汉语组合型切分歧义字段消歧方法研究

廉竹钧

（北京语言文化大学语言信息处理研究所，北京，100083）

E-mail：lianzj@blcu.edu.cn

摘要：本文提出如下的汉语组合型切分歧义消歧策略：对分、合两种切分形式分布比较均匀的组合型切分歧义字段采用决策表算法消歧；对分、合两种切分形式分布悬殊的组合型切分歧义字段采用人工规则+默认切分形式的方法消歧。本文选择 22 个典型的组合型切分歧义字段作为实验对象，其中 17 个分、合分布均匀的，5 个分、合分布悬殊的。开放测试的结果是：前 17 个和后 5 个的平均准确率分别为 87.82%和 97.70%。

关键词：汉语自动分词；组合型切分歧义；决策表；相似度

引言

分词歧义是汉语自动分词的难点之一。从构成形式来看，汉语自动分词的歧义包括两种：交集型歧义和组合型歧义。组合型歧义又称多义组合型歧义或多义型歧义。如：“将来”

a. 他明天 将 来 上海。　　b. 他 将来 肯定是个大画家。

组合型切分歧义的定义如下（Sun M.S. and Benjamin K.T.(1995)）：

定义 1：汉字串 AB 被称作多义组合型切分歧义，如果满足（1）A、B、AB 同时为词；（2）中文文本中至少存在一个前后语境 C，在 C 的约束下，A、B 在语法和语义上都成立。

其中条件 2 的判断需要一个极大的语料库，这在理论上成立，但在实际中则难以做到。所以本文对条件 2 的判断主要根据人的判断。具体说来，本文中讨论的组合型切分歧义字段是先根据条件 1 由程序从语料中自动获取，然后根据条件 2 进行人工甄别而获得的。

切分歧义研究的重点长期以来主要放在交集型歧义问题上，因为人们一般都有这样一个认识：与交集型歧义相比，组合型歧义在实际语料中出现的机会极少。但实际上，很多词形（word type）在某个特定语境下都有可能成为组合型歧义字段。即，在一个具体文本中实际出现的组合型切分歧义可能并不多，但潜在的此类歧义却非常多。从这个角度来说，组合型切分歧义的消解是一个不宜忽略的问题。此外，组合型歧义的消歧要考虑字段与其上下文的句法和语义关系，研究起来比交集型歧义更难。已有的组合型歧义消歧方法研究有：

（1）在切分和标注一体化的概率模型中进行组合型歧义的消歧。白拴虎（1995）。这种方法对于交集型歧义的处理比较有效，但对于组合型歧义的处理效果则不太理想。因为包含两种切分形式的两条路径不等长，在计算路径概率时词数少的路径占优，所以结果往往会选择合的形式；（2）采用统计和规则相结合的方法进行消歧。郑家恒、吴芳芳（1999）。其统计方法等同于选择概率最大的切分形式，小概率的切分形式将会被忽略，而通过人工编写切分规则进行消歧，不仅费事费力，而且很难覆盖所有的语言现象；（3）采用基于向量空间模型的统计方法进行消歧。孙茂松（2001）。该方法将组合型歧义消歧问题与 WSD 问题相提并论抓住了组合型歧义消歧问题的实质，不过采用向量空间法为基本模型，只考虑特征

词的信息，不可避免地会遇到严重的数据稀疏问题，为此而不得不采取一些补救手段。

1 本文研究策略的提出

我们首先从98年1月的人民日报语料(由北京大学计算语言学研究所和富士通研究开发中心有限公司共同制作）中采集到615个组合型歧义字段，分析了它们的消歧条件和分布情况，然后在此基础上提出本文的研究策略。

1.1 歧义字段的分类

我们从两个角度对组合型切分歧义字段进行了分类。第一个分类标准是根据消歧信息与歧义字段距离的远近，可分为两类（分类一）：（1）利用近距离信息的。如，歧义字段“年内”只要获知前一个词的词性为数词，就可确定其应采用分的形式；（2）利用远距离信息的。如，歧义字段“的话”，如果句首有“如果”或者“如果说”与之呼应，则采用合的切分形式的可能性较大。在615个字段中，38个字段（占总数的6%）的消歧可利用远距离信息。第二个分类标准是根据两种切分形式在文本中的分布状况，分为两类（分类二）：（1）分布比较均匀的；（2）分布悬殊的。①

1.2 处理策略的提出

对分类一的分析发现，大多数字段都是利用近距离信息进行消歧，而且对于某一个字段来说，它所能利用的消歧信息并不一定只局限于其中的一种，而往往是远近结合的（如，某个歧义字段的一种切分形式需远距离信息，而另一种切分形式则只需近距离信息，或该字段的一些远距离信息和近距离信息相辅相成，共同提示歧义字段的正确切分形式）。因此，在本文的研究中，我们将暂不考虑远距离信息，选择一种能充分利用近距离信息进行消歧的算法：决策表算法（Yarowsky，1994）。该算法的理论依据是：歧义字段在特定的上下文中是非歧义的，歧义字段的各搭配词（包括标点符号）往往为歧义字段词性和词义的确定提供有力而且一致的线索。算法的基本过程如下：

