 交集型歧义消除模型进行切分的准确率

	1	2	3	4	5	6
正确切分数	118	127	139	149	160	177
总字数	120	132	140	156	164	190
准确率	0.983	0.962	0.993	0.955	0.976	0.932

图 2 为 MM 法、RMM 法和交集型歧义消除模型切分准确率的对比。从图 2 中可以看出，交集型歧义消除模型对文本切分中的交集型歧义消除准确率比单独使用正向最大匹配法和逆向最大匹配法的切分准确率要高。

图 3 是交集型歧义消除模型与 MM 法、RMM 法在文本切分效率上的对比。从图 3 中可以看出，交集型歧义消除模型虽然较 MM 法和 RMM 法额外使用了不单独成词语素表，但在效率上并没有明显的降低。

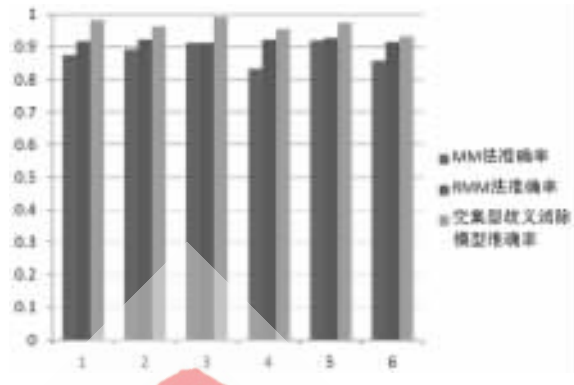

图 2 MM 法、RMM 法和交集型歧义消除模型切分准确率的对比

图 3 交集型歧义消除模型、MM 法和 RMM 法文本切分效率对比

通过以上的实验可以看出，交集型歧义消除模型可以很好地发现并解决中文语言环境下的交集型歧义问题，并且具有较高的效率和准确率。根据实验数据可知，本系统的分词结果准确率比单纯使用正向最大匹配法和逆向最大匹配法高得多；另一方面，由于使用了不单独成词语素表，本文算法的分词效率较原始的正向最大匹配法和逆向最大匹配法有略微的降低。但结合效率和准确性来进行整体分析，可以看到交集型歧义消除模型对于解决中文分词中的交集型歧义是非常有价值的。

本文基于不单独成词语素表及常用的分词方法提供了一个中文分词中的交集型歧义的解决方案。实验结果表明，交集型歧义消除模型能够很好地解决中文分词中的交集型歧义问题，希望本文的研究成果能够对中文分词歧义消除领域的发展起到一定的推动作用。

参考文献

[1] 孙茂松，邹嘉彦. 汉语自动分词研究评述[J]. 当代语言学，2001(1):22-32.

[2] 麦范金，王挺.基于双向最大匹配和 HMM 的分词消歧模型[J].现代图书情报技术，2008(8):37-41.

[3] 施彤年，卢忠良，荣融，等.多类多标签汉语文本自动分类的研究[J]. 情报学报，2003,22(3):306-309.

[4] 邹海山，吴勇，吴月珠，等.中文搜索引擎中的中文信息处理技术[J]. 计算机应用研究，2000(12).

[5] 赵伟,戴新宇,尹存燕,等.一种规则与统计相结合的汉语分词方法[J]. 计算机应用研究, 2004(3):23-25.
[6] 刘颖.计算语言学[M].北京:清华大学出版社,2002.
[7] 梁南元.书面汉语自动分词系统——CDWS[J]. 中文信息学报,1987(2):44-52.
[8] 一种 Hash 高速分词算法[J].解放军理工大学学报(自然科学版),2004,5(2):40-42.

(收稿日期:2012-12-26)

作者简介:

李春雨,女,1976 年生,硕士研究生,主要研究方向:通信系统开发

王勇,男,1986 年生,硕士研究生,主要研究方向:计算机软件。