(19)中华人民共和国国家知识产权局

(12)发明专利申请

(10)申请公布号 CN 112836506 A
(43)申请公布日 2021.05.25

(21)申请号 202110206745.3

(22)申请日 2021.02.24

(71)申请人 中国人民解放军国防科技大学
地址 410073 湖南省长沙市开福区德雅路109号

(72)发明人 魏急波 赵海涛 张亦弛 熊俊 张姣

(74)专利代理机构 长沙国科天河知识产权代理有限公司 43225
代理人 邱轶

(51)Int.Cl.
G06F 40/284(2020.01)
G06F 40/30(2020.01)
G06N 3/04(2006.01)

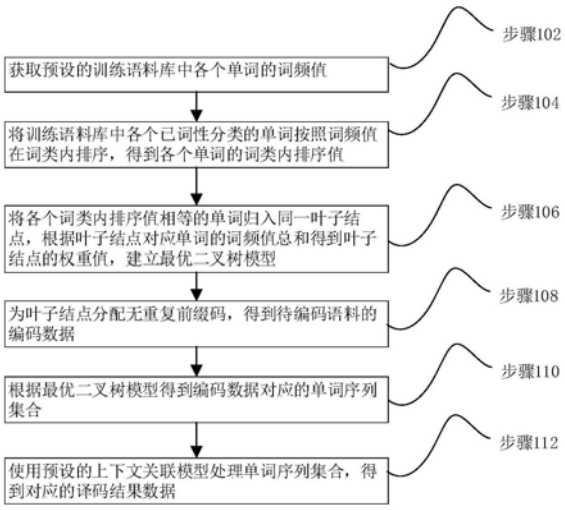
权利要求书2页 说明书11页 附图6页

(54)发明名称

一种基于上下文语义的信源编译码方法和装置

(57)摘要

本申请涉及一种基于上下文语义的信源编译码方法和装置。所述方法包括:在编码端,根据词频对训练语料库的单词按照词类进行分别排序,将不同词类中排序相同的单词合并为一个叶子结点,并根据一个叶子结点中所有单词的词频总和得到该叶子结点的权重值,生成最优二叉树模型,为叶子结点分配无重复前缀码。根据叶子结点的无重复前缀码得到语料的编码数据。在译码端,根据编码数据在二叉树模型中得到对应的候选单词序列集合,根据上下文关联关系得到上下文关系最紧密的单词序列作为译码结果。本申请在编码过程中加入了语义维度,译码时利用上下文的语义关联得到最优译码结果,能够实现高效的语义信息表达传输能力,并节省传输开销。



CN 112836506 A

1.一种基于上下文语义的信源编译码方法,其特征在于,所述方法包括:

在编码端:

获取预设的训练语料库中各个单词的词频值;

将所述训练语料库中各个已词性分类的单词按照所述词频值在词类内排序,得到各个单词的词类内排序值;

将各个所述词类内排序值相等的单词归入同一叶子结点,根据所述叶子结点对应单词的词频值总和得到所述叶子结点的权重值,建立最优二叉树模型;

为所述叶子结点分配无重复前缀码,得到待编码语料的编码数据;

在译码端:

根据所述最优二叉树模型得到所述编码数据对应的单词序列集合;

使用预设的上下文关联模型处理所述单词序列集合,得到对应的译码结果数据。

2.根据权利要求1所述的方法,其特征在于,将所述训练语料库中各个已词性分类的单词按照所述词频值在词类内排序,得到各个单词的词类内排序值的步骤包括:

将所述训练语料库中的单词按照词性进行分类,得到对应的词性分类;所述词性分类包括名词分类、动词分类、形容词分类、副词分类和连词分类;

在所述词性分类中,按照所述词频值由高到低的顺序,得到各个单词的词类内排序值。

3.根据权利要求1所述的方法,其特征在于,建立最优二叉树模型的方式包括:

获取当前权重值最低的第一叶子结点和第二叶子结点,合并所述第一叶子结点和所述第二叶子结点得到第三叶子结点;

根据所述第一叶子结点和所述第二叶子结点的权重值的和,得到所述第三叶子结点的权重值。

4.根据权利要求3所述的方法,其特征在于,为所述叶子结点分配无重复前缀码的方式包括:

比较所述第一叶子结点和所述第二叶子结点的权重值,根据比较结果分别得到所述第一叶子结点和所述第二叶子结点的标签值;

根据所述最优二叉树模型中从根结点到所述第一叶子结点经历的所有叶子结点的标签值序列,得到所述第一叶子结点的无重复前缀码。

5.根据权利要求1所述的方法,其特征在于,使用预设的上下文关联模型处理所述单词序列集合,得到对应的译码结果数据的方式包括:

获取所述训练语料库中各个单词间的上下文语义关联特征;

根据上下文语义关联特征得到联合出现概率值最高的单词序列作为译码结果数据。

6.根据权利要求5所述的方法,其特征在于,获取所述训练语料库中单词间的上下文语义关联特征的方式包括:

使用基于LSTM的神经网络模型学习所述训练语料库中各个单词间的上下文语义关联特征。

7.根据权利要求6所述的方法,其特征在于,得到联合出现概率值最高的单词序列的方式包括:

使用N-gram模型对所述单词序列集合中单词序列的联合概率分布建模;

当所述单词序列的长度小于上下文窗口的预设值时,根据所述N-gram模型,使用枚举

法得到联合出现概率值最高的单词序列；

当所述单词序列的长度大于上下文窗口的预设值时，根据所述N-gram模型，使用状态压缩动态规划算法得到联合出现概率值最高的单词序列。

8.一种基于上下文语义的信源编译码装置，其特征在于，所述装置包括：

编码模块，用于获取预设的训练语料库中各个单词的词频值，将所述训练语料库中各个已词性分类的单词按照所述词频值在词类内排序，得到各个单词的词类内排序值，将各个所述词类内排序值相等的单词归入同一叶子结点，根据所述叶子结点对应单词的词频值总和得到所述叶子结点的权重值，建立最优二叉树模型，为所述叶子结点分配无重复前缀码，得到待编码语料的编码数据；

译码模块，用于根据所述最优二叉树模型得到所述编码数据对应的单词序列集合，使用预设的上下文关联模型处理所述单词序列集合，得到对应的译码结果数据。

9.一种计算机设备，包括存储器和处理器，所述存储器存储有计算机程序，其特征在于，所述处理器执行所述计算机程序时实现权利要求1至7中任一项所述方法的步骤。

10.一种计算机可读存储介质，其上存储有计算机程序，其特征在于，所述计算机程序被处理器执行时实现权利要求1至7中任一项所述方法的步骤。

一种基于上下文语义的信源编译码方法和装置

技术领域

[0001] 本申请涉及智能体语义通信技术领域，特别是涉及一种基于上下文语义的信源编译码方法和装置。

背景技术

[0002] 随着通信设备的智能化水平和对外界的认知能力不断增强，智能体语义通信已经成为通信领域的一大研究趋势。语义通信的核心在于准确传输数据含义或内容通信，而非以通信符号的准确传输为目标。

[0003] 目前在发送端根据通信目的和历史通信过程的理解分析对需要传输的数据进行处理，从而源头上避免大量冗余的传输。在接收端根据上下文、先验信息以及个体通信的目的等知识，对接收信号进行智能纠错和恢复。但语义通信中还存在许多有待解决的基础问题，例如如何通过有效的编码方式实现高效的语义信息表达等。

