基于最大熵模型的语义块切分

谢法奎[1,2],张 全[2]
XIE Fa-kui[1,2],ZHANG Quan[2]

1.中国科学院 研究生院,北京 100039
2.中国科学院 声学研究所,北京 100190
1.Graduate University of Chinese Academy of Sciences,Beijing 100039,China
2.Institute of Acoustics,Chinese Academy of Sciences,Beijing 100190,China
E-mail:holinax@tom.com

XIE Fa-kui,ZHANG Quan.Semantic chunks segmentation based on maximum entropy model.Computer Engineering and Applications,2009,45(26):118-120.

Abstract: Semantic chunks segmentaion is an important task in the Hierarchical Network of Concepts(HNC) theory.To deal with this problem,this paper adopts a new method based on statistical modeling.And forms some feature templates with word,POS,concept,and selects features by a incremental way.Finally,construct a semantic chunks segmentation system based on a maximum entropy model.The experiment is taken on HNC corpus,and the result shows that the model works well,the open test precision and recall are 83.78% and 91.17% respectively.

Key words: maximum entropy model;sematic chunk;Hierarchical Network Concepts(HNC)

摘 要:语义块切分是HNC理论的重要课题,与以往的处理策略不同,采用统计建模的方法来解决这一问题。采用词语、词性、概念等信息组成特征模板,并应用增量方法进行特征选择,构建了一个基于最大熵模型的语义块切分系统。在HNC标注语料库上的测试取得了较好的效果,开放测试的正确率和召回率分别达到了83.78%和91.17%。

关键词:最大熵模型;语义块;概念层次网络
DOI:10.3778/j.issn.1002-8331.2009.26.035 **文章编号:**1002-8331(2009)26-0118-03 **文献标识码:**A **中图分类号:**TP391

1 前言

在HNC(概念层次网络 Hierarchical Network of Concepts的简称)理论中,语义块是语句的下一级语义构成单元,相应的,语义块切分是指把一个语句切分成若干个语义块。语义块切分是HNC理论研究的重要内容,与句类分析和语义块内部构成分析有密切的关系,是这两项工作的基础。随着HNC理论研究的深入和大规模语料机器处理技术的发展,语义块切分的重要性正逐渐体现出来。

以往的语义块切分处理研究主要是作为句类分析的一个步骤进行,采用一种基于规则的方案,依靠人工总结来构建规则库,这需要投入大量的人力。该文的思路有所不同,希望利用已有的HNC语义标注语料库,通过基于统计的方法来解决这一问题。时至今日,在自然语言处理领域,已经出现了许多成熟的统计语言模型,其中,最大熵模型以其简洁、通用、易于移植等优点而被广泛地采用。从HNC标注语料库中提取训练和测试数据,利用最大熵原理构建模型,具体分析语义块切分的特点,构造了多种特征模板,并采用增量法进行特征选择,实现了一个基于最大熵模型的语义块切分系统。

2 语义块切分的任务

HNC理论建立了自然语言语句的深层语义结构表示模式,发现了57组基本表示式,运用这些表示式及其混合便可以描述任何句子的语义概念结构。语义块是语句的下一级语义概念构成单元,语义块切分的目标就是确定该句类中各语义块在语言空间中的边界。

以上语义块切分的含义是清晰的,但实际情况要复杂得多,这主要是因为:语义块成为语句的唯一的构成单元,这有利于句类空间知识的净化,但同时也引起了语义块内部构成的复杂化。其中一个典型情况是,为描述语句嵌套的句子,要求语义块能够包含语句,为此HNC引入了句蜕和块扩的概念。句蜕是指一个语句蜕化为语义块或语义块的一部分;块扩是指语义块扩展为一个或多个语句。句蜕有两种基本类型:原型句蜕和要素句蜕。从形式的角度看,语句、语义块、句蜕、块扩这些结构,形成了一套可循环嵌套的表述体系。下面给出一个实例。

原文:随着我国进入工业化中期阶段,经济社会发展对林业要求也在发生根本性的变化。

其语言空间的标注信息为:随着 { 我国 | 进入 | 工业化

基金项目:国家重点基础研究发展规划(973)(the National Grand Fundamental Research 973 Program of China under Grant No.2004CB318104);中科院声学所知识创新工程项目(the Knowledge Innovation Engineering Project of Institute of Acoustics,CAS under Grant No.O654091431)。
作者简介:谢法奎(1981-),男,硕士,主要研究方向:自然语言理解;张全,研究员,博导。
收稿日期:2008-05-19 **修回日期:**2008-08-12

