HNC语义标注模型的构建

谢法奎[1,2] 张 全[2]
(中国科学院研究生院 北京 100039)[1] (中国科学院声学研究所 北京 100190)[2]

摘 要 介绍一种基于HNC理论的、人机结合的汉语语料语义标注模型。首先分析了HNC语义标注的内容,在此基础上定义了标注的流程。因标注十分复杂,在流程的主要环节使用机器标注来帮助人工标注。具体地说,在语义块切分问题上采用最大熵模型,其正确率和召回率分别达到了83.78%和91.17%;在句类判断问题上采用基于实例的模型,其正确率达到了51.64%。运用此标注模型建设了HNC语义标注语料库,目前语料规模已达到40万字。

关键词 概念层次网络,语料库,最大熵模型
中图法分类号 TP391 **文献标识码** A

Novel HNC Conceptual Tagging Model for Corpus

XIE Fa-kui[1,2] ZHANG Quan[2]
(Graduate School of the Chinese Academy of Sciences, Beijing 100039, China)1
(Institute of Acoustics, Chinese Academy of Sciences, Beijing 100190, China)[2]

Abstract This paper introduced a novel conceptual tagging model for corpus which is based on the Hierarchical Network of Concepts (HNC) theory, and which benefits from manual work and automatic machine. Firstly, the contents of tagging were given, and the process of tagging was defined. For the complexity of the process, some machine tagging ways were used to help manual work. A maximum entropy model was adopted to deal with the problem of semantic chunks segmentation, and the test precision and recall are 83.78 % and 91.17 %. An example based model was adopted to deal with the problem of sentence category parsing, and the test precision is 51.64 %. Relying on the model, a HNC corpus was constructed, which currently reaches 400,000 characters.

Keywords HNC, Corpus, Maximum entropy model

1 引言

语料标注是语料库建设的重要内容,对自然语言处理具有重要的意义。在不同的理论指导下,语料标注的内容自然会有很大的差异。如果从句法理论出发,标注的内容主要是词性、短语和句法树构成等。HNC(概念层次网络 Hierarchical Network of Concepts 的简称)理论有自己的特色,它深入到语义层面,建立了自然语言的概念空间。相应地,基于HNC理论的语料标注就是以这种语言概念空间为基础来描述自然语言的概念结构。我们在长期的研究和实践中逐渐形成了HNC语义标注体系,并建设了HNC语义标注语料库,为理论研究和具体工作提供服务。在理论方面,能够全面检验HNC理论和方法如HNC三有限假设;能够根据需要提供大量生语料、熟语料,帮助研究人员具体考察实际语料,进行HNC及相关的处理算法研究,如制定HNC处理策略、填写知识库等;能够通过对语言现象的统计,得到句类分布方图、句类分析难点分布方图、语句理解度分布方图等,对HNC交互引擎的发展有极大的促进作用

以下本文将介绍HNC语义标注的内容,并给出标注的流程。因标注过程十分复杂,我们构造了一种人机结合的标注模型,在流程的主要环节使用机器标注来帮助人工标注。

2 HNC语义标注的内容

语句本身存在着概念结构,HNC理论对此进行了深入的研究,形成了以句类为核心内容,以语句、语义块、句蜕、块扩等概念单元为依托的语句深层语义结构表述模式。语义块是语句的下一级语义概念构成单元;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。从形式上看,语句、语义块、句蜕、块扩这些结构,形成了一套可循环嵌套的语句结构表述体系。HNC标注就是针对这些概念单元,以符号化的方式表述其语言空间和语言概念空间信息。

具体地说,HNC标注分为语言空间标注和语言概念空间标注两部分内容:

(1)语言空间标注是指在语料文本上进行标注,给出语句的形式结构,指明各结构单元的类型与边界。具体内容包括:语句边界标注、语义块边界标注、句蜕的标注、块扩的标注、词语优先组合、主语义块分离现象的标注等。语言空间标注在

到稿日期:2008-06-25 本文受国家973项目"自然语言理解的交互引擎研究"(2004CB318104),中国科学院声学研究所"所长择优基金"(GS13SJJ04)资助。