发明内容

[0004] 基于此，有必要针对上述技术问题，提供一种基于上下文语义的信源编译码方法和装置。

[0005] 一种基于上下文语义的信源编译码方法，所述方法包括：

[0006] 在编码端：

[0007] 获取预设的训练语料库中各个单词的词频值。

[0008] 将训练语料库中各个已词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值。

[0009] 将各个词类内排序值相等的单词归入同一叶子结点，根据叶子结点对应单词的词频值总和得到叶子结点的权重值，建立最优二叉树模型。

[0010] 为叶子结点分配无重复前缀码，得到待编码语料的编码数据。

[0011] 在译码端：

[0012] 根据最优二叉树模型得到编码数据对应的单词序列集合。

[0013] 使用预设的上下文关联模型处理单词序列集合，得到对应的译码结果数据。

[0014] 其中一个实施例中，将训练语料库中各个已词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值的步骤包括：

[0015] 将训练语料库中的单词按照词性进行分类，得到对应的词性分类。词性分类包括名词分类、动词分类、形容词分类、副词分类和连词分类。

[0016] 在词性分类内，按照词频值由高到低的顺序，得到对应的单词序列，根据单词序列，得到各个单词的词类内排序值。

[0017] 其中一个实施例中，建立最优二叉树模型的方式包括：

[0018] 获取当前权重值最低的第一叶子结点和第二叶子结点，合并第一叶子结点和第二叶子结点得到第三叶子结点。

[0019] 根据第一叶子结点和第二叶子结点的权重值的和,得到第三叶子结点的权重值。

[0020] 其中一个实施例中,为叶子结点分配无重复前缀码的方式包括:

[0021] 比较第一叶子结点和第二叶子结点的权重值,根据比较结果分别得到第一叶子结点和第二叶子结点的标签值。

[0022] 根据最优二叉树模型中从根结点到第一叶子结点经历的所有叶子结点的标签值序列,得到第一叶子结点的无重复前缀码。

[0023] 其中一个实施例中,使用预设的上下文关联模型处理单词序列集合,得到对应的译码结果数据的方式包括:

[0024] 获取训练语料库中各个单词间的上下文语义关联特征。

[0025] 根据上下文语义关联特征,从单词序列集合中得到联合出现概率值最高的单词序列,得到对应的译码结果数据。

[0026] 其中一个实施例中,获取训练语料库中各个单词间的上下文语义关联特征的方式包括:

[0027] 使用基于LSTM的神经网络模型学习训练语料库中各个单词间的上下文语义关联特征。

[0028] 其中一个实施例中,从单词序列集合中得到联合出现概率值最高的单词序列的方式包括:

[0029] 使用N-gram模型对单词序列集合中单词序列的联合概率分布建模。

[0030] 当单词序列的长度小于上下文窗口的预设值时,根据N-gram模型,使用枚举法得到联合出现概率值最高的单词序列。

[0031] 当单词序列的长度大于上下文窗口的预设值时,根据N-gram模型,使用状态压缩动态规划算法得到联合出现概率值最高的单词序列。

[0032] 一种基于上下文语义的信源编译码装置,包括:

[0033] 编码模块,用于获取预设的训练语料库中各个单词的词频值,将训练语料库中各个已词性分类的单词在分类中按照词频值在词类内排序,得到各个单词的词类内排序值,将各个词类内排序值相等的单词归入同一叶子结点,根据叶子结点对应单词的词频值总和得到叶子结点的权重值,建立最优二叉树模型,为叶子结点分配无重复前缀码,得到待编码语料的编码数据。

[0034] 译码模块,用于根据最优二叉树模型得到编码数据对应的单词序列集合,使用预设的上下文关联模型处理单词序列集合,得到对应的译码结果数据。

[0035] 一种计算机设备,包括存储器和处理器,所述存储器存储有计算机程序,所述处理器执行所述计算机程序时实现以下步骤:

[0036] 在编码端:

[0037] 获取预设的训练语料库中各个单词的词频值。

[0038] 将训练语料库中各个已词性分类的单词在分类中按照词频值在词类内排序,得到各个单词的词类内排序值。

[0039] 将各个词类内排序值相等的单词归入同一叶子结点,根据叶子结点对应单词的词频值总和得到叶子结点的权重值,建立最优二叉树模型。

[0040] 为叶子结点分配无重复前缀码,得到待编码语料的编码数据。

[0041] 在译码端：

[0042] 根据最优二叉树模型得到编码数据对应的单词序列集合。

[0043] 使用预设的上下文关联模型处理单词序列集合，得到对应的译码结果数据。

[0044] 一种计算机可读存储介质，其上存储有计算机程序，所述计算机程序被处理器执行时实现以下步骤：

[0045] 在编码端：

[0046] 获取预设的训练语料库中各个单词的词频值。

[0047] 将训练语料库中各个已词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值。

[0048] 将各个词类内排序值相等的单词归入同一叶子结点，根据叶子结点对应单词的词频值总和得到叶子结点的权重值，建立最优二叉树模型。

[0049] 为叶子结点分配无重复前缀码，得到待编码语料的编码数据。

[0050] 在译码端：

[0051] 根据最优二叉树模型得到编码数据对应的单词序列集合。

[0052] 使用预设的上下文关联模型处理单词序列集合，得到对应的译码结果数据。

[0053] 与现有技术相比，上述一种基于上下文语义的信源编译码方法、装置、计算机设备和存储介质，在编码端，根据单词词频对训练语料库的单词在词类内进行排序，将各个词类中排序相同的单词合并为一个叶子结点，并根据一个叶子结点中所有单词的词频总和得到该叶子结点的权重值，生成最优二叉树模型，并为各个叶子结点分配无重复前缀码。根据最优二叉树模型中各叶子结点的无重复前缀码，得到待编码语料的编码数据；在译码端，根据编码数据在二叉树模型中得到对应的候选单词集合，根据上下文关联关系，从候选单词集合中得到对应的结果作为译码结果。本申请通过对单词进行分类排序，在编译码过程中加入了语义维度，在译码时利用上下文的语义关联得到最优译码结果，能够实现高效的语义信息表达传输和信息恢复能力，并节省传输开销。

附图说明

[0054] 图1为一个实施例中一种基于上下文语义的信源编译码方法的步骤图；

[0055] 图2为一个实施例中一种基于上下文语义的信源编译码方法中译码端的数据处理流程示意图；

[0056] 图3为一个实施例中基于LSTM神经网络的结构示意图；

[0057] 图4为一个实施例中使用N-gram模型计算单词序列联合概率值的流程示意图；

[0058] 图5为本申请提供的一种基于上下文语义的信源编译码方法和Huffman编码方法的性能曲线图；

[0059] 图6为本申请提供的一种基于上下文语义的信源编译码方法在上下文窗口为3时的性能曲线图；

[0060] 图7为本申请提供的一种基于上下文语义的信源编译码方法在上下文窗口为4时的性能曲线图；

[0061] 图8为本申请提供的一种基于上下文语义的信源编译码方法在上下文窗口为5时的性能曲线图；

[0062] 图9为一个实施例中计算机设备的内部结构图。

具体实施方式

[0063] 为了使本申请的目的、技术方案及优点更加清楚明白，以下结合附图及实施例，对本申请进行进一步详细说明。应当理解，此处描述的具体实施例仅仅用以解释本申请，并不用于限定本申请。