（1）获取歧义字段训练集，从大规模语料（已分词并标有词性的标准语料）中获取目标歧义字段的前后各k个词的上下文，并将采用相同切分形式的上下文例句保存在一起。

（2）统计搭配信息。统计上下文中出现的每个搭配信息在合与分两种切分形式下的分布情况。分析发现，近距离信息通常包括词形信息和词性信息，因此，在考虑搭配信息时，我们将同时考虑上下文中的搭配词信息和搭配词的词性信息。这不但可以避免严重的数据稀疏问题，也可以更多地利用语言单位组合的语言学性质。

关于上下文窗口的大小，可以预见，如果开得太大，会在取得较远信息的同时，带进许多无用信息，造成噪音干扰。本文分别取左右各3、4、5个词，对9个歧义字段进行开放测试，k=4时，平均准确率最高。因此，本文上下文大小为目标歧义字段左右各取4个词。

Yarowsky(1994)和孙茂松(2001)的研究都表明搭配信息在上下文中所处的位置对消歧是有影响的，因此，在统计时，我们将位置因素也考虑在内（本文约定“+1”、“+2”、“+3-4”分别为目标词右边第一个、第二个和第三个到第四个之间的位置，“-1”、“-2”、“-4-3”分别为目标词左边第一个、第二个和第三个到第四个之间的位置）。

（3）获取决策表。利用公式（1）计算每个搭配信息的相似度（loglikelihood），根据相似度的大小进行排序，并按照大小顺序进入决策表。

---

①为了度量组合型歧义字段的两种切分形式在语料中分布的差异，我们定义分布均匀度这一概念来度量这一差异：$c=\frac{\max(c1,c2)}{\min(c1,c2)}$。c1，c2分别为合和分的形式的出现次数，c≥1。当c1=c2时，c=1。c的值越接近“1”，该歧义字段合、分切分形式的分布越均匀。本文根据训练语料的规模和内容，以c=10作为界限来区分第1类（分布比较均匀的）和第2类（分布悬殊的）。在615组歧义字段中，c>10的有191组，占总数的31%。

$$Abs(Log(\frac{Pr(Case_1 \mid Collocation_i)}{Pr(Case_2 \mid Collocation_i)})) \qquad (1)$$ ①

（4）决策表的运用。决策表中的规则只有前面几条的作用比较大和有效。本文在应用决策表时，选取决策表的前 10 条。此外，仅靠决策表获得的数条规则是不可能应付千变万化的自然语言的，因此，我们又加设一条规则：当前面 10 条规则都无法对当前歧义字段进行消歧时，如果在同一段落中，已出现过相同的歧义字段，则采用与前面歧义字段相同的切分形式；如没有，则采用规定的默认切分形式（默认切分形式的规定见 2.1）。

对 1.1 中的分类二，第 1 类（分布比较均匀的）采用决策表算法无疑可以自动获取两种切分形式的有关规则，而第 2 类（分布悬殊的），如果一个歧义字段的分布非常悬殊，如一个字段的一种切分形式在文本中占 99%，另一种切分形式占 1%，那么从实用的角度考虑，可以忽略低概率的那种切分形式。如果两种切分形式的分布差异较大时，由于歧义字段出现概率低的那种形式在训练语料中很难搜集到足够多的信息来自动获得切分规则，此时，就需依赖于人的语言知识来人工制定切分规则。

综上所述，我们的研究将采用如下方法：对“合”、“分”分布差异悬殊的歧义字段，由人工总结出现概率低的形式的规则，先利用这些规则处理歧义字段，规则不能解决的一律切分为出现概率高的形式；对“合”、“分”分布比较均匀的歧义字段，采用决策表算法得到若干切分规则，运用这些规则处理歧义字段。

2. 实验报告与分析

实验所用的训练语料是北大与富士通合作制作的 1998 年 1 月的人民日报语料。研究进行的测试为开放测试，测试语料是 1998 年 7 月的人民日报语料，先利用程序进行了自动初始标注：分词并标注有可能性最大的词性。初始标注采用和训练语料相同的词语分词和词性标注规范。初始标注时，所有组合型切分歧义都处理成合的情形。

为了便于问题的研究，我们在 1.1 中对组合型切分歧义字段进行的分类的基础上，选择了 22 个典型歧义字段作为实验用字段。合、分两种切分形式分布比较均匀的实验用字段为 17 个：大小、学会、一道、上来、前后、至今、更是、的话、不要、不管、上去、总会、最好、内在、东西、个人、市区。合、分两种切分形式分布差异悬殊的实验用字段为 5 个：三国、一生、支队、人才、最近。

