一个汉语短语自动界定模型*

周 强

(北京大学计算语言学研究所 北京 100871)

摘要 本文提出了一个汉语短语自动界定模型,它通过基于统计的自动界定处理,利用通过错误驱动自动学习而得到的调整规则进行界定情况局部调整,利用人工总结的全局调整规则进行精调整等3个处理阶段,可以较好地确定一句经过正确切分和词性标注处理的汉语句子中不同短语的边界位置,从而为进一步的汉语短语自动划分和标注处理打下了良好的基础. 对1 000多句句子的实验结果表明,模型的界定正确率达到了96.33%(封闭测试)、94.54%(开放测试).

关键词 汉语短语界定模型,短语划分,语料库自动标注.

汉语短语的界定是指给定一句经过正确切分和词性标注的汉语句子,如何确定短语的边界位置,即哪个词语处于短语的左边界,哪个处于右边界,哪个处于中间位置,此问题的正确解决,对汉语短语的自动划分和标注具有重要意义.

在汉语中,某些虚词,如:助词、介词、连词、副词等,在短语中的位置一般比较固定. 而一些实词,包括动词、名词、形容词等,在短语中的位置则比较灵活. 但是通过利用上下文词类信息,考察不同的词类组合模式,我们还是可以找到一些确定短语边界的规律的. 对大量的人工划分语料进行类似的信息统计,就可以为自动界定短语提供许多有用的数据.

在对此问题进行了深入研究的基础上,我们提出了一个汉语短语自动界定模型,它分为以下3个处理阶段,通过统计处理和机器学习,并结合人的丰富的语言学知识,取得了较好的处理效果.

(1)利用从树库语料中统计得到的数据,构造统计模型,进行短语自动界定处理.

(2)将自动处理结果和人工校对结果相比较,发现错误事例,从中自动归纳界定情况局部调整规则,然后,将自动习得的规则运用于统计处理结果,以达到降低错误率的目的.

(3)总结汉语短语的远距离依赖现象,依据人的语言学知识,归纳总结一些界定情况全局调整规则,将这些规则运用于自动调整结果,以减少误调整次数,进一步降低处理错误率.

1 统计模型设计

1.1 基本统计模型

* 本文研究得到国家自然科学基金资助. 作者周强,1967年生,博士生,主要研究领域为语料库语言学,机器翻译,计算语言学.

本文通讯联系人:周强,北京 100871,北京大学计算语言学研究所

本文1995-09-14收到修改稿

© 中国科学院软件研究所 http://www.jos.org.cn

令 $S=\langle W,T\rangle$ 为短语分析的原始输入句子，其中 $W=w_1,w_2,\ldots,w_n$ 为句子中的词语串，$T=t_1,t_2,\ldots,t_n$ 为各词语的词类标记串. 设 $B=b_1,b_2,\ldots,b_n$ 为句子中每个(词语/词类)对所对应的短语划分情况，b_i 可取值{0—不分(w_i/t_i)，1—左分$([w_i/t_i)$，2—右分$(w_i/t_i])$}. 这样短语界定的工作就变成寻找一个划分点序列 B'，使得：

$$B'=\arg\max_{B\in\beta}P(B|S)=\arg\max_{B\in\beta}P(S|B)P(B) \tag{1}$$

假设词语和词类信息对短语界定的作用是独立的，则可得：

$$P(S|B)=P(W|B)P(T|B) \tag{2}$$

假设每个划分点的确定是独立的，并且只与局部的词语词类信息相关. 在此条件下，对式(2)进行简化，可得到式(3)，其中 Ti 表示所考虑的上下文词类信息.

$$P(S|B)=\prod_{i=1}^{n}P(wi|bi)P(Ti|bi) \tag{3}$$

而对于 $P(B)$，利用 bigram 模型进行简化，得到：

$$P(B)=\prod_{i=1}^{n}P(bi|bi-1) \tag{4}$$

综合式(1)～(4)，可得：

$$B'=\arg\max\prod_{i=1}^{n}P(wi|bi)P(Ti|bi)P(bi|bi-1) \tag{5}$$

通过 MLE 方法估计概率参数，利用 Viterbi 算法计算最佳路径，就可以完成对一个句子的短语自动界定.

1.2 数据统计和参量估计

从树库语料中，我们可以统计得到一组用于界定短语的数据，包括：

(1)词语与界定情况的共现频度：$[w,w],w$

它们反映了不同词语的特殊性，对提高统计模型的处理正确率起了重要作用.