谢法奎(1981-),硕士,主要研究方向为自然语言理解,E-mail:holinax@tom.com;张 全 研究员,博士生导师。

实质上是一种结构标注，其目标是给出自然语言对应的由概念单元组成的树形分支结构。

(2)语言概念空间标注给出各概念单元在语言概念空间的抽象语义信息，主要包括：句类、格式、语义块共享方式、语义块类型等。具体如表1所列。

表1 概念单元属性表

概念单元	属性	属性意义
语句	sc	句类
	format	格式
	connect	语义块共享方式
语义块	cat	类型
	role	角色
	red	是否含有非句蜕成分
	plug	是否为插入辅块
句蜕	cat	类型
语义块分离	cat	类型

针对以上标注内容，我们制定了HNC标注规范，它包含了一套非常详细的标注符号体系。标注实例如下：

原文：主要理由是近期一系列利好政策对股市的影响逐步增强，人民币升值的预期将刺激股市稳步上涨。

语言空间：主要理由 || 是 || {〈近期一系列利好政策 | 对股市 | 的影响〉| 逐步增强 }，{ 人民币升值的预期 | 将刺激 | [# 股市 || 稳步上涨 #] }。

语言概念空间：

jDJ#DC=[{Y401J}{X03J}]##YBC=〈! 111X0J〉[#X03BC#]=[Y4J]

3 人机结合的标注模型

根据以上介绍，我们明确了HNC语义标注的内容。与其它形式的标注相比较，HNC语义标注的一个显著特点是标注内容丰富，其目标是将句子的语言空间和语言概念空间信息全部标注出来。这从一开始便决定了HNC语义标注不可能一蹴而就，必须将它拆分为若干个子问题，有步骤有次序地进行。因而本文明确规定了标注的流程，采用一种自上而下的标注方式，先标注大的语言单位，再标注小的语言单位。具体地说，标注流程的主要步骤如下：

(1)语句切分：将一个句子切分为若干个语句。对每个语句，转(2)和(3)。

(2)语句属性标注：判断语句的句类、格式、语义块共享方式。

(3)语义块切分：将一个语句切分为若干个语义块。对每个语义块，转(4)和(5)。

(4)语义块属性标注：判断语义块的类型、角色等。

(5)语义块内部结构分析：判断语义块是否包含句蜕、块扩、语义块分离等内部结构，如果存在，指明其在语言空间的边界。对句蜕、块扩内的语言单元，返回(1)进行处理。

在明确了HNC标注的流程后，便可以开展语料标注工作。在以往的实际工作中，语料标注主要依赖于人工，这需要投入大量的人力，我们迫切希望引入机器标注来帮助人的工作。在目前的条件下，构造完全自动的机器标注模型是不切实际的，这是由HNC标注的特点决定的。HNC标注过程非常复杂，各标注环节相互关联，又相对独立，难以统一，因而如果要构造机器标注模型，那么只能针对每一个环节分别处理。即便这一点做到了，实际使用时每一个环节的机器标注错误会直接影响后面环节的机器标注操作，如此层层累积错误，在整体上难以取得理想的标注效果。

因而在实际标注中，一定的人工投入是必需的。这里关键的问题是如何寻找一个合适的人工介入点，以最少人工投入，获得最佳的整体处理效果。为此，我们提出了人机结合的标注思想，结合人工标注和机器标注。具体的标注方案是：根据标注流程，将标注任务纵向分解为多个子任务，针对各个子任务，构造机器标注模型，给出建议性解答，而后由人来检查修改。模型如图1所示。

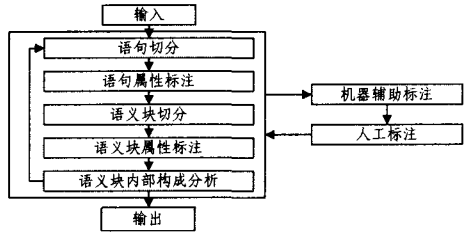


图1 HNC语义标注模型

4 机器辅助标注研究

根据以上的标注模型，标注流程包含了多个子任务：语句切分、句类判断、格式判断、语义块共享方式判断、语义块切分、语义块角色判断、语义块内部构成分析等，这些任务都需要构造相应的机器标注模型。本文主要讨论其中的两个任务：语义块切分和句类判断。之所以选择这两个任务，是因为它们在实际标注中出现频率最高，如果能够在这两个问题上有所突破，将极大地减轻用户的工作量。