中期阶段 } ~‖,< 经济社会发展 | 对林业 | 的要求 > ‖ 也在发生根本性的变化。

这种语句嵌套的情况带来了语义块切分的困难,难以找到区别语义块所处层级的上下文特征,如果直接针对复杂句子进行语义块切分,难度太大。因而只处理同级的语义块切分情况,即只处理处于同一语句内的语义块。将语义块切分的层级分有4种:句子级、块扩级、原型句蜕级、要素句蜕级。句子级的语句是指直接归属于句子的语句,类似的,其他3种是归属于块扩和原型句蜕、要素句蜕的语句。目标就是对这4种层级的语句进行语义块切分。从上例句中按照层级分离出3个语句,语义块切分结果如下:

句子级:[随着 SD0],[SD1] [也在发生根本性的变化。]

原型句蜕级:[我国] [进入] [工业化中期阶段]

要素句蜕级:[经济社会发展] [对林业] [的要求]

一般的说,句蜕和块扩内的语句与外层语句在语义上并没有直接关联,因而可以使用符号来代替。这里约定使用CE代表块扩,SD0代表原型句蜕,SD1代表要素句蜕。

3　语义块切分模型总体设计

最大熵模型的主要任务是分类,为便于模型的使用,需要把语义块切分问题转化为词语序列的标注问题。为此,使用语义块类型和边界标志共同组成语义块的标注符号,如表1和表2。如果只使用边界标志,分类类型太少,不能有效利用上下文信息,效果不好。

表1　块类型

块类型	含义
EK	特征语义块
FK	辅语义块
JK	广义对象语义块
SEP	语义块的间隔块

表2　边界标记

边界标记	含义
B	开始
E	结束
BE	一个词语作为一个语义块
I	内部

下面给出一个实例。

语义块切分:[我国] [进入] [工业化 中期 阶段]

语义块标记:我国/JK_BE　进入/EK_BE　工业化/JK_B

中期/JK_I　　阶段/JK_E

采用了4种块类型。语义块分为主语义块和辅语义块,主语义块又可以分为特征语义块和广义对象语义块。广义对象语义还可以细分为作用者语义块、对象语义块、内容语义块,但它们没有明显的标志和特征,这里并不适宜。因而一共采用3种语义块分类。另外,语句之间和语义块之间有些成分不适于划归到语义块中,因而单独作为语义块的间隔块,这里主要是指“,”标点符号。

采用了4种边界标记。同时采用B和E标记是为了突出语义块的起始和结尾,这对于特征语义块和辅语义块尤为重要。加入BE标记是为了突出独词语义块。当然,在机器标注时会产生边界矛盾,这里约定:凡是标记为B,E,BE的词语,都认为是语义块的边界。

基于最大熵模型,构造了完整的语义块切分模型,总体框架如图1。

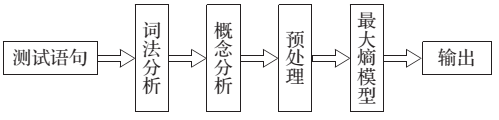

图1　模型总体结构图

具体的说,模型分为以下几部分:

(1)词法分析:即进行分词和词性标注。语义块在形式上与短语等级相似,是介于词语和语句之间的中间层次,词法分析是语义块切分的基础。

(2)概念分析:在HNC理论中,词语有相应的深层语义概念,概念信息对语义块切分有重要作用。

(3)预处理:考查语义块的边界特点,去除冗余信息,提高关键信息的关联性,缩短有效统计特征的上下文长度,改善最大熵模型的效果。主要有以下几方面处理:除少数特殊词汇外,隐藏副词、连词等;对数字等特殊文字以统一符号替代;隐藏引号等对称标点内的文字。