(2)词类与界定情况的共现频度，包括：

① 1 个词类的情况：$[t,t],t$

② 2 个词类的情况：$[t_i\ t_{i+1},t_i[t_{i+1},t_{i-1}\ t_i],t_{i-1}]\ t_i,\ t_i\ t_{i+1},t_{i-1}\ t_i$

它们是统计模型的基本数据. 利用这些数据，通过 MLE 方法进行参量估计，可以得到：

$$P(wi|bi)=f(wi,bi)/f(bi) \tag{6}$$

$$\begin{aligned}P(Ti|bi)&=\max[P(ti,ti+1|bi),P(ti-1,ti|bi)]\\&=\max[f(bi,ti,ti+1)/f(bi),f(ti-1,ti,bi)/f(bi)]\end{aligned} \tag{7}$$

其中(7)式充分考虑了 ti 的上下文信息对界定情况的影响，并利用 backing—off 方法对式(7)进行了数据平滑(Smoothing)处理，即：

当 $P(Ti|bi)=0$ 时，取 $P(Ti|bi)=P(ti|bi)=f(ti,bi)/f(bi)$ (8)

2 错误驱动的调整规则自动学习

将经过统计模型自动处理的界定结果与人工校对结果相比较，就可以发现错误的界定及其相应的上下文信息，它们组成了一组错误事例，这是调整规则学习的生长点.

另外，从“汉语语法电子词典”[1]中，我们可以得到每个词语丰富的句法特征信息，这是

© 中国科学院软件研究所 http://www.jos.org.cn

规则学习的基础,从中可以归纳总结出许多有用的规则约束条件.

2.1 调整规则的形式和层次组织

为正确地调整相对于某个词的界定情况,需参考一定的上下文信息.目前,我们只考虑左右各一个词语的语法特征信息(即观察窗口大小 $W=3$),从而形成了以下的调整规则基本模式:

$$\langle FR_L, FR_M, FR_R\rangle :: B_{me} \rightarrow B_{mc}, P_{err}$$

其中,FR_L,FR_M 和 FR_R 表示左词语、中间词语和右词语的特征约束(Feature Restriction),B_{me}和 B_{mc}分别表示相对于中间词语的错误界定和正确界定情况,P_{err}则反映了在规则的特征约束条件下,界定 B_{me}出错的可能性大小.在实际使用过程中,可选定一个概率阈值 T,将满足条件:$P_{err}>T$ 的调整规则作用于统计处理结果,可收到较好的处理效果.

根据特征约束深度的不同,可以把同一词类约束下的调整规则组织成图 1 所示的层次树结构.其中树的根节点是最概括的词类约束$\langle C_L, C_M, C_R\rangle$,而树的叶节点,则罗列了所有由不同词语组$\langle W_L, W_M, W_R\rangle$组成的错误事例,中间的节点则反映了不同层次的特征约束向量(FRV)的信息:越靠近根节点,则特征约束向量所具有的特征越具有概括性,能覆盖更多的错误事例;越靠近叶节点,则特征约束向量的概括性越弱,特殊性越强.

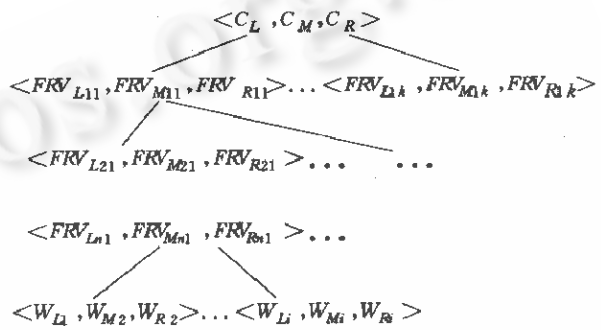

图1 调整规则的层次树组织

将不同词类约束下的规则组织成一个树表,就形成了调整规则的整体结构.而如何从大量的错误事例中归纳出不同层次的特征约束向量,则是错误驱动的规则自动学习所要解决的主要问题.

2.2 基本学习流程

目前的规则自动学习流程主要分为 2 个阶段:

(1)学习阶段:增加树表信息,自动学习规则结构树

Ⅰ. 比较自动处理结果和人工校对结果,得到错误事例.

Ⅱ. 根据词类约束信息,将错误事例插入树表中的适当位置,错误事例计数器 Ne 加 1.

Ⅲ. 如果某个词类约束下的错误事例数目超过了归纳学习启动阈值,则进行自动归纳学习:①检索语法电子词典,得到有关词语的句法特征信息.②通过自动归纳,生成或更新调整规则的层次结构树.③路径上的错误事例计数器 Ne 加 1.

(2)验证阶段:利用正确事例搜索规则树,计算 P_{err}

Ⅰ. 得到正确事例,检索有关词语的句法特征信息.

