文章编号:1001 - 9081(2008)07 - 1686 - 03

基于规则挖掘和 Naïve Bayes 方法的组合型歧义字段切分

张严虎,潘璐璐,彭子平,张靖波,于中华
(四川大学 计算机学院, 成都 610064)
(yuzhonghua@cs.scu.edu.cn)

摘 要:组合型歧义字段切分是中文自动分词的难点之一。在对现有方法进行深入分析的基础上,提出了一种新的切分算法。该算法自动从训练语料中挖掘词语搭配规则和语法规则,基于这些规则和 Naïve Bayes 模型综合决策进行组合型歧义字段切分。充分的实验表明,相对于文献中的研究结果,该算法对组合型歧义字段切分的准确率提高了大约 8%。

关键词:中文分词;组合型歧义;词语搭配规则;语法规则
中图分类号: TP18 **文献标志码**:A

Resolving combinational ambiguity in Chinese word segmentation based on rule mining and Naïve Bayes method

ZHANG Yan-hu, PAN Lu-lu, PENG Zi-ping, ZHANG Jing-bo, YU Zhong-hua
(*College of Computer Science, Sichuan University, Chengdu Sichuan* 610064, *China*)

Abstract: Combinational ambiguity is one of the most difficult problems for Chinese word segmentation. After in-depth analysis of the other algorithms in literature, the paper proposed a new segmentation algorithm. The algorithm automatically mined word collocation rules and grammar rules from training corpus, and then made integrated decisions to resolve combinational ambiguity based on the mined rules and Naïve Bayes method. Extensive experiments show that the proposed algorithm obtains an accuracy increase of 8% against the related works.

Key words: chinese word segmentation; combinational ambiguity; word collocation rules; grammar rules

0 引言

中文自动分词是中文信息处理的基础工作,后续深层次的分析处理,如机器翻译(MT)、语音合成、自动分类、自动摘要、自动校对等,都直接依赖自动分词的结果。自动分词对中文搜索引擎也有巨大影响,分词的正确与否,常常影响搜索结果的相关度排序。

对于中文自动分词,尽管研究者们已经做了大量工作,提出了一系列算法[1,2,5],也设计了一些原型系统,但这些算法和系统还无法满足中文信息处理的实际需要。造成这种困难的主要原因是中文分词的两大难题——歧义切分和新词识别。主要的切分歧义可以归结为两类[4]:交集型歧义和组合型歧义。

目前,汉语分词歧义的研究大多集中在交集型歧义的消解上,根据相关上下文信息予以解决,而对组合型歧义的研究相对较少。

文献[5]主要依靠人工总结出来的规则对组合型歧义进行消解,尽管达到了比较高的准确率,但这种方法只能处理有限的语言现象,不具有一般性。文献[2]将组合型歧义切分问题看成是词义消歧(Word Sense Disambiguation, WSD)问题,基本思想是根据歧义字段语境中共现词语的分布特征来进行消歧,利用向量空间模型,通过距离计算,对组合型切分歧义进行消解。这种解决方法具有一般性,但其结果仍不尽如人意。文献[1]在从分样本集和从合样本集中分别统计前语境和后语境中出现的词频,形成四个语境词频表,基于这些词频进行切分歧义消解。然而,由于中文词组数以几十万计,上述方法需要极大的训练语料,以便涵盖尽可能多的词组,否则在测试中会遇到大量训练语料中不存在的词组,严重影响准确率。

本文在对现有方法进行深入分析的基础上,提出了一种基于规则挖掘和 Naïve Bayes 方法的组合型歧义字段切分算法,以进一步提高消歧的准确率。

1 组合型歧义字段切分模型设计

研究者将组合型歧义定义为:

定义 1 给定任意汉字字段 AB,如果 A、B 均为词表中的词,且切分 A 及 B 在真实语境中均能实现,则称 AB 为组合型歧义切分字段(简称组合型歧义)。

组合型歧义的实例如下:

1)我必须要学会用头脑打球

2)中国少数民族戏剧学会颁发的荣誉

在实例 1 中"学会"应分,实例 2 中"学会"应合。由上面的例子可以发现,对于组合型歧义,词汇的使用并不是孤立的。一方面,它会和别的词构成习惯性和典型性的搭配关系;