机器标注模型的具体实现可以基于规则或已有的HNC研究成果，而本文主要着眼于利用已有的HNC标注语料，通过统计模型的方法来解决。具体地说，本文采用最大熵模型解决语义块切分问题，采用基于实例的方法解决句类判断问题。

4.1 基于最大熵模型的语义块切分模型

语义块切分是指把一个语句切分成若干个语义块，它是HNC语义标注的一个较为基础的子任务。我们基于最大熵模型，构造了完整的语义块切分模型，大致包括以下几部分：

(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。本系统直接采用已有的自动词法分析系统。

(2)概念分析：在HNC理论中，词语有相应的深层语义概念，概念的局部联想脉络和句类的全局联想脉络有着密切的关联，概念信息对语义块切分有重要作用。

(3)预处理：考查HNC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，改善最大熵模型的训练和测试效果。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。

模型的总体框架如图2所示。

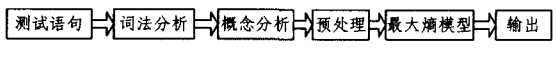


图2 语义块切分模型结构图

最大熵模型是一个比较成熟的统计模型，适合于解决分类问题，以其简洁、通用、易于移植等优点而被广泛采用。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，形式如下：

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

最大熵模型的关键在于构造特征集合，在本文中最直接的特征是词语的上下文环境。总的来说，可利用的已有信息如下：

(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，有很多值得借鉴之处。

(3)HNC 概念。HNC 概念是指词语在 HNC 概念网络节点的对应，具体来说就是概念表达式。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，前面位置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，我们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。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。本文采用了 16 种复合模板，限于篇幅，这里不详细叙述了。利用训练集中的样本数据，对所有的模板进行实例化，便得到具体的特征函数，依次循环，便得到了最大熵模型的候选特征集。例如，样本数据为"… 是/v || 宏伟/a …"，对于复合模板(W0,POS+1)，可以抽取特征信息"是 a EK_BE"，相应的特征函数如下：

$$f(x,y)=\begin{cases}1, & y\text{为“}EK\_BE\text{”，且 }w0\text{ 为"是"，}POS+1\text{ 为“}a\text{”}\\ 0, & \text{否则}\end{cases}$$

如上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和测试是沉重的负担，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。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，经过试验，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，这种方法的缺点主要是模型训练费时较多。本文采用增量特征选择方法，具体的算法描述请参见文献[1]。最终我们抽取出约 3000 个特征，作为模型的特征集。

最大熵模型的训练数据和测试数据均来源于 HNC 语义标注语料库。我们从中选取 150 篇文章约 6500 句作为训练数据，另外选取 15 篇文章约 500 句作为测试数据。测试结果如表 2 所列。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 $\beta=1$。

我们将语句分为 4 级：句子级、块扩级、原型句蜕级、要素句蜕级，虽然这 4 级语句在 HNC 语义结构上大致等同，但句类分布是不同的，在具体语言特点上也有很大出入。我们将这 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它语句更多，训练数据最充分。

表 2　语义块切分模型的测试结果

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

我们根据系统的处理结果，分析总结了典型错误。不可避免，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。具体的改进措施还有待于进一步深入研究。

4.2　基于实例的句类判断模型

句类判断顾名思义是指判断一个语句的句类。句类是语言概念空间的核心信息，句类判断是 HNC 语义标注的核心任务。我们采用基于实例的方法来解决句类判断问题。HNC 标注语料库中已包含大量熟语料，我们从中抽取各个层级的语句，作为模型的句类实例库。对一个目标语句，在实例库中经过相似度计算，求取几种与目标语句最相似的语句，以它们的句类代码作为输出。