Ⅱ. 搜索规则树,得到满足语境特征约束的最深节点.

Ⅲ. 在路径上的所有正确事例计数器 Nc 上加 1.

这样,对每条习得的规则,就有:$P_{err}=Ne/(Ne+Nc)$

2.3 GBM 模型及其改进

GBM(generalization—based memory)模型是 Michael Lebowitz 提出的一种概念自动

© 中国科学院软件研究所 http://www.jos.org.cn

学习工具[2],用于从大量复杂的例子(Examples)中归纳出重要的普遍性概念(General Concepts).它具有识别和定义多重概念、进行增量(Incremental)学习和处理大规模例子的能力,可以很好地适应汉语短语界定的调整规则自动学习的要求.为此,我们吸收了 GBM 的基本处理思想,并对此进行了一些改进,形成了一个效率较高的调整规则自动学习模型.下面首先简单地介绍一下 GBM 模型的节点结构和基本算法,然后提出我们的改进措施.

2.3.1 GBM 的结构组织和基本算法

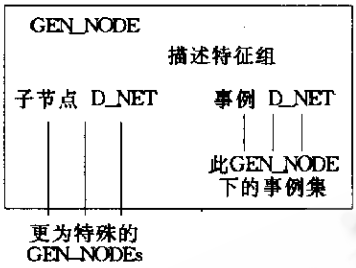

图2 GEN_NODE的结构

GBM 实际上是有一层或多层的概括节点 GEN_NODEs 所组成的一棵树(或一个网络).在 GEN_NODE 中,则包含了子节点和事例(Instances)的信息,它们通过区分网络(Discrimination Networks, D—NETs)而标识.图 2 给出了 GEN_NODE 的基本结构.

GBM 处理的基本流程为:

(1)接受一个新的输入事例(具有一组特征).

(2)搜索 GBM,得到能描述这个事例的最特殊的 GEN_NODE(Search_GBM 函数).

(3)将新事例加入 GBM 中,若条件许可,就进行概括处理(Update_GBM 函数).

有关搜索函数 Search—GBM()和更新函数 Update—GBM()的细节详见文献[2].

2.3.2 我们的改进措施

为使 GBM 模型更好地运用于短语界定调整规则的自动学习,我们对此模型进行了以下改进:

(1)节点结构的调整

为适应调整规则特征约束向量归纳概括的要求,对 GEN_NODE 的结构作了以下调整:①将描述特征组改为描述特征向量组,保存归纳得到的观察窗口中的特征约束信息;②将区分网络 D_NETs 扩充为向量结构,便于进行特征向量检索;③事例集中保存了观察窗口中所有词语的其它信息(未归纳的).另外,还在 GEN_NDOE 中增加了一些新内容,如:正确界定情况 B_{mc},错误事例计数器 Ne,正确事例计数器 Nc 等.

(2)设置子节点再概括机制

对最初 GBM 的处理结果进行分析,我们发现了这样一种现象:同一层次的 GEN_NODE数目很多,它们一般包含了两三个事例,并且具有许多重复的概括特征.究其原因,主要是由于词语的句法特征比较多,其中只要有 1 个与 GEN_NODE 的描述特征不匹配,就可能导致产生新的子节点.为此,我们增加了 1 个子节点的再概括处理过程(步骤 4),以提高特征向量的概括性,减低特征的冗余度.

3 实验结果分析

本实验所用的统计和测试语料主要选自汉英机器翻译研究的测试题库.它们句型多样,不同短语组合的分布也很广,而且句子较短,便于进行自动分析处理.语料的规模为 1 434 个汉语句子,约 11 830 个词,平均句长为 8.25 词/句.对这些句子进行人工短语划分和标

© 中国科学院软件研究所 http://www.jos.org.cn

注,并进行了多次校对,形成了一个准确度较高的由人工标注的树库(Treebank).它们是进行统计处理、调整规则自动学习和人工总结以及系统性能评估的基础.

对于系统的整体处理性能,我们主要考虑了以下几个性能指标:

①界定情况错误率

它反映了经过模型处理的界定情况与正确界定的差异比率,包括3个处理阶段的结果:统计模型处理(SP)、自动学习规则校正(AT)和人工总结规则校正(RT).

②规则误调整率

将实验语料分成2个部分:1 300个句子组成封闭测试语料,用于得到统计数据、进行调整规则自动学习和人工总结;另外143个句子组成开放测试语料.得到了以下的实验结果(表1、表2):

表1　界定处理错误率

	SP	SP+AT	SP+AT+RT
封闭测试(%)	8.188	4.562	3.674
开放测试(%)	8.803	6.690	5.458

表2　规则误调整率

	AT	RT
封闭测试(%)	14.202	7.965
开放测试(%)	21.053	6.250

其中规则归纳学习启动阈值LT=10,调整阈值PT=0.75.