收稿日期:2008 - 01 - 14;**修回日期**:2008 - 03 - 19。 **基金项目**:国家自然科学基金资助项目(60473071);高等学校博士学科点专项科研基金 SRFDP 资助项目(20020610007 号);四川大学计算机学院青年基金资助项目。

作者简介:张严虎(1982 -),男,浙江温州人,硕士研究生,主要研究方向:数据挖掘、自然语言处理; 潘璐璐(1982 -),女,重庆人,硕士研究生,主要研究方向:数据挖掘、自然语言处理; 彭子平(1981 -),男,重庆人,硕士研究生,主要研究方向:数据挖掘、自然语言处理; 张靖波(1984 -),男,内蒙古赤峰人,硕士研究生,主要研究方向:数据挖掘、自然语言处理; 于中华(1967 -),男,黑龙江齐齐哈尔人,副教授,博士,主要研究方向:数据挖掘、自然语言处理。

另一方面,其前后语境的其他信息也可能预示了该组合歧义的切分与否。对于一些典型的搭配关系,绝大多数时候只表现为一种方式,或者合,或者分。

总之,局部上下文信息蕴含了组合歧义该分或合的信息,因此,如何深入挖掘该信息,挖掘的信息如何处理成了切分的关键。

本文从词的搭配和词性的搭配这两方面着手,进行上下文信息挖掘,提取词的搭配规则和词性搭配规则,并结合 Naïve Bayes 方法建立歧义字段切分模型。

为方便对某词以及词性的出现位置作标记,本文规定,以组合型歧义字段的出现位置为原点,左边第一个词表示为 Word - 1,词性(part of speech)表示为 POS - 1,右边第一个词表示为 Word + 1,词性表示为 POS + 1,以此类推,左边第 n 个词为 Word - n,词性为 POS - n,右边第 n 个词为 Word + n,词性为 POS + n。

1.1　词的搭配规则

特定词语出现在歧义字段的局部前后文时,由于与歧义字段中不同切分词语构成搭配的强度不同,因此对歧义的消解具有预示作用。例如,"要学会"该分,"要 - 1"与"学会"构成搭配,"学会了"该分,"学会"与"了 + 1"构成搭配,"的才能"该合,"的 - 1"与"才能"构成搭配。当出现以上搭配时,几乎总是表现为一种切分方式,因此,可以认为它们是一种典型的搭配关系。这里,前后文中词出现的位置不同,可能对组合型歧义字段的切分有不同的影响。例如,"工程 学会 要 完成 这个 任务"中,"要"的位置变到 + 1, "学会"变成合。因此,词的位置信息对于切分结果非常重要。

为了对上述论断进行验证,设置窗口大小为 2,考查了 - 1和 + 1 两个位置的词和歧义字段的搭配情况,如表 1 所示。

表 1　"学会"训练语料中部分词的搭配情况

词及其位置	出现次数	合	分	从合率/%	从分率/%	强势切分准确率/%
要 - 1	88	0	88	0.00	100.00	100.00
必须 - 1	13	0	13	0.00	100.00	100.00
了 + 1	100	0	100	0.00	100.00	100.00
用 + 1	18	0	18	0.00	100.00	100.00
教育 - 1	9	9	0	100.00	0.00	100.00
会员 + 1	6	6	0	100.00	0.00	100.00

通过表 1 可以发现,存在一些词的搭配规则,对歧义字段的切分具有很大的影响。在本文的词语搭配规则挖掘算法中,设定了阈值 α, 若搭配的强势切分正确率大于等于 α, 则输出该搭配作为切分的词语规则。此外,设定搭配出现次数阈值 β, 若出现次数大于等于 β, 则认为该规则可信,否则认为该规则偶然性较大,抛弃该规则。本文实验中,这些参数设置为 $\alpha = 0.95, \beta = 5$,表 1 中所有词的搭配都符合阈值要求。