[0064] 在一个实施例中，如图1所示，提供了一种基于上下文语义的信源编译码方法，用于编码端和译码端。

[0065] 在编码端包括以下步骤：

[0066] 步骤102，获取预设的训练语料库中各个单词的词频值。

[0067] 步骤104，将训练语料库中各个已词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值。

[0068] 具体地，已词性分类的单词是指训练语料库中的所有单词($\mathcal{N}$个)已经按其词性标记分成$\mathcal{P}$个词性分类，具体包括名词类η、动词类v、形容词类a等。然后，将各个词性分类中的单词按照其词频值由高到低降序排列，得到各个单词在对应词性分类中的词类内排序值。

[0069] 步骤106，将各个词类内排序值相等的单词归入同一叶子结点，根据叶子结点对应单词的词频值总和得到叶子结点的权重值，建立最优二叉树模型。

[0070] 将各个词性分类的单词序列汇中处于同一位置(即具有相等词类内排序值)的单词合并，归入同一叶子结点。具体地，各个词性分类中词频值最高的单词组成一个叶子结点$A_0=(\eta_0,v_0,a_0,...)$，次高的组成一个叶子$A_1=(\eta_1,v_1,a_1,...)$，以此类推得到M个叶子结点$A_i=(\eta_i,v_i,a_i,...)$，i＝0,...,M-1，$M\leq\mathcal{N}$。每个叶子结点的权重为该叶子结点所含的所有单词的词频值的和。根据得到的叶子结点建立最优二叉树模型。

[0071] 进一步地，建立最优二叉树模型的方式包括：获取当前权重值最低的第一叶子结点和第二叶子结点，合并第一叶子结点和第二叶子结点得到第三叶子结点。根据第一叶子结点和第二叶子结点的权重值的和，得到第三叶子结点的权重值。

[0072] 步骤108，为叶子结点分配无重复前缀码，得到待编码语料的编码数据。

[0073] 无重复前缀码是指，根据最优二叉树模型中各个叶子结点的无重复前缀码，可以得到各个叶子结点对应的唯一编码数据。

[0074] 举例来说，将训练语料库中的所有$\mathcal{N}$个单词按其词性标记分成$\mathcal{P}=4$个大类，即名词类η、动词类v、形容词类a和其他类o。设在训练语料库中名词类的‘time’(出现1597次)，动词类的‘is’(10108次)，形容词类的‘new’(1635次)和其他类的‘the’(69968次)分别在四类中出现频次最多，得到叶子结点A_0为{time,is,new,the}，且其权重为四个词的频次相加之和83308。依次类推得到所有叶子结节点，每次取权重最低的两个叶子结点合并生成新的叶子结点，以两个叶子结点的权重的求和值得到新的叶子结点的权重，自底向上构建最优二叉树。同时，根据需要合并的两个叶子结点的权重值大小，给这两个叶子结点分别设置标签‘1’和‘0’，直到给最优二叉树的M个叶子结点都分配了码字，该码字是从根结点到该叶子结点的标签的序列，获得的编码就是该叶子结点的无重复前缀码。

[0075] 在译码端包括以下步骤：

[0076] 步骤110，根据最优二叉树模型得到编码数据对应的单词序列集合。

[0077] 步骤112，使用预设的上下文关联模型处理单词序列集合，得到对应的译码结果数据。

[0078] 译码端收到的是一组编码，根据编码可以在最优二叉树模型中得到对应的叶子结点。由于每一个叶子结点对应于一组单词，因此每一个无重复前缀码可以得到对应的单词集合。由于表达语义时，上下文单词之间是存在上下文关联的。因此使用上下文关联模型可以得到上下文单词同时出现联合概率最高作为对应的译码结果。

[0079] 本实施例通过对单词进行分类排序，在编译码过程中加入了语义维度，在译码时利用上下文的语义关联作为先验知识优化码字的分配和实现智能信息恢复，从对应的单词序列集合中得到最优译码结果，能够实现高效的语义信息表达传输能力，并节省传输开销。

[0080] 其中一个实施例中，为叶子结点分配无重复前缀码的方式包括：

[0081] 比较第一叶子结点和第二叶子结点的权重值，根据比较结果分别得到第一叶子结点和第二叶子结点的标签值。

[0082] 根据最优二叉树模型中从根结点到第一叶子结点经历的所有叶子结点的标签值序列，得到第一叶子结点的无重复前缀码。

[0083] 具体地，比较要合并两个叶子结点的权重，将权重值较高的叶子结点的标签值设为1，将权重值较低的叶子结点的标签值设为0。迭代合并过程直到最后只剩下两个叶子结点合并形成根结点，在这一过程中为最优二叉树中所有的叶子结点设置标签值。根据从根结点到某个叶子结点的路径上的所有叶子结点的标签值序列，就可以得到该叶子结点的无重复前缀码。标签值的设置方式还可以根据编码需要调整，只要确保能够区分两个需要合并的叶子结点。

[0084] 本实施例基于最优二叉树的生成过程提供了一种简单的无重复前缀码分配方式，具有编码方式简单、容易实现的特点。

[0085] 其中一个实施例中，如图2所示，利用N-gram模型和基于多层LSTM神经网络模型分别刻画和学习上下文间的相关性，并利用状态压缩动态规划方法将一系列相邻的单词作为上下文一起译码，针对一个编码对应多个单词的情况得到全局最优解。本实施例中从单词序列集合中得到联合出现概率值最高的单词序列的方式包括：

[0086] 步骤202，使用基于LSTM的神经网络模型学习训练语料库中各个单词间的上下文语义关联特征。

[0087] 步骤204，使用N-gram模型建模单词序列集合中单词序列的联合概率分布。

[0088] 具体地，对应$\mathcal{P}$个词性分类，一个无重复前缀码最多对应$\mathcal{P}$个单词，所以一个长度为n的单词序列s，最多可有$n^{\mathcal{P}}$种排列组合。计算每种排列组合的概率值$P(w_1,w_2,\ldots,w_n)$。本实施例利用N-gram模型对联合概率$Pr(w_1w_2\ldots w_n)$进行建模，其过程为一个马尔科夫链$Pr(w_1w_2\ldots w_n)=Pr(w_1)Pr(w_2|w_1)\ldots Pr(w_n|w_{n-1}\ldots w_2w_1)$，即$\Pr(s)=\prod_{i=1}^{n}\Pr(w_i|w_1w_2\ldots w_{i-1})$，式中每个单词的出现与前面的历史字符是相关的。然而，随着单词出现位置之间的距离增加，距离越远的两个单词出现概率的相关性逐渐降低。因此，

其马尔可夫假设是单词序列中的每个字符仅与前面N个历史字符是相关的，则可将联合概率公式简化为$\Pr(s) \approx \prod_{i=1}^{n} \Pr(w_i \mid w_{i-N} \ldots w_{i-1})$。

[0089] 其中，上下文语义关联特征$\Pr(w_i \mid w_{i-N} \ldots w_{i-1})$可由深度网络学习出来。如图3所示，本实施例使用多层LSTM网络，包括LSTM层I(256个节点)和II(256个节点)，Dense层I(256个节点，非线性激活函数为Relu)和Dense层II(词库中单词的数量为该层的节点数，非线性激活函数为Softmax)组成。多层LSTM神经网络输入为需要预测的中心词的周围几个词的one-hot向量(one-hot向量为“一位有效”编码，即对$\mathcal{N}$个状态进行编码，每个状态都是独一无二的。在任意时候，只有一位有效(取1)，其余位置取0的$\mathcal{N}$维向量。)$w_{Input} = [w_{i-\lceil L/2 \rceil} \ldots w_{i-1}, w_{i+1} \ldots w_{i+\lceil L/2 \rceil - 1}]$，多层LSTM神经网络的输出为预测目标函数的one-hot向量$w_{0nput} = w_i$。它的原理是利用中心词的前后L个上下文单词来预测中心词。通过梯度下降方法使训练中多层LSTM网络的损失函数最小化(损失函数为$E = -\log \Pr(w_{0utput} \mid w_{Input})$)，此时网络输出层即所求的基于上下文推导中心词的概率。多层LSTM网络利用中心词的前后L个词$w_{Input} = [w_{i-\lceil L/2 \rceil} \ldots w_{i-1}, w_{i+1} \ldots w_{i+\lceil L/2 \rceil - 1}]$来预测中心词$w_{0nput} = w_i$。网络输出层的激活函数为Softmax函数，该函数将多个神经元的输出映射到(0,1)区间，输出值即为所求概率。网络通过梯度下降方法训练使多层LSTM网络的损失函数$E = -\log \Pr(w_{0utput} \mid w_{Input})$最小化。