分析自动界定和调整处理结果,可以发现一部分是由于错误调整所引起的,但绝大多数还是那些没有经过调整处理的错误界定.表3列出了一些界定实例,其中错误界定通过下划线标出,而正确界定则以数值形式描述(见1.1节).

表3　部分界定实例

1...[很多/m 人/n][来/v (1)寄/v]东西/n]

2[中国/n (1)科学/n]技术/n]情报/n]学会/n][是/v[一个/m 群众组织/n]

3[我们/r[要/v[输送/v[大批/b[有/v[文化/n(2) 的/u 劳动者/n]

4[他/r [跟/p [你/r(0)一样/u(2) 勇敢/a]

5[我们/r[一定/d[能/v[进一步/d (1)认识/v]和/c[掌握/v(2)[语言/n 结构/n]的/u 规律/n]

对那些错误进行分类,可以发现主要由以下原因造成:

(1)自动归约的不完全;

(2)歧义结构的优先选择,如:表3的例1、例2;

(3)复杂结构和固定搭配问题,如:表3的例3、例4;

(4)并列结构的分析难点,如:例5;

对此模型,还有2个重要的问题是:

①统计模型的收敛性问题:即树库语料达到什么规模可以使统计模型的处理性能达到稳定.

②归纳学习的性能增长极限问题:即需要多少错误实例才能习得合理的调整规则.

下面通过2个实验对此进行初步的分析.

3.1　统计模型的收敛性分析

将1 300个句子,从100个句子出发,每次增加100句,形成一个封闭测试语料序列集(共13个元素),而以134个句子组成共同的开放测试语料.记录每次处理的界定错误率,得到了图3的结果.

© 中国科学院软件研究所 http://www.jos.org.cn

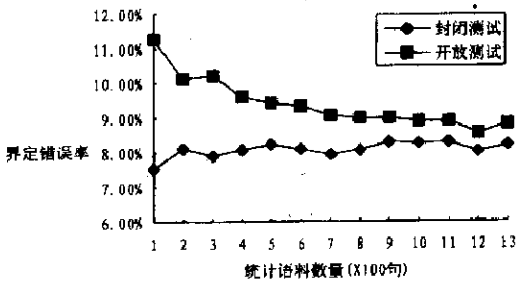

图 3 统计模型处理性能分析

从图 3 中可以发现：(1)封闭测试的界定错误率在 8.10%附近作微小的波动，基本保持稳定.(2)开放测试的界定错误率，则在经历了较大幅度的下降之后，在达到 1 000 句统计语料的规模后逐渐趋于稳定.

当然，如果有更多的树库语料(如 5 000 个句子)，界定错误率的变化趋势可能会更明显.但在目前，我们也可以初步得出结论：对目前的统计模型，1 000 个左右正确标注句子的统计数据基本上可以满足处理要求了.

3.2 自动归纳学习的性能分析

对于归纳学习性能的测试，则采用了从 200 个句子出发，每次增加 200 个的封闭测试序列集(共 7 个元素)，开放测试语料仍为 134 个句子.利用从 1 300 个句子中统计得到的有关数据，对每个封闭测试语料进行统计界定处理和自动归纳学习，并利用习得的规则进行自动调整处理，得到了表 4 的结果(另见图 4、5).

表 4 自动规则学习实验结果

语料数量(×100)	2	4	6	8	10	12	13
归纳的错误事例数	10	35	111	158	260	350	384
未归纳的错误事例数	142	249	302	375	409	466	492
习得的 GBM 规则数	5	19	54	74	111	146	151
词类约束规则数	85	128	159	180	201	223	232
SP 界定错误率(封闭)	8.994	8.460	8.204	8.185	8.171	8.289	8.188
SP+AT 界定错误率(封闭)	5.089	5.005	4.589	4.561	4.470	4.561	4.561
SP+AT 界定错误率(开放)	8.011	7.394	7.042	7.130	6.690	6.778	6.690
AT 误调整率(封闭)	0.000	5.426	9.502	11.111	12.137	14.079	14.202
AT 误调整率(开放)	36.667	32.500	29.546	31.915	27.273	26.191	21.050