1.2　词性的搭配

除了上述词的搭配关系,还存在词性的搭配,下面分为两种情况讨论。

第一种情况:只与前后文中相邻的一个词构成搭配

"小孩/n 学会 走路/v","机械/n 学会 颁发/v","计算机/n 学会 发表/v"

在上述例句中,"学会"分别与"/n - 1"、"/v + 1"搭配。同样,词性位置信息对于切分结果来说也是非常重要的。

第二种情况:与前后文中的多个词构成搭配

"美国/n 心理/n 学会 最近/t 发表/v 文章/n"

"走访/v 了/u 上海市/n 化学/n 化工/n 学会"

上述例子中包含(/n - 2,/n - 1)搭配,"学会"为合。为了验证这不是偶然现象,从搜狗语料库[8]中统计了包含"学会"的部分例句,共有 776 句,含(/n - 2,/n - 1)搭配的有 87 句,"学会"为合的 78 次,占 89.66%,为分的 9 次,占 10.34%。可见这不是偶然现象,这种词性的搭配关系也可以作为切分的依据。

为了对上述规律进行验证,设定窗口大小为 4,考查了三种形式的搭配,分别为(- 2, - 1),(- 1, + 1),(+ 1, + 2)。通过对"学会"训练语料的分析,提取出部分搭配,列于表 2 中。

表 2　"学会"训练预料中部分词性搭配情况

词性的搭配	出现次数	分	合	从分率/%	从合率/%	强势切分准确率/%
(/v + 1, NULL + 2)	121	117	4	96.69	3.31	96.69
(/d - 1,/v + 1)	37	37	0	100.00	0.00	100.00
(/v - 1,/v + 1)	87	83	4	95.40	4.60	95.40
(/v - 2,/n - 1)	39	28	11	71.79	28.21	71.79
(/n - 1,/v + 1)	47	25	22	53.19	46.71	53.19

注:NULL - n,NULL + n 表示不存在"学会"前面/后面的第 n 位词的词性。

由表 2 可知,确实存在一些词性搭配对歧义字段的切分具有较大的影响。在本文的词性搭配规则挖掘算法中,设定了阈值 α, 若词性搭配的强势切分正确率大于 α, 则输出该搭配作为歧义切分的语法规则。此外,设定搭配出现次数阈值 β, 若出现次数大于等于 β, 则认为该规则可信,否则认为该规则偶然性较大,抛弃该规则。本文实验中,这些参数设置为 $\alpha = 0.85, \beta = 5$。基于这些参数,可以从表 2 中提取出如下语法规则:

(/v + 1,/NULL + 2)分

(/d - 1,/v + 1)分

(/v - 1,/v + 1) 分

1.3　Naïve Bayes 方法

前文的词搭配规则和语法搭配规则并不能涵盖所有的组合型歧义字段切分问题。例如,表 2 中的(/n - 1,/v + 1),若采取强势切分策略,准确率只有 53.19%,对于总体准确率有很大影响,应该将其抛弃。为了解决词语搭配规则和语法搭配规则覆盖面窄的缺陷,本文提出基于 Naïve Bayes 的方法。

基于 Bayes 定理的分类规则为[3]:

$$C = \underset{c_j}{\operatorname{argmax}} P(C_j \mid w_1, \cdots, w_n) = \underset{c_j}{\operatorname{argmax}} P(C_j) \prod_{i=1}^{n} P(w_i \mid C_j)$$

由于组合型歧义字段只有分与合两类,因此可以归结为两类分类问题。设定窗口大小为 2,取 POS - 1,POS + 1 两个点(- 1 及 + 1 两个位置词的词性)作为分类特征,则

$$C = \operatorname{argmax} P(C_i) \times P(POS-1 \mid C_i) \times P(POS+1 \mid C_i)$$

转化为求 $Value_{分}$, $Value_{合}$:

$$Value_{分} = P(C_{分}) \times P(POS-1 \mid C_{分}) \times P(POS+1 \mid C_{分})$$

$$Value_{合} = P(C_{合}) \times P(POS-1 \mid C_{合}) \times P(POS+1 \mid C_{合})$$

分别计算 $Value_{分}$ 和 $Value_{合}$, 分类规则为:

1)若 $Value_{分} >= Value_{合}$,则分。

2)若 $Value_{分} < Value_{合}$,则合。

举例说明:

1)人们/n 需/v 学会/n 的/u 四/m 种/q 气质/n

2)营养/n 学会/n 主席/n 在/p 历经/v 实际/a 调查/v 后/f

表3 “学会”训练样本中部分词性出现频率表

词性	出现次数	分	合	$C_{分}$个数	$C_{合}$个数
/v-1	155	129	26	533	143
/u+1	97	86	11	533	143
/n-1	141	49	92	533	143
/n+1	63	15	48	533	143

通过对“学会”训练样本的学习,得到表3,可以计算各个词性对应的 $Value_{分}$ 和 $Value_{合}$ 如下:

1) $Value_{分} = 533/(533+143) \times 129/533 \times 86/533 = 0.030790$;

$Value_{合} = 143/(533+143) \times 26/143 \times 11/143 = 0.002959$;

$Value_{分} > Value_{合}$, Naïve Bayes 模型判断为分,消歧正确。

2) $Value_{分} = 533/(533+143) \times 49/533 \times 15/533 = 0.002040$;

$Value_{合} = 533/(533+143) \times 92/143 \times 48/143 = 0.045682$;

$Value_{分} < Value_{合}$, Naïve Bayes 模型判断为合,消歧正确。

2 算法步骤

2.1 考查对象

前文对算法中的一些关键点进行了详细的分析和介绍,下面给出算法的具体流程。本文将选取文献[2]中的高频组合型歧义字段“学会”、“才能”、“个人”、“正当“和“上来”作为实验对象。根据[2]的实验结果,在20个典型且常用的组合型歧义字段中,切分正确率最差的分别是“学会”(86.65%)、“个人”(93.16%)、“上来”(84.14%)、“才能”(92.94%)、“正当”(92.44%),这也是本文选择它们作为实验对象的原因。

2.2 主要流程

从一个规模为53.9 MB的搜狗语料库(http://www.sogou.com/labs/dl/t.html)中抽取含有上述组合型歧义字段的例句,一部分作为训练语料,另一部分作为测试语料,对这些例句采用中科院分词系统ICTCLAS进行切分。

对歧义字段,ICTCLAS可能会产生两种类型的切分:

1)歧义字段不切分为单独的词(少数)

例如:

①中华/n 预防/v 医学会/n 妇女/n 保健/b 分会/n

②江苏省/n 昆虫学/n 会/v 的/u 副/b 理事长/n

2)歧义字段被切分为单独的词(多数)

例如:

③我们/r 要/v 学会/v 一些/m 技巧/n

只抽取属于第二种切分类型的例句构成训练语料,对这些语料中组合型歧义字段的切分错误手工进行了修正。测试语料包含所有两种切分类型。为方便与文献[1]进行实验对比,将属于第一类切分的测试例句称为第一类测试样本,属于第二类切分的测试例句称为第二类测试样本。

主要流程如下:

1)使用中科院分词系统ICTCLAS对语料进行分词和词性标注,划分一部分作为训练语料,另一部分作为测试语料。

2)统计训练语料中词及词性的搭配情况,抽取词的搭配规则和语法规则。

3)逐句读入测试样本,首先判断是否属于第一类测试样本,若是,则输出ICTCLAS的切分结果。

4)如果第3步不能消歧,则使用词的搭配规则进行消歧。若存在两个或两个以上词语搭配规则获得满足,则选择在训练语料上出错最少的作为消歧依据。

5)若没有词语搭配规则获得满足,则判断是否存在语法规则获得满足,若是,则基于语法规则进行消歧。若存在两个或两个以上语法规则获得满足,则选择在训练语料上出错最少的作为消歧依据。

6)如果第3、4和5步均不能消歧,则基于Naïve Bayes方法消歧。

3 实验结果

为了验证算法的消歧效果,对测试样本中的组合型歧义字段进行了手工切分,以这些手工切分结果作为衡量算法输出正确与否的标准。表4给出了本文算法在两类测试样本上的综合准确确率的消歧实验结果。