[0090] 步骤206，如图4所示，当单词序列的长度小于上下文窗口的预设值时，根据N-gram模型，使用枚举法得到联合出现概率值最高的单词序列。

[0091] 设一个上下文窗口的大小为N，$N \in Z^+$。对于一个长度为n的单词序列$s = (w_1, w_2, \ldots, w_n)$，$n \in Z^+$，当$n \leqslant N$时，用枚举算法在$n^{\mathcal{P}}$种排列组合S中找到上下文相关性最强的一种s*作为译码结果，即$s^* = \operatorname{argmax}_{s \in S} \Pr(s)$。当$\Pr(w_1 w_2 \ldots w_n)$概率值的最大值为P时，即所对应的序列作为译码结果，该过程又可以表述为$P[(w_1^k \ldots w_n^k)] \triangleq \max \sum_{i=1}^{n} \ln \Pr(w_i \mid w_1 \ldots w_{i-1})$。

[0092] 步骤208，当单词序列的长度n大于上下文窗口大小预设值N时，根据N-gram模型，使用状态压缩动态规划算法得到联合出现概率值最高的单词序列，该状态转移过程又可以表述为：

[0093] $$P[\mathcal{S}_i(k^1 \ldots k^N)] \overset{\text{def}}{=} \max_{s \in \mathcal{S}_i(k^1 \ldots k^N)} \ln \Pr(\mathbf{s}) = \max_{w_{i-N}^{l}} \{P[\mathcal{S}_{i-1}(l\, k^1 \ldots k^{N-1})] + \ln \Pr(w_i^{k^N} \mid w_{i-N}^{l} \ldots w_{i-1}^{k^{N-1}})\}$$

[0094] 当单词序列的长度大于上下文窗口大小预设值n＞N时，用状态压缩动态规划算法先求解最小子问题的解，即单词序列中前N个单词组合概率值，再逐渐增加子问题的规模，即逐渐考虑前N+1个字词的全局最优组合，其过程为：

$$P[\mathcal{S}_{N+1}(k^1...k^N)] \overset{\text{def}}{=} \max_{\mathbf{s}\in\mathcal{S}_{N+1}(k^1...k^N)} \ln \Pr(\mathbf{s}) = \max_{(w_1...w_{N+1})} \sum_{i=1}^{N+1} \ln \Pr(w_i \mid w_{i-N}...w_{i-1})$$

[0095]
$$= \max_{w_1^l} \{\sum_{i=1}^{N} \ln \Pr(w_i^{k^{i-1}} \mid w_1^l...w_{i-1}^{k^{i-2}}) + \ln \Pr(w_{N+1}^{k^N} \mid w_1^l...w_N^{k^{N-1}})\}$$
$$= \max_{w_1^l} \{P[\mathcal{S}_N(l\, k^1...k^{N-1})] + \ln \Pr(w_{N+1}^{k^N} \mid w_1^l...w_N^{k^{N-1}})\}$$

[0096] 以此类推,前N+2个字词的全局最优组合直到子问题i考虑前i个字词的全局最优组合,一直到求解出序列n个字词的全局最优解,具体过程为:

[0097] (1)首先计算最小子问题(即i=N)的所有概率值,记为$P[(w_1^k...w_N^k)]$,此时概率值的计算公式为$P[(w_1^k...w_N^k)] \triangleq \max \sum_{i=1}^{N} \ln \Pr(w_i \mid w_{i-N+1}...w_{i-1})$。

[0098] (2)递归求解(i>N)每个子问题时,需要用到上一个子问题i-1多个最优子序列的概率值。即第i个子问题的状态的最优概率值$P[S_i(k^1...k^N)]$等于选择第i-N位的单词w_{i-N}^l,使相对应子问题i-1的最优概率值$P[S_{i-1}(l\ k^1...k^{N-1})]$和由上文前N个单词$(w_{i-N}^l w_{i-N+1}^{k^1}...w_{i-1}^{k^{N-1}})$推出下个词为$w_i^{k^N}$的概率值之和的最大值,即状态转移公式为

[0099] $$P[\mathcal{S}_i(k^1...k^N)] \overset{\text{def}}{=} \max_{w_{i-N}^l} \{P[\mathcal{S}_{i-1}(l\, k^1...k^{N-1})] + \ln \Pr(w_i^{k^N} \mid w_{i-N}^l...w_{i-1}^{k^{N-1}})\}$$

[0100] 进一步地,当马尔可夫假设单词序列中的每个字符仅与前面N个历史字符是相关时,单词序列联合概率公式建模可以表示为$\Pr(\mathbf{s}) = \prod_{i=1}^{n} \Pr(w_i \mid w_{i-N+1}...w_{i-1})$,最小子问题的最大概率值可以表示为$P[(w_1^k...w_N^k)] \triangleq \max \sum_{i=1}^{N} \ln \Pr(w_i \mid w_{i-N+1}...w_{i-1})$,第i子问题的最大概率值$P[S_i(k^1...k^N)] \triangleq \max \sum_{j=1}^{i} \ln \Pr(w_j \mid w_{j-N}...w_{j-1})$可以写成

$P[S_i(k^1...k^N)] \triangleq \max_{w_{i-N}^l} \{P[S_{i-1}(l\, k^1...k^{N-1})] + \ln \Pr(w_i^{k^N} \mid w_{i-N}^l w_{i-N+1}^{k^1}...w_{i-1}^{k^{N-1}})\}$,即第i子问题的最大概率值可以分解成多个更小的子问题,并且其中子问题是重叠子问题,利用状态压缩动态规划算法将子问题的多个最优结果存储在一个表格中,可以避免重复计算,降低从大量的潜在组合中寻找全局最优解的时间复杂度。

[0101] 为说明本申请技术效果,基于上述一个实施例中提供的方法对Brown字词库进行测试。按照词性将Brown字词库分成四类,具体包括30632个名词、10392个动词、8054个形容词和4331个其他类的字词。相比于对每个单词进行编码,需要的编码数量从53409下降到30632个。图5为本申请中方法和Huffman编码方法的性能比较,可以看到对同一个语料库中字词进行编码时,本申请的编码方法会比Huffman编码的动态平均码长长度更短,而且差距会随着需要编码的字符数的增加而增大,验证了算法的有效性。

[0102] 图6至图8分别为本申请方法的上下文窗口长度为3、4、5时的仿真结果。可以看到,

当上下文窗口大小等于或大于基于LSTM神经网络学习时的特征窗口大小时，本申请提供的方法的语义相似度可以达到峰值并保持稳定。随着上下文窗口的增加，语义相似度得分会增加；随着特征窗口大小的增加，语义相似度得分也会提高。