(4)最大熵模型:根据上下文特征信息,给出当前位置的语义块标注符号。这是模型的核心部分。

4　最大熵模型的构建

4.1　最大熵模型介绍

最大熵模型的基本思想是:将已知事实作为制约条件,求得可使熵最大化的概率分布作为正确的概率分布。具体地说,假设模型存在 k 个特征 $f_1,\cdots,f_k$,相应存在 k 个约束,满足这 k 个约束的模型集合为:

$$P=\{p|E_p(f_i)=E_{\tilde{p}}(f_i),i\in\{1,\cdots,k\}\}$$

选取集合中熵值最大的模型作为目标模型,即

$$p^*=\arg\max_{p\in P}H(p)$$

满足上式的解如下:

$$p^*(y|x)=\frac{1}{Z(x)}\exp\Big(\sum_i\lambda_i f(x,y)\Big)$$

其中,$Z(x)$是归一化因子,λ 为特征的权重参数。最大熵模型的建模问题便转化为根据训练数据求解参数 λ 的问题。具体的估算方法有很多,这里采用GIS算法[1]。

4.2　特征表示

最大熵模型的关键在于构造特征集合。总的来说,可利用的已知信息如下:

(1)词。词语本身是最直接的信息。

(2)词性。词性在语法上表现为约束和联系,是重要的上下文信息。从理论上讲,HNC与传统的语法理论是不相关的,但在局部上下文环境下,有很多方面是可以借用的。

(3)HNC概念。HNC概念是指词语在HNC概念网络节点的对应,具体的说就是概念表达式。这里只取节点库的一部分,主要是语言逻辑概念,其包括语义块标志概念、语义块内部连接构成概念等,是语义块切分的重要线索。

(4)语义块标注信息。系统采用从左到右的标注顺序,前面的标注信息对当前位置的标注有重要的提示作用。

根据如上信息,定义了模型的特征模板。如果只考虑一种信息,且特征长度为1,则形成原子特征模板。该文采用的原子特征模板如表3。

表3　原子特征模板

原子模板	含义
$W0;W+1;W-1;W+2;W-2$	词语
$POS0;POS+1;POS-1;POS-2;POS+2$	词性
$C0;C+1;C-1;C+2;C-2$	概念
$Tag-1;Tag-2$	语义块标注

仅仅使用原子模板,不足以表征上下文环境信息。综合使用多个特征,拉长特征长度,便形成了复合模板。该文采用的复合模板如表4。

表4 复合特征模板

序号	复合模板	序号	复合模板
1	*W*-1,*W*0	9	*POS*-2,*POS*-1,*POS*0
2	*W*0,*W*+1	10	*POS*0,*POS*+1,*POS*+2
3	*W*-1,*W*0,*W*+1	11	*Tag*-2,*Tag*-1
4	*POS*0,*POS*+1	12	*W*0,*POS*+1
5	*POS*-1,*POS*0	13	*POS*0,*W*+1
6	*POS*-2,*POS*-1	14	*POS*-1,*W*0
7	*POS*+1,*POS*+2	15	*W*-1,*POS*0
8	*POS*-1,*POS*0,*POS*+1	16	*Tag*-1,*POS*0

一个训练样本(x,y),对模板进行实例化,即可得到具体的特征函数。训练样本集合对所有的模板进行实例化,便得到候选特征集。例如,样本数据为"… 是/v ‖ 宏伟/a …",对于复合模板(*W*0,*POS*+1),可以抽取特征信息"是 a EK_BE",特征函数如下:

$$f(x,y)=\begin{cases}1, y\text{ 为“EK\_BE”,且 }W0\text{ 为“是”,}POS+1\text{ 为“a”}\\0,\text{否则}\end{cases}$$

4.3 特征选择

由特征模板实例化得到的候选特征数量巨大,过多的特征对模型的训练是沉重的负担,而且并不是所有的特征对模型都有贡献,采纳全部特征并不能保证效果最好,因而要对候选特征集合进行筛选,选出价值较高的特征,这便是特征选择。

常用的特征选择方法有基于频次的特征选择方法和增量特征选择方法。基于频次的特征选择方法给定一个阈值*K*,模型只考虑在训练样本集中出现次数大于*K*次的特征。这种方法简单明了,但不能得到一个最小的特征集合,经过实验,要达到理想的效果,仍然需要采用大量的特征。增量特征选择方法能够得到较小的特征集,系统测试速度很快,同时这种方法也有缺点,主要是模型训练费时较多。采用增量特征选择方法,其原理如下。