2.1 运用决策表消歧

如前所述，该方法针对合、分两种切分形式分布差异比较均匀的歧义字段。设在训练语料中出现次数较多的切分形式为 CASE1，出现次数较少的为 CASE2。在获得的决策表中，第一条规则都是 CASE1 的规则，原因在于充足的训练集信息保证其具有较高的相似度。

分析获得的决策表，我们发现：（1）分布均匀度 c<=2.5 时，决策表前 10 条规则中 CASE1 和 CASE2 的比例基本持平，分布比较均匀；（2）2.5<c<=4 时，CASE2 的数目少于 4 个；（3）4<c<=10 时，CASE2 的数目少于 2 个。后两种情况是因为训练集中 CASE2 的例句太少，得到的搭配信息的相似度很难超过 CASE1，使决策表前 10 条规则中 CASE1 占据绝对优势。因此，这两种情况需对决策表进行一些调整。调整后，三种情况运用的决策表分别如下：

情况（1），取决策表中的前 10 条规则，默认切分形式采用与第一条规则相反的切分形式（既 CASE2 的规则）；情况（2），将第一条 CASE2 的规则提升到第二位，第二条 CASE2 的规则提升到第五位。取调整

① $Pr(Case_1 \mid Collocation_i)$ 表示某个搭配信息出现的情况下切分形式 1 出现的概率，$Pr(Case_2 \mid Collocation_i)$ 表示该搭配信息出现的情况下切分形式 2 的出现概率。在搭配信息的分布数据表中常常会出现零的情况，在计算相似度时，需对分布数据进行平滑。本文采用了一种简单的数据平滑方法，即为分子和分母分别加一个参数 α：m/n -> (m+α)/(n+α)（m、n 分别为搭配信息在歧义字段合与分两种形式的上下文中的出现次数）。

后的决策表中的前 10 条规则，默认切分形式采用与第一条规则相反的切分形式（既 CASE2 的规则）；情况（3），将第一条 CASE2 的规则提升到第二位，第二条 CASE2 的规则提升到第五位。取调整后的决策表中的前 10 条规则，默认切分形式采用与第一条规则相同的切分形式（既 CASE1 的规则）。

以下 17 个组合型歧义字段开放测试的结果证实这样的调整是比较有效的。

（1）c<=2.5（平均准确率为 89.45%）

表 2-1　c<=2.5 的字段的实验结果

分布均匀度 c	歧义字段	基本准确率（%）	消歧准确率（%）	消歧准确率-基本准确率[①]（%）
1.04	大小	66.67	86.67	20
1.19	学会	62.12	87.88	25.76
1.33	一道	53.13	90.63	37.5
1.5	上来	79.25	93.40	14.15
2.0	前后	80.95	85.71	4.76
2.5	至今	68.29	92.68	24.39

（2）2.5<c<=4（决策表调整前的平均准确率为 77.38%；调整后的平均准确率为 77.48%）

表 2-2　2.5<c<=4 的字段的实验结果

分布均匀度 c	歧义字段	基本准确率（%）	消歧准确率(%)		消歧准确率-基本准确率（%）	
			决策表调整以前	决策表调整以后	决策表调整以前	决策表调整以后
2.63	更是	58.54	73.17	75.61	14.63	17.07
2.86	的话	72.22	79.16	75	7.54	2.78
3.1	不要	95.96	79.80	81.82	-16.16	-14.14

（3）4<c<=10（决策表调整前的平均准确率为 85.44%；调整后的平均准确率为 90.49%）

表 2-3　4<c<=10 的字段的实验结果

分布均匀度 c	歧义字段	基本准确率（%）	消歧准确率(%)		消歧准确率-基本准确率（%）	
			决策表调整以前	决策表调整以后	决策表调整以前	决策表调整以后
4.27	不管	80.49	82.93	87.80	2.44	7.31
4.67	上去	77.42	80.64	90.32	3.22	12.9
5.83	总会	80	80	85	0	5
7.0	最好	86.36	92.05	94.32	6.14	7.72
7.0	内在	91.43	91.43	91.43	0	0
8.58	东西	82.57	84.40	84.40	1.83	1.83
8.74	个人	77.33	84.73	98.57	7.4	21.24
8.83	市区	90.48	87.30	92.06	-3.18	1.58

2.2 运用人工规则消歧

[①]歧义字段在测试语料中出现次数多的那种切分形式占该歧义字段总数的比例可看作基本准确率。设利用消歧规则进行消歧后取得的准确率为消歧准确率，则消歧准确率与基本准确率之差反映消歧规则所起的实际作用的大小。