[0103] 应该理解的是，虽然图1的流程图中的各个步骤按照箭头的指示依次显示，但是这些步骤并不是必然按照箭头指示的顺序依次执行。除非本文中有明确的说明，这些步骤的执行并没有严格的顺序限制，这些步骤可以以其它的顺序执行。而且，图1中的至少一部分步骤可以包括多个子步骤或者多个阶段，这些子步骤或者阶段并不必然是在同一时刻执行完成，而是可以在不同的时刻执行，这些子步骤或者阶段的执行顺序也不必然是依次进行，而是可以与其它步骤或者其它步骤的子步骤或者阶段的至少一部分轮流或者交替地执行。

[0104] 在一个实施例中，提供了一种基于上下文语义的信源编译码装置，包括：

[0105] 编码模块，用于获取预设的训练语料库中各个单词的词频值，将训练语料库中各个词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值，将各个词类内排序值相等的单词归入同一叶子结点，根据叶子结点对应单词的词频值总和得到叶子结点的权重值，建立最优二叉树模型，为叶子结点分配无重复前缀码，得到待编码语料的编码数据。

[0106] 译码模块，用于根据最优二叉树模型得到编码数据对应的单词序列集合，使用预设的上下文关联模型得到对应的译码结果数据。

[0107] 其中一个实施例中，编码模块用于将训练语料库中的单词按照词性进行分类，得到对应的词性分类。词性分类包括名词分类、动词分类、形容词分类、副词分类和连词分类。在词性分类内，按照词频值由高到低的顺序，得到各个单词的词类内排序值。

[0108] 其中一个实施例中，编码模块用于获取当前权重值最低的第一叶子结点和第二叶子结点，合并第一叶子结点和第二叶子结点得到第三叶子结点。根据第一叶子结点和第二叶子结点的权重值的和，得到第三叶子结点的权重值。

[0109] 其中一个实施例中，编码模块用于比较第一叶子结点和第二叶子结点的权重值，根据比较结果分别得到第一叶子结点和第二叶子结点的标签值。根据最优二叉树模型中从根结点到第一叶子结点经历的所有叶子结点的标签值序列，得到第一叶子结点的无重复前缀码。

[0110] 其中一个实施例中，译码模块用于获取训练语料库中各个单词间的上下文语义关联特征。根据上下文语义关联特征得到联合出现概率值最高的单词序列，得到对应的译码结果数据。

[0111] 其中一个实施例中，译码模块用于使用基于LSTM神经网络模型学习训练语料库中各个单词间的上下文语义关联特征。

[0112] 其中一个实施例中，编码模块用于使用N-gram模型对上下文单词序列的联合概率分布建模。当单词序列的长度小于上下文窗口的预设值时，根据N-gram模型，使用枚举法得到联合出现概率值最高的单词序列。当单词序列的长度大于上下文窗口的预设值时，根据N-gram模型，使用状态压缩动态规划算法得到联合出现概率值最高的单词序列。

[0113] 关于一种基于上下文语义的信源编译码装置的具体限定可以参见上文中对于一种基于上下文语义的信源编译码方法的限定，在此不再赘述。上述一种基于上下文语义的信源编译码装置中的各个模块可全部或部分通过软件、硬件及其组合来实现。上述各模块

可以硬件形式内嵌于或独立于计算机设备中的处理器中，也可以以软件形式存储于计算机设备中的存储器中，以便于处理器调用执行以上各个模块对应的操作。

[0114] 在一个实施例中，提供了一种计算机设备，该计算机设备可以是服务器，其内部结构图可以如图9所示。该计算机设备包括通过系统总线连接的处理器、存储器、网络接口和数据库。其中，该计算机设备的处理器用于提供计算和控制能力。该计算机设备的存储器包括非易失性存储介质、内存储器。该非易失性存储介质存储有操作系统、计算机程序和数据库。该内存储器为非易失性存储介质中的操作系统和计算机程序的运行提供环境。该计算机设备的数据库用于存储最优二叉树模型、上下文关联模型数据。该计算机设备的网络接口用于与外部的终端通过网络连接通信。该计算机程序被处理器执行时以实现一种基于上下文语义的信源编译码方法。

[0115] 本领域技术人员可以理解，图9中示出的结构，仅仅是与本申请方案相关的部分结构的框图，并不构成对本申请方案所应用于其上的计算机设备的限定，具体的计算机设备可以包括比图中所示更多或更少的部件，或者组合某些部件，或者具有不同的部件布置。

[0116] 在一个实施例中，提供了一种计算机设备，包括存储器和处理器，该存储器存储有计算机程序，该处理器执行计算机程序时实现以下步骤：

[0117] 在编码端：

[0118] 获取预设的训练语料库中各个单词的词频值。

[0119] 将训练语料库中各个已词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值。

[0120] 根据词类内排序值相等的单词归入到同一叶子结点，根据叶子结点对应单词的词频值总和得到叶子结点的权重值，建立最优二叉树模型。

[0121] 为叶子结点分配无重复前缀码，得到待编码语料的编码数据。

[0122] 在译码端：

[0123] 根据最优二叉树模型得到编码数据对应的候选单词集合。

[0124] 使用预设的上下文关联模型得到对应的译码结果数据。

[0125] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：将训练语料库中的单词按照词性进行分类，得到对应的词性分类。词性分类包括名词分类、动词分类、形容词分类、副词分类和连词分类。在词性分类内，按照词频值由高到低的顺序得到各个单词的词类内排序值。

[0126] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：获取当前权重值最低的第一叶子结点和第二叶子结点，合并第一叶子结点和第二叶子结点得到第三叶子结点。根据第一叶子结点和第二叶子结点的权重值的和，得到第三叶子结点的权重值。

[0127] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：比较第一叶子结点和第二叶子结点的权重值，根据比较结果分别得到第一叶子结点和第二叶子结点的标签值。根据最优二叉树模型中从根结点到第一叶子结点经历的所有叶子结点的标签值序列，得到第一叶子结点的无重复前缀码。

[0128] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：获取训练语料库中各个单词间的上下文语义关联特征。根据上下文语义关联特征得到联合出现概率值最高的单词序列，得到对应的译码结果数据。

[0129] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：使用基于LSTM神经网络模型学习训练语料库中各个单词间的上下文语义关联特征。

[0130] 在一个实施例中，处理器执行计算机程序时还实现以下步骤：使用N-gram模型对上下文单词序列的联合概率分布建模。当单词序列的长度小于上下文窗口预设值时，根据N-gram模型，使用枚举法得到联合出现概率值最高的单词序列。当单词序列的长度大于上下文窗口预设值时，根据N-gram模型，使用状态压缩动态规划算法得到联合出现概率值最高的单词序列。

[0131] 在一个实施例中，提供了一种计算机可读存储介质，其上存储有计算机程序，计算机程序被处理器执行时实现以下步骤：

[0132] 在编码端：

[0133] 获取预设的训练语料库中各个单词的词频值。

[0134] 将训练语料库中各个词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值。

[0135] 将各个词类内排序值相等的单词归入同一叶子结点，根据叶子结点对应单词的词频值总和得到叶子结点的权重值，建立最优二叉树模型。

[0136] 为叶子结点分配无重复前缀码，得到待编码语料的编码数据。

[0137] 在译码端：

[0138] 根据最优二叉树模型得到编码数据对应的单词序列集合。

[0139] 使用预设的上下文关联模型得到对应的译码结果数据。

[0140] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：将训练语料库中的单词按照词性进行分类，得到对应的词性分类。词性分类包括名词分类、动词分类、形容词分类、副词分类和连词分类。在词性分类内，按照词频值由高到低的顺序，得到对应的单词序列，根据单词序列，得到各个单词的词类内排序值。