文献[2]的测试样本同样中包含了第一类和第二类测试样本。本文算法的综合准确率(包括第一、二类测试样本)及与文献[2]实验结果对比见表4。准确率提升最大的是“上来”和“学会”,分别提升了12.34%和10.24%,其余几个歧义字段也有不同程度的提高。五个歧义字段平均切分正确率提高了8.07%,提升效果相当显著。

表4 本文算法与文献[2]实验结果对比

歧义字段	训练样本数	测试样本数	本文算法切分正确/错误	本文算法准确率/%	文献[2]算法准确率/%	与文献[2]算法相比提升/%
学会	676	321	311/10	96.89	86.65	10.24
个人	447	858	848/10	98.83	93.16	5.67
上来	292	227	219/8	96.48	84.14	12.34
才能	589	827	823/4	99.52	92.94	6.58
正当	223	150	147/3	98.00	92.44	5.56
平均					89.87	8.07

文献[1]对第二类测试样本进行了实验,选取的歧义字段只包含本文中的“学会”、“才能”和“上来”,达到的准确率分别为96.51%,96.83%和91.21%。而本文在第二类测试样本的实验结果:“学会”97.72%,“才能”99.62%,“上来”95.51%。比较这三个歧义字段,本文算法的准确率分别提高了1.21%,2.79%和4.3%。

4 实验分析

4.1 三种消歧方法的作用

本文基于三种策略综合进行歧义消解,即词语搭配规则、语法规则和Naïve Bayes方法,这三种策略对于正确切分起到了不同程度的作用。

(下转第1704页)

实验中,使用延迟来模拟每个Web服务具有的不同执行成本。WSMSME系统和Web服务运行在不同的机器上。另外每个Web服务组合由一系列的顺序的Web服务组成(其他的Web服务组合模式暂时不考虑)。

3.2 实验与结果分析

在3.1节原型系统中还实现了随机算法(RANDOM),用它和本文中动态规划算法(表示为OPTIMIZER)进行比较。随机算法采用的是将Web服务组合随机分成 个子序列进行执行的策略。

我们开发了19个Web服务,每个服务的执行成本在30到160之间,并从1到15动态增加引擎数。实验结果如图3所示,随着引擎数的增加,两种算法都能够使最大负载不断减小,但OPTIMIZER算法的性能始终比随机算法要好。

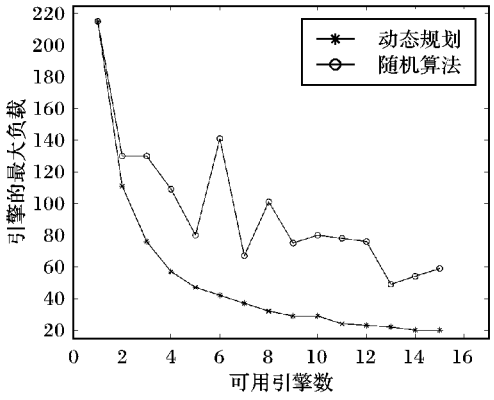

图3 引擎的最大负载

4 结语

基于多引擎的Web服务管理系统,弥补了单引擎架构的缺点,同时解决了WSMSME的调度执行优化问题。在Web服务分派和执行前,将其分成流水执行的子序列来最小化引擎的最大负载。然而这只是对WSMSME系统最基本的优化,并且实验的设计趋于理想化,下一步的工作应将算法扩展到其他更复杂的Web服务组合模式,并考虑使用更复杂的机制来表述Web服务的执行成本,尝试在Web服务组合语言如BPEL4WS方面做更多的努力。

参考文献:

[1] 马晓轩,林学练. Web服务性能优化的研究[J]. 计算机工程与应用,2005,41(8):19-20.

[2] MCILRAITH S A, MARTIN D L. Bringing semantics to Web services [J]. IEEE Intelligent Systems, 2003, 18(1): 90-93.

[3] Lü WEI-FENG, YU JIAN-JUN. pService: Peer to Peer based Web services discovery and matching [C]// Proceedings of 2nd International Conference on System and NetWorks Communications (ICSNC 2007). Washington, DC. IEEE Press, 2007: 54-54.