如前所述，该方法针对合、分两种切分形式分布差异悬殊的歧义字段。由人总结5个实验字段的消歧规则。消歧规则及结果如下列两表：

表2-4 c>10的字段的消歧规则

歧义字段	消歧规则
三国	If(w_{+1}="演义"‖"时期"‖"时代"),then(合的形式);Else(分的形式)
一生	If((w_{-1}！="的")‖(w_{+1}!="里"‖"中")&&(同句范围内，歧义字段后有"就")),then(分的形式);Else(合的形式)
支队	If(w_{-1}="这"‖"那"‖"多"),then(分的形式);Else(合的形式)
人才	If(w_{-1}="个"),then(采用分的形式);Else if(同句范围内，歧义字段前有"只有"),then(分的形式);Else(合的形式)
最近	If(w_{+1}="处"‖"距离"),then(分的形式);Else if(同句范围内，歧义字段前有"离"或"距"或"距离"), then(分的形式);Else(合的形式)

（注：w_{+1}指目标歧义字段右一位置上的词，w_{-1}指目标歧义字段左一位置上的词）

表2-5 c>10的字段的实验结果

歧义字段	基本准确率（%）	消歧准确率（%）	消歧准确率-基本准确率（%）
三国	96.36	100	3.64
一生	100	91.43	-8.57
支队	96.72	98.36	1.64
人才	99.07	99.38	0.31
最近	97.70	99.34	1.64

2.3 实验分析

利用决策表算法进行消歧的17个实验歧义字段的平均准确率为87.82%。就目前的实验结果来说，决策表算法的效果还不太理想，但是该算法的特点在于：①算法的实现比较容易。②算法同时利用了词和词性的信息进行消歧。③利用统计方法获取规则并使之具有先后顺序，较之由人工总结规则并由人决定规则的先后顺序，不但省时省力，而且更合理。④由决策表算法获得的规则可为其它应用直接使用。分析消歧错误的原因，我们发现主要由两点引起：1）训练语料的规模不够大，内容不够丰富，训练语料所用的标记集不够细化。2）算法只利用了上下文中线性的词或词性信息，如果利用部分语义信息将会有所帮助。

利用人工规则+默认切分形式的方法进行消歧的5个实验歧义字段的平均准确率为97.70%。表面看来，所得准确率一般都很高，但与基本准确率一比，提高得却并不多。这是因为这类歧义字段出现概率小的那种切分形式的例句很少，因此，有些规则没有总结出来，有些规则总结出来了，但又不具普遍性。这类合、分两种切分形式分布悬殊的字段具有不可预测性，且总数占所有组合型歧义字段的比例并不少，采用人工规则+默认切分形式的方法进行消歧就目前来说基本能满足某些应用的要求，但如要进一步提高消歧精度，则还需进行更多的分析和研究。

3 结束语

本文针对两种分布情况不同的组合型切分歧义字段，提出了不同的处理策略。实验中只处理了22个典型字段，而该问题的实际情况还要复杂得多，所以本文所做的工作只是迈出了第一步。

参考文献：

[1] Sun. M.S. and Benjamin K.T., 1995, Ambiguity resolution in Chinese word segmentation. Proceedings of the 10th Asia Conference on Language. Information and Computation,121-126. Hong Kong

[2] Yarowsky, David, 1994, Decision Lists for Lexical Ambiguity Resolution: Application to Accent Restoration in Spanish and French, 32nd Annual Meeting of the Association for Computational Linguistics, Las Cruces, NM

[3] 白拴虎，1995，汉语词切分及词性自动标注一体化方法，《计算语言学进展与应用》北京：清华大学出版社，56-61

[4] 孙茂松，2001，汉语自动分词研究的若干最新进展，中国中文信息学会20周年学术会议，清华大学出版社，20-41

[5] 郑家恒、吴芳芳，1999，多义歧义切分方法研究，《计算语言学文集》，北京：清华大学出版社，129-134

致谢 本文的研究、写作过程中得到孙宏林老师的悉心指导和热情帮助，宋柔老师和杨尔弘老师也给我提出了很多有益的建议，在此一并表示感谢。

A Study on the Disambiguation of Combinatorial Ambiguities in Chinese Word Segmentation

Lian Zhujun

(Beijing Language & Culture University, Beijing 100083, China)

E-mail: lianzj@blcu.edu.cn

Abstract: Two different approaches to the problem are proposed for dealing with two kinds of the cases in this thesis: 1) Decision list algorithm is deployed for the cases whose two segmented forms have even distribution in the text; 2) Rules devised by humans are applied for tackling the cases whose two segmented forms have uneven distribution.22 typical examples are chosen in our experiment, including 17 evenly distributed and 5 unevenly distributed. The average accuracies for the two kinds of examples are 87.82% and 97.70% respectively.

Keywords: Chinese word segmentation, combinatorial ambiguity, decision list , log likelihood