[0141] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：获取当前权重值最低的第一叶子结点和第二叶子结点，合并第一叶子结点和第二叶子结点得到第三叶子结点。根据第一叶子结点和第二叶子结点的权重值的和，得到第三叶子结点的权重值。

[0142] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：比较第一叶子结点和第二叶子结点的权重值，根据比较结果分别得到第一叶子结点和第二叶子结点的标签值。根据最优二叉树模型中从根结点到第一叶子结点经历的所有叶子结点的标签值序列，得到第一叶子结点的无重复前缀码。

[0143] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：获取训练语料库中各个单词间的上下文语义关联特征。根据上下文语义关联特征得到联合出现概率值最高的单词序列，得到对应的译码结果数据。

[0144] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：使用基于LSTM神经网络模型学习训练语料库中上下文单词间的上下文语义关联特征。

[0145] 在一个实施例中，计算机程序被处理器执行时还实现以下步骤：使用N-gram模型对上下文单词序列的联合概率分布建模。当单词序列的长度小于预设上下文窗口值时，根据N-gram模型，使用枚举法得到联合概率值最高的上下文单词序列。当单词序列的长度大于预设值时，根据N-gram模型，使用状态压缩动态规划算法得到联合出现概率值最高的上

下文单词序列。

[0146] 本领域普通技术人员可以理解实现上述实施例方法中的全部或部分流程，是可以通过计算机程序来指令相关的硬件来完成，所述的计算机程序可存储于一非易失性计算机可读取存储介质中，该计算机程序在执行时，可包括如上述各方法的实施例的流程。其中，本申请所提供的各实施例中所使用的对存储器、存储、数据库或其它介质的任何引用，均可包括非易失性和/或易失性存储器。非易失性存储器可包括只读存储器(ROM)、可编程ROM(PROM)、电可编程ROM(EPROM)、电可擦除可编程ROM(EEPROM)或闪存。易失性存储器可包括随机存取存储器(RAM)或者外部高速缓冲存储器。作为说明而非局限，RAM以多种形式可得，诸如静态RAM(SRAM)、动态RAM(DRAM)、同步DRAM(SDRAM)、双数据率SDRAM(DDRSDRAM)、增强型SDRAM(ESDRAM)、同步链路(Synchlink)DRAM(SLDRAM)、存储器总线(Rambus)直接RAM(RDRAM)、直接存储器总线动态RAM(DRDRAM)、以及存储器总线动态RAM(RDRAM)等。

[0147] 以上实施例的各技术特征可以进行任意的组合，为使描述简洁，未对上述实施例中的各个技术特征所有可能的组合都进行描述，然而，只要这些技术特征的组合不存在矛盾，都应当认为是本说明书记载的范围。

[0148] 以上所述实施例仅表达了本申请的几种实施方式，其描述较为具体和详细，但并不能因此而理解为对发明专利范围的限制。应当指出的是，对于本领域的普通技术人员来说，在不脱离本申请构思的前提下，还可以做出若干变形和改进，这些都属于本申请的保护范围。因此，本申请专利的保护范围应以所附权利要求为准。

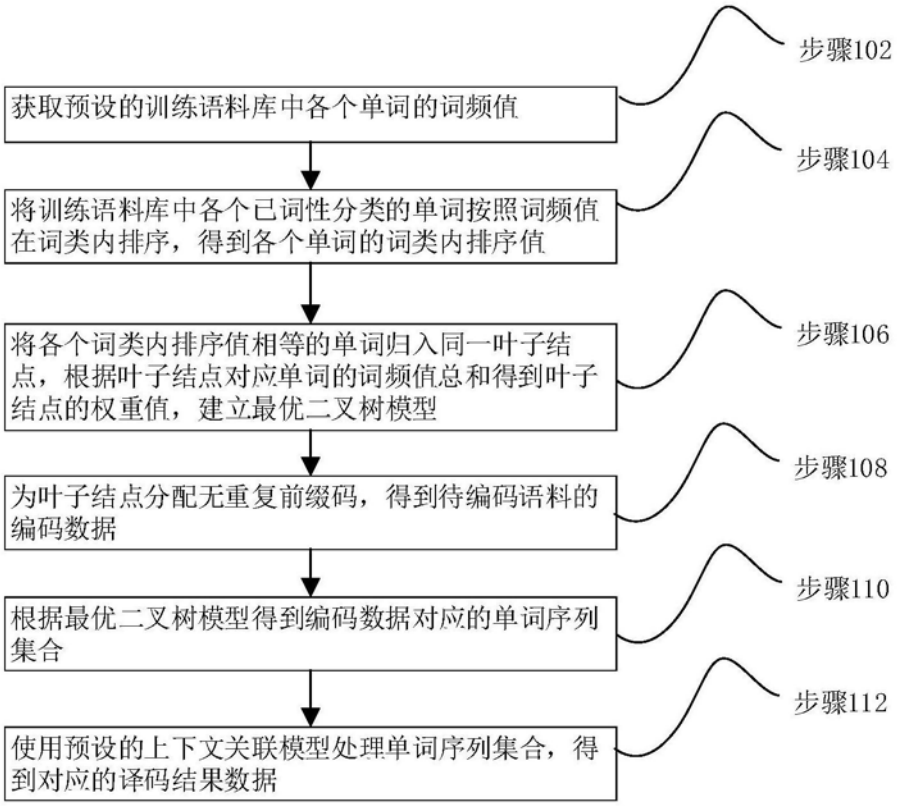

步骤102
获取预设的训练语料库中各个单词的词频值
步骤104
将训练语料库中各个已词性分类的单词按照词频值在词类内排序，得到各个单词的词类内排序值
步骤106
将各个词类内排序值相等的单词归入同一叶子结点，根据叶子结点对应单词的词频值总和得到叶子结点的权重值，建立最优二叉树模型
步骤108
为叶子结点分配无重复前缀码，得到待编码语料的编码数据
步骤110
根据最优二叉树模型得到编码数据对应的单词序列集合
步骤112
使用预设的上下文关联模型处理单词序列集合，得到对应的译码结果数据