模型的关键在于如何计算两个语句的相似度，计算方法如下：首先，如果两个语句的语义块数量不相等，则该两语句的相似度为 0，这是因为一种句类的语句的语义块数量是固定的。例外的情况是语义块省略和语义块共享的现象，对此我们通过补充或省略一个语义块来达到数量的一致。其次，如果两个语句的语义块数量相等，我们可以通过语义块来计算它们的相似度。有两个语句 S_1 和 S_2，其中 S_1 语句由 n 个语义块(C_{11}，C_{12}，…，C_{1n})组成，S_2 语句由 n 个语义块(C_{21}，C_{22}，…，C_{2n})组成。如果这两个语句的句类一致，我们要求它们的语义块一一对应，同时我们对这 n 对语义块的对位关系不做任何限制，因而一共有 $1*2*\cdots*n$ 种对位关系，我们以 A 来表示对位关系集合。在一种对位关系 $a=a_1,a_2,\cdots,a_n$ 下(其中 a_i 的取值为从 0 到 n 的整数，且互不相等)，语句 S_1 和语句 S_2 的相似度为

$$\mathrm{Sim}_a(S_1,S_2)=\prod_{i=1}^{n}\mathrm{Sim}(C_{1i},C_{2a_i}) \quad (2)$$

我们取其最大者作为两语句的相似度，即

$$\mathrm{Sim}(S_1,S_2)=\max_{a\in A}\mathrm{Sim}_a(S_1,S_2) \quad (3)$$

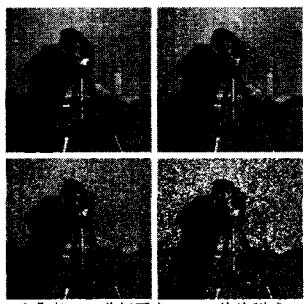
如此，语句相似度问题便转化为语义块相似度问题。对不同种类的语义块，语义块相似度计算方法是不同的。对于特征语义块，我们严格匹配特征关键词，是一个{0,1}二值函数；对于广义对象语义块，我们通过概念关联来计算，具体的计算方法请参见文献[6]。

我们从 HNC 标注语料库中选取 2 万个语句作为实例库，另选 200 个语句进行测试，测试结果如表 3 所列。

本模型并非真正的句类分析模型，目的只是通过一些经

(下转第 268 页)

当于加密的逆过程。所以，加密系统对明文的敏感性会影响到受损密文解密后的图像质量，要提高抗损伤能力，必不可免地需要适度降低系统的算法强度。本测试对密文图像分别添加了 Gauss 噪声、椒盐噪声和几何擦除等人为干扰，解密后的图像仍可在一定程度上恢复，图 8 为加密 4 轮时受损密图的解密结果。



(1)无噪声；(2) 剪切两个20×20的块测试；(3) 椒盐噪声测试(p=0.05)；(4) Gauss噪声测试($\mu=0, \sigma=0.01$)

图 8 不同干扰情况下受损密文的解密图像(加/解密轮次 n=4)

其中，Gauss 噪声对图像质量恢复的影响比较严重，它随机地改变像素的值，因此解密时需要将所有扩散作用的过程去掉方能解密；而擦除和双极性脉冲噪声(椒盐噪声)的影响要小一些，可通过滤波及增强等手段进一步还原图像。随着噪声强度和损失面积的增大，受损图像的恢复能力明显下降。需要说明的是，如果算法中使用了灰度扩散操作，那么受损图像将难以恢复。使用时应视不同应用场合，在算法强度和图像恢复能力之间加以权衡。

结束语 对经典 Standard 映射的混沌特性和动态 S 盒密码学特性进行了详细的分析，并根据所得到的结论，提出了一种结合混沌映射和动态 S 盒的 Feistel 网络结构图像加密算法。该算法交替地使用置乱扩散操作进行多轮次的加密。相对于单一混沌映射的加密方法，此算法具有更高的密码强度，且解密算法与加密算法具有相同的结构，不需另行设计。此外，由于采用了 Feistel 型的网络结构，本算法可同时完成对两块明文图像(即 L_0 和 R_0)的加密，这有助提高系统吞吐率，符合图像加密的特点，而且引入动态 S 盒非线性运算进一步提高了算法的安全性。最后的实验和安全性分析结果表明：该加密算法具有很大的密钥空间，对密钥十分敏感，对多种攻击手段都具有较好的免疫性，适用于软件加密系统，而且很容易移植到硬件平台上，具有良好的应用前景。