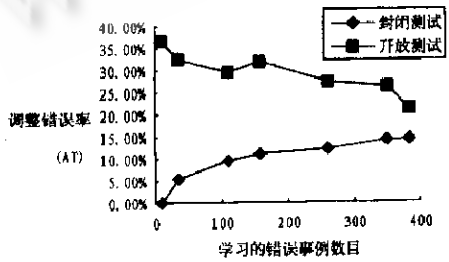

图 4 学习事例与误调整率关系

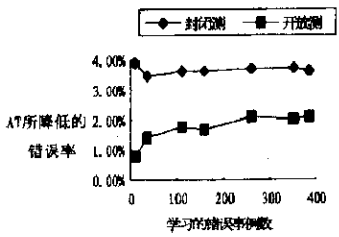

图 5 学习事例与降低的界定错误率

© 中国科学院软件研究所 http://www.jos.org.cn

尽管由于用于归纳学习的错误事例数量的限制,规则学习的性能增长并没有趋于稳定,但从表 4(图 4、5)中,我们还是可以发现许多有意义的发展趋势:

(1)随着处理语料数量的不断增大,归纳的错误事例在总的错误事例中所占的比例也在不断增大,而新出现的词类约束规则数目则在不断减少.这表明统计模型的处理错误开始逐渐集中于某几个比较典型的词类组合上,从而更有利于自动归纳的处理.

(2)图 4 中,封闭测试误调整率的逐渐增加和开放测试误调整率的逐渐降低,从一个侧面反映了自动习得规则描述能力的不断提高.当处理语料较少时,由于错误事例的词类组合分布分散,并缺少足够的正确验证事例,使习得规则的 Perr 不能很好地反映语言事实,因此,虽然对封闭语料调整得很好,但一旦应用于开放语料,误调整率就大大增加了;随着处理语料的不断扩大,错误事例开始逐渐集中于几个典型的词类组合上,不同的正确验证事例数量也在不断增大,从而使习得规则的 Perr 更符合语言事实,这样特征阈值 PT 就开始发生作用,使开放测试和封闭测试的误调整率接近 1−PT.

(3)图 5 中,自动习得的规则对开放语料调整能力的不断提高,则从另一侧面反映了规则描述能力的增强.随着进行归纳的错误事例数量的增大及再概括机制的使用,使习得规则的概括性越来越高,其中的 Perr 也更能反映实际的语言事实,从而能将更多的未经学习的错误事例正确地调整过来.

4　结　语

本文介绍了一个汉语短语自动界定模型,它具有以下几个特点:

(1)普遍性知识和特殊性知识相结合;

(2)局部调整规则学习和全局调整规则总结相结合;

(3)人机处理相结合.

从目前的实验结果看,它已经取得了很好的处理效果.在以后的研究中,我们将在以下几方面对此模型进行改进,以进一步提高它的处理性能:

(1)提取更多的语境特征,改进统计模型(类似文献[3]的处理).

(2)利用语义信息,提高调整规则的调整约束能力.

(3)逐步扩大处理语料的规模,从中发现并总结更多的全局调整规则.

致谢　北京大学计算语言学研究所的许多老师和同学为语料的人工标注和校对做了大量工作,我的导师俞士汶教授以及不知名的审稿者都对论文的初稿提出了许多宝贵的意见,这里一并表示感谢.

参考文献

1　俞士汶,朱学锋,郭锐.现代汉语语法电子词典的概要与设计.In:Proc. of ICCIP'92, 1992.186～191.

2　Michael Lebowitz. Concept learning in a rich input domain: generalization－rased memory. In: Michalski R S, Carbonell I G, Mitchell T M eds., Machine Learning: a Artificial Intelligence Approach, Chapter 8, 1986.193～214.

3　Lin Y C, Chiang T H, Su K Y. Automatic model refinement－with and application to tagging. In: Proc. of COL-

© 中国科学院软件研究所 http://www.jos.org.cn

ING—94, 1994,1:148~153.

A MODEL FOR AUTOMATIC PREDICTION OF CHINESE PHRASE BOUNDARY LOCATION

Zhou Qiang

(*Institute of Computational Linguistics Beijing University Beijing* 100871)

Abstract Phrase boundary location provides an important information for bracketing and tagging the phrase automatically. This paper describes an experimental model for the automatic prediction of the phrase boundary location. It consists of three processing stages: first, automatically identify the phrase boundaries using statistics from treebank; then, post—tune the results using local tuning rules generated by an error—driven the machine learning method; at last, refine the results of the last two stages with the overall tuning rules summarized by man. Experimental results on a corpus of 1 434 sentences demonstrate a high rate of the success for predicting the phrase boundary (96.33% correct the prediction for the close testing and 94.54% correct the prediction for open testing).

Key words Predicting phrase boundary, phrase bracketing, corpus annotation.

© 中国科学院软件研究所 http://www.jos.org.cn