设*F*是候选特征集合,*S*为有效特征集合。增量法的思想是:首先置*S*为空,从*F*-*S*中抽取一个特征增益最大的候选特征,作为一个有效特征加入*S*,如此循环,便可以得到有效特征集。特征*f*的特征增益的计算方法为

$$\Delta L(S,f)\equiv L(p_{S\cup f})-L(p_S)$$

其中*L*是模型的对数似然。可以证明,*f*的特征增益越大,由$S\cup f$确定的模型$p(x,y)$与经验概率模型之间的Kullback-Leibler距离越小,即模型的质量越高。

以上为增量法的基本思路,实际执行时,每次加入新特征*f*,需要重新计算模型的参数和分布,算法复杂度太高,精确的计算难以实现。因而只能采取近似的处理方式,假设加入候选特征*f*后的模型分布的形式如下:

$$p^{\alpha}_{S,f}=\frac{1}{Z_{\alpha}(x)}p_S(y|x)\mathrm{e}^{\alpha f(x,y)}$$

其中*Z*为归一化因子。该式中只有一个实数参数α,使特征增益最大化,即取值

$$\arg\max_{\alpha}\left(L(p^{\alpha}_{S,f})-L(p_S)\right)$$

通过这种近似手段,无需使用GIS算法计算模型参数便可以直接得到模型分布,使特征选择问题转变为一维极值问题,降低了计算复杂度。具体的算法描述请参见文献[2]。

依据上述特征选择算法,抽取出约3 000个特征(这里指句子级模型),作为模型的特征集。

5 测试及结果

训练数据和测试数据均来源于HNC语义标注语料库,语料库中每一篇文章均以句子为单位进行HNC语义结构标注,可以从标注信息中抽取出不同层级的语句以及语句内的语义块信息。从中选取了150篇文章约6 500句作为训练数据,另外选取15篇文章约500句作为测试数据。需要注意的是,为避免同篇章内的句间干涉,不以句子为单位,而以篇章为单位来选取语料。

根据上文分析,将语句分为4级:句子级、块扩级、原型句蜕级、要素句蜕级,这4级语句在句类分布是不同的,在具体语言特点上也是有很大出入的,句子级语句一般较为饱满,而句蜕级语句因其子句身份一般较为短小,言语简练,块扩级语句介于二者之间。这些不同的特点会影响到模型的特征集,如果统一训练和测试,会出现特征干扰,难以取得理想的效果。因而将这4级语句的语料分开处理,分别进行模型的训练和测试。

测试结果如表5。在计算正确率和召回率时,针对的是语义块序列中任2个语义块交界处的识别情况。在计算*F*值时取β=1。结果表明,句子级语句的语义块切分的效果最好,这主要是因为句子级语句的语料比其他语句更多,因而训练数据最充分。

表5 测试结果

语句的级别	正确率	召回率	*F*
句子级	83.78	91.17	87.32
原型句蜕级	77.94	85.83	81.69
要素句蜕级	80.14	80.74	80.44
块扩级	78.17	81.91	80.00

根据系统的处理结果,分析总结了典型错误。不可避免的,前期处理中的分词、词性标注、概念分析中存在一些错误。除此之外,语义与形式的不一致是导致模型判断失误的重要原因。HNC是基于语义的理论,一些特殊句类对语义块有特殊的要求,尤其是特征语义块以及对象内容语义块,因而更多依赖局部形式特征的统计模型在遇到这种情况时会判断错误。下面是一个典型实例:

原文:"经济社会发展对林业的要求也在发生根本性的变化"

机器切分:[经济社会发展对林业的要求] [也在发生]
[根本性的变化]

正确切分:[经济社会发展对林业的要求]
[也在发生根本性的变化]

错误原因:句类是D1Y*11J,语句的EK关键词是"变化",而不是形式上的"发生"。

目前模型对概念特征的运用是有限度的,更多的是使用局部的词语和词性特征,要在现有模型框架下解决上述问题,需要引用特殊的语义特征和远距特征,但这类特征难以总结和构造,具体的解决方法还有待继续研究。

6 结束语

以上介绍了语义块切分的概念,并利用最大熵模型构造了

(下转130页)