图1

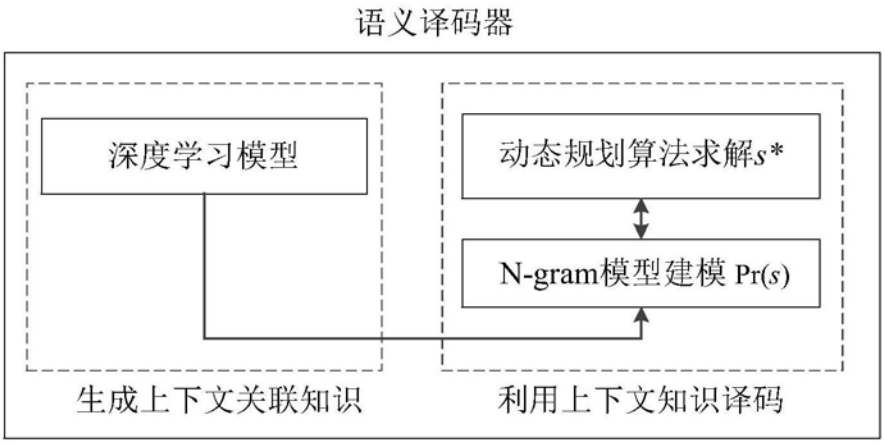

语义译码器
深度学习模型
动态规划算法求解s*
N-gram模型建模 Pr(s)
生成上下文关联知识
利用上下文知识译码

图2

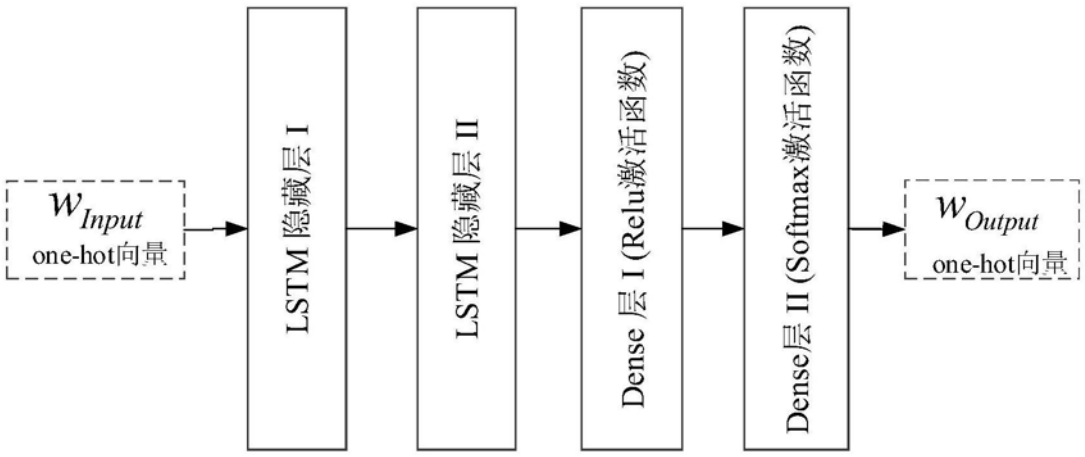

w_{Input}
one-hot向量
LSTM 隐藏层 I
LSTM 隐藏层 II
Dense 层 I (Relu激活函数)
Dense层 II (Softmax激活函数)
w_{Output}
one-hot向量

图3

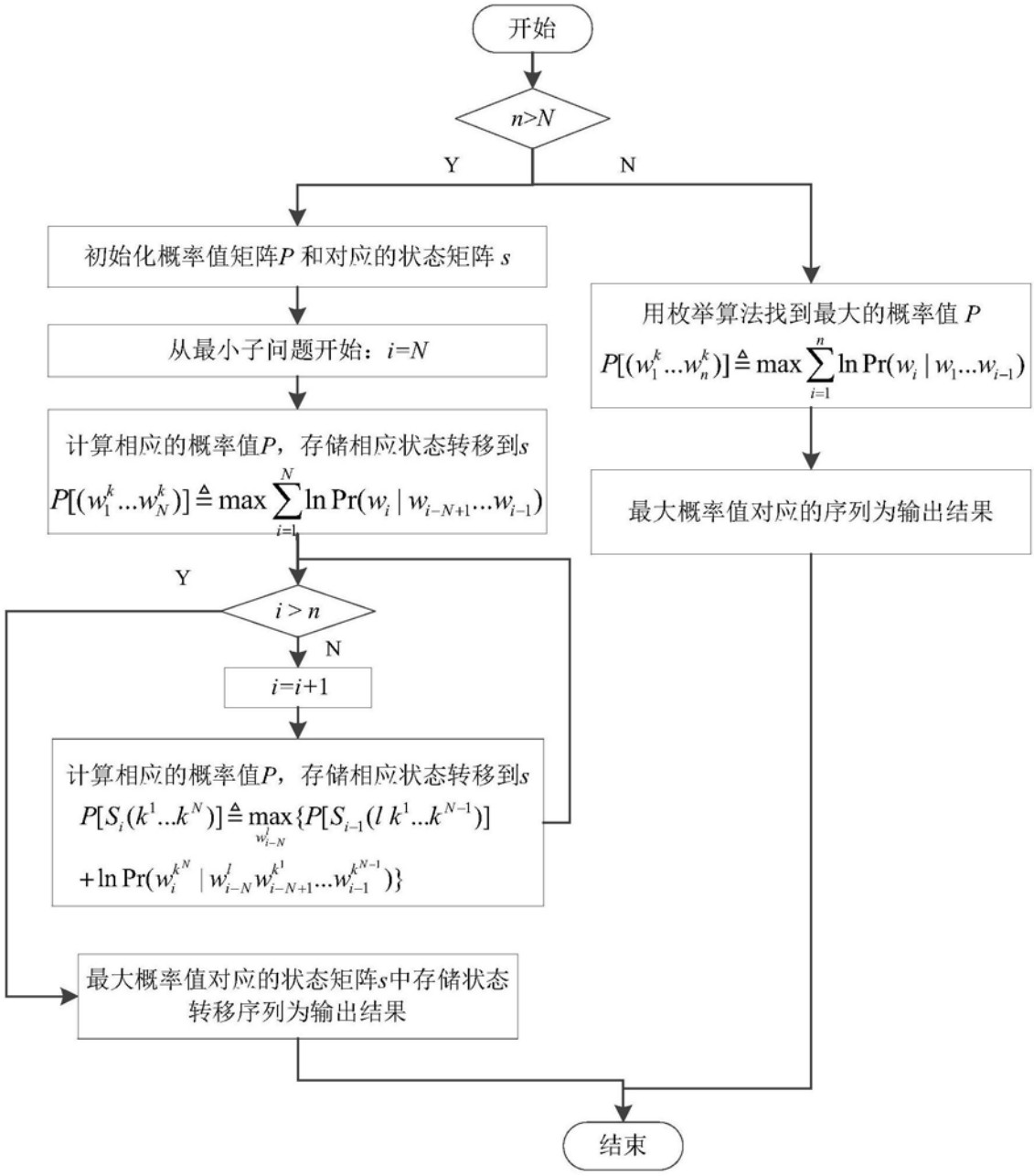

开始
n>N
Y
N
初始化概率值矩阵P 和对应的状态矩阵 s
从最小子问题开始：i=N
计算相应的概率值P，存储相应状态转移到s
$P[(w_1^k...w_N^k)] \triangleq \max \sum_{i=1}^{N} \ln \Pr(w_i \mid w_{i-N+1}...w_{i-1})$
用枚举算法找到最大的概率值 P
$P[(w_1^k...w_n^k)] \triangleq \max \sum_{i=1}^{n} \ln \Pr(w_i \mid w_1...w_{i-1})$
最大概率值对应的序列为输出结果
Y
i > n
N
i=i+1
计算相应的概率值P，存储相应状态转移到s
$P[S_i(k^1...k^N)] \triangleq \max_{w_{i-N}^l} \{P[S_{i-1}(l\,k^1...k^{N-1})]$
$+\ln \Pr(w_i^{k^N} \mid w_{i-N}^l w_{i-N+1}^{k^1}...w_{i-1}^{k^{N-1}})\}$
最大概率值对应的状态矩阵s中存储状态转移序列为输出结果
结束