参 考 文 献

[1] 彭军，张伟，杨治明，等. 一种基于 Feistel 网络的反馈式分组混沌密码的研究[J]. 计算机科学，2006，33(1)

[2] Fridrich J. Symmetric ciphers based on two. dimensional chaotie maps[J]. IntJ Bifureatlon and Chaos，1998，8(6)：1259

[3] Wong K-W，Kwok BS-H，Law W-S. A fast image encryption scheme based on chaotic standard map[J]. Physics Letters A Prn：22/12/2007

[4] Chen Guo，Chen Yong，Liao Xiaofeng. An extended method for obtaining S-boxes based on three-dimensional chaotic Baker maps[J]. Chaos，Solitons and Fractals，2007，31：571-579

[5] Xiao Di，Liao Xiaofeng，Wong K W. An efficient entire chaos-based scheme for deniable authentication[J]. Chaos，Solitons and Fractals，2005，23：1327-1331

[6] Xiang Tao，Wong K-W，Liao Xiaofeng. A Novel Symmetrical Cryptosystem based on Discretized Two-dimensional Chaotic Map[J]. Physics Letters A，2007，364(3/4)：252-258

[7] Pareek N K，Vinod Patidar K K. Sud Image encryption using chaotic logistic map[J]. Image and Vision Computing，2006，24，926-934

[8] Zhang Linhua，Liao Xiaofeng，Wang Xuebing. An image encryption approach based on chaotic maps[J]. Chaos，Solitons and Fractals，2005，24：759-765

[9] Xiang Tao，Liao Xiao feng，Tang Guoping，et al. A novel block cryptosystem based on iterating a chaotic map[J]. Physics Letters A，2006，349：109-115

[10] Chen Guanrong，Mao Yaobin，Chui C K. A symmetric image encryption scheme based on 3D chaotic cat maps[J]. Chaos，Solitons and Fractals，2004，21：749-761

(上接第 240 页)

验的方法获得一个可能的估计值，减少标注人员的工作，因而可以返回多个结果以供标注人员参考。由测试结果来看，取 $N=3$ 最为合适。

表 3 句类判断模型的测试结果

返回结果的数量 N	正确率
1	34.74 %
2	45.54 %
3	51.64 %
4	53.05 %
5	54.46 %

语料库的数据还在不断扩大之中，经过测试，目前模型正确率的上限是 75%左右，随着语料库建设不断扩大，这一问题会逐渐得到解决。

结束语 在标注规范指导下，应用人机结合的标注模型，我们建设了句子级的 HNC 语义标注语料库，目前语料库规模已达到 40 万字。我们开发了完整的语料库管理加工应用系统，为语料标注人员的工作搭建了平台。未来的工作包括以下几方面：进一步增加语料库的规模，为统计模型提供更充分的数据；目前已有的机器辅助标注模型在标注准确率上需要提高，尤其是句类判断模型；在更多的标注子任务上实现机器辅助标注。

参 考 文 献

[1] Berger A L，Pietra S A D，Pietra V J D. A maximum entropy approach to natural language processing[J]. Computational Linguistics，1996，22(1)：1-36

[2] Darroch J N，Ratcliff D. Generalized iterative scaling for log-linear models[J]. The Annals of Mathematical Statistics，1972，43(5)：1470-1480

[3] 黄曾阳. HNC 理论概要[J]. 中文信息学报，1997(4)：11-20

[4] 李素建，刘群，杨志峰. 基于最大熵模型的组块分析[J]. 计算机学报，2003，26(12)：1722-1727

[5] 周雅倩，郭以昆，黄萱菁，等. 基于最大熵模型的中英文基本名词短语识别[J]. 计算机研究与发展，2003，40(3)：440-446

[6] 张运良，张全. 基于 HNC 理论的语义相关度计算方法[J]. 计算机工程与应用，2005，34(41)：1-3