[4] SHU GAO, RANA O F, NICK J, *et al*. Ontology-based semantic matchmaking approach [J]. Advances in Engineering Software, 2007, 38(1): 59-67.

[5] 蒋运承,史忠植. 多主体服务组合的优化策略[J]. 计算机工程与应用,2004,40(6):1-3.

[6] HONG W, STONEBRAKER M. Optimization of parallel query execution plans in XPRS [C]// Proceedings of the First International Conference on Parallel and Distributed Information Systems. Washington, DC. IEEE Press, 1991: 218-225.

[7] VIGLAS S, NAUGHTON J F, BURGER J. Maximizing the output rate of multi-join queries over streaming information sources [C]// Proceedings of the 2003 International Conference on Very Large Data Bases. San Fransisco: Morgan Kaufmann Publishers, 2003: 285-296.

(上接第1688页)

统计本文的测试结果,在消歧过程中,词语搭配规则起了巨大作用,涵盖了67.3.%的歧义字段切分,其正确率很高,对于"学会"达到100%。

此外,不使用语法规则,单独使用Naïve Bayes切分的正确率为71%。如果先采用语法规则切分,符合切分条件的有50句,正确率达到92%,再对剩余的5句基于Naïve Bayes切分,正确率为60%。综合而言,单独采用Naïve Bayes正确率为71%,语法规则和Naïve Bayes相结合的方式正确率为89%,提高了18%。由此可见,语法规则的引入优于单纯的Naïve Bayes方法。

4.2 模型分析

同样都是利用上下文信息进行切分,文献[1-2]只考虑了歧义字段前后文中词的出现频率,而忽略了词的搭配和词性信息对于切分的影响。本文算法通过提取词语搭配规则、语法规则和Naïve Bayes方法充分挖掘并利用上下文信息,在歧义字段的切分上拥有更全面和准确的信息和依据。在通用性上,本文模型从一般情况出发,模拟人脑判断歧义字段的思维过程,首先考虑词的搭配,再从词性出发,理论依据充分。

5 结语

组合型歧义字段切分是中文自动分词的难点之一,长期以来一直没有完全解决,成为提高分词准确率,从而进一步推动中文信息处理应用的瓶颈。本文在对现有方法进行深入分析的基础上,提出了一种基于规则挖掘和Naïve Bayes方法的组合型歧义字段切分算法,该算法自动从训练语料中挖掘词语搭配规则和语法规则,然后基于这些规则和Naïve Bayes模型综合决策,对组合型歧义字段进行切分消歧。充分的实验结果表明,相对于文献中的研究结果,本文算法对组合型歧义字段切分的准确率提高了大约8%。

作为下一步工作,将考虑扩大前后文窗口的范围,使上下文的信息量更大,进一步提高切分的正确率。当然,随着前后文窗口的扩大,噪声也会增加,如何在扩大前后文的情况下减小噪声的影响是需要重点考虑的问题。

参考文献:

[1] 曲维光,吉根林,穗志方,等. 基于语境信息的组合型分词歧义消解方法[J]. 计算机工程, 2006,32(17):74-76.

[2] 肖云,孙茂松,邹嘉彦. 利用上下文信息解决汉语自动分词中的组合型歧义[J]. 计算机工程与应用,2001,37(19):87-89.

[3] 李静梅,孙丽华,张巧荣,等. 一种文本处理中的朴素贝叶斯分类器[J]. 哈尔滨工程大学学报,2003,24(1):71-74.

[4] 廉竹钧. 汉语组合型切分歧义字段消歧方法研究[C]//第一届学生计算语言学研讨会. 北京:北京语言文化大学,2002:65-70.

[5] 郑家恒,吴芳芳. 多义型歧义字段切分研究[M]. 北京:清华大学出版社, 1999:129-134.

[6] 曾华琳, 李堂秋. 一种基于提取上下文信息的分词算法[J]. 计算机应用,2005,25(9):2025-2027.

[7] 冯素琴,陈惠明. 一种自组织的汉语组合型歧义消歧方法[J]. 计算机工程与设计,2007,28(3):737-739.