图4

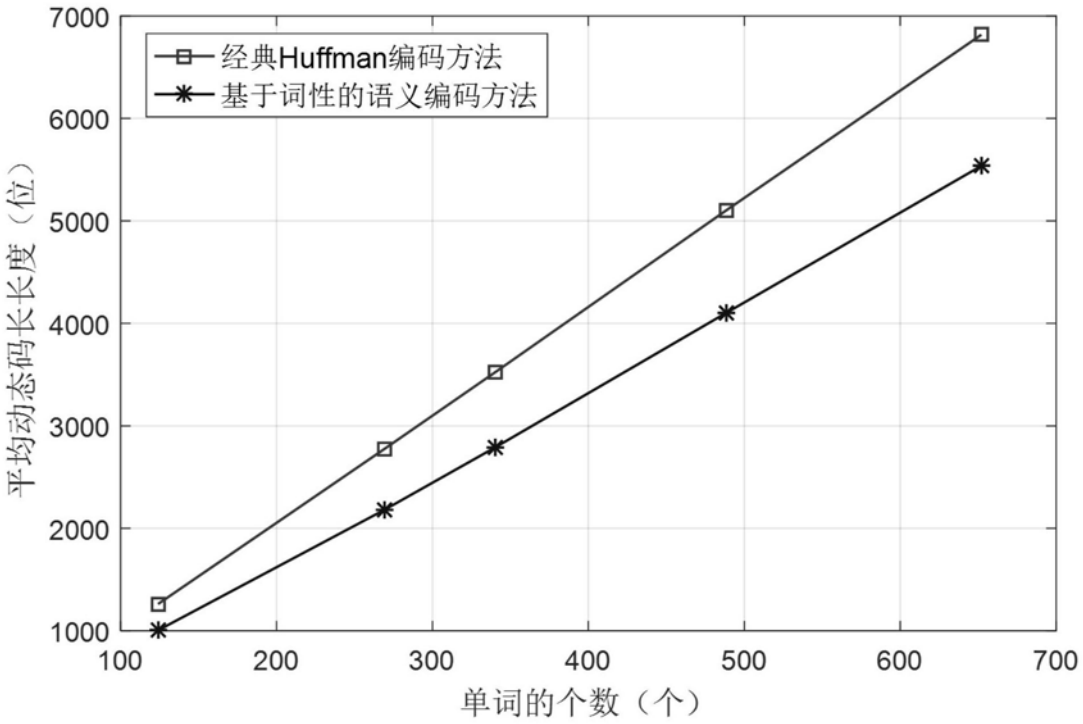

经典Huffman编码方法
基于词性的语义编码方法
平均动态码长长度（位）
单词的个数（个）
7000
6000
5000
4000
3000
2000
1000
100
200
300
400
500
600
700

图5

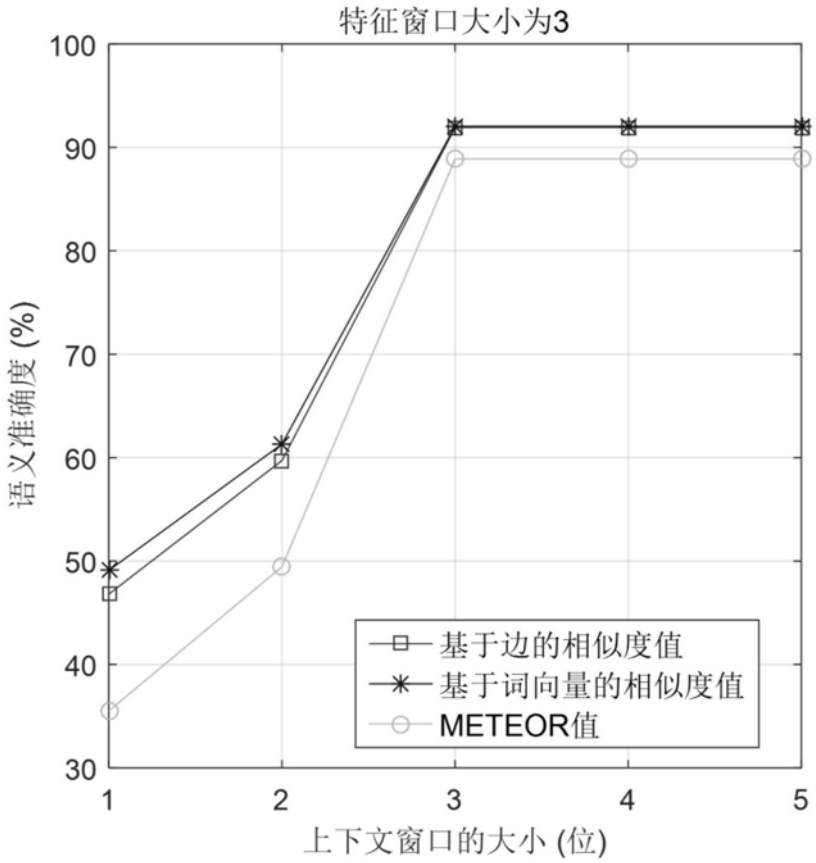

特征窗口大小为3
基于边的相似度值
基于词向量的相似度值
METEOR值
语义准确度 (%)
上下文窗口的大小 (位)
100
90
80
70
60
50
40
30
1
2
3
4
5

图6

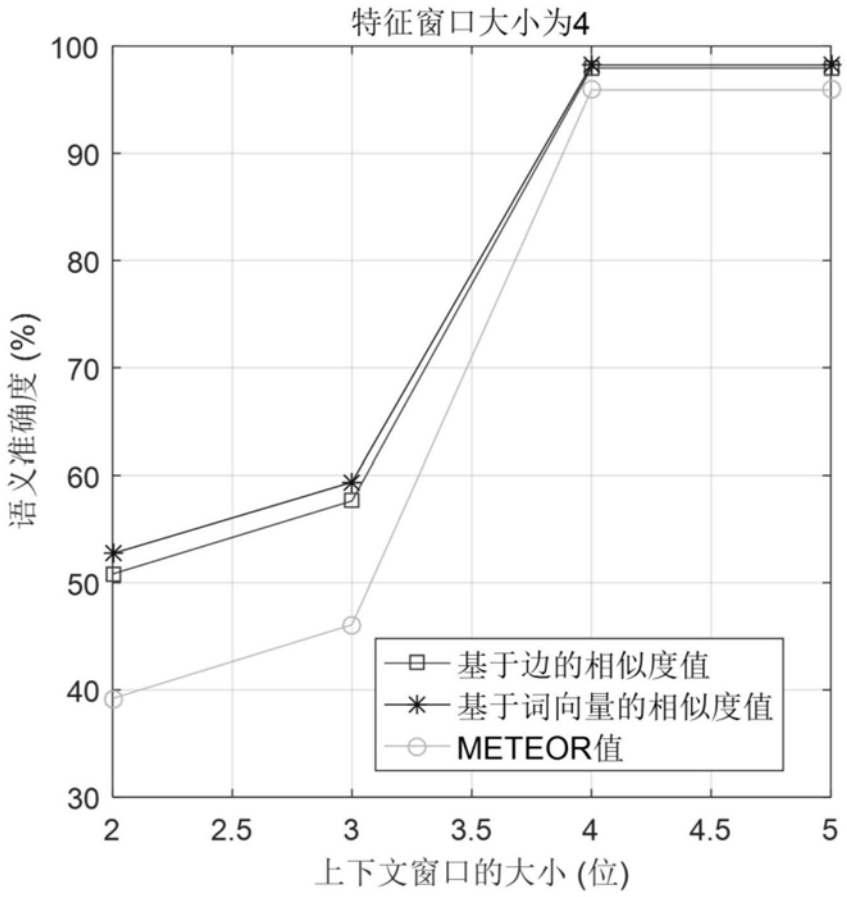

特征窗口大小为4
语义准确度 (%)
上下文窗口的大小 (位)
基于边的相似度值
基于词向量的相似度值
METEOR值
100
90
80
70
60
50
40
30
2
2.5
3
3.5
4
4.5
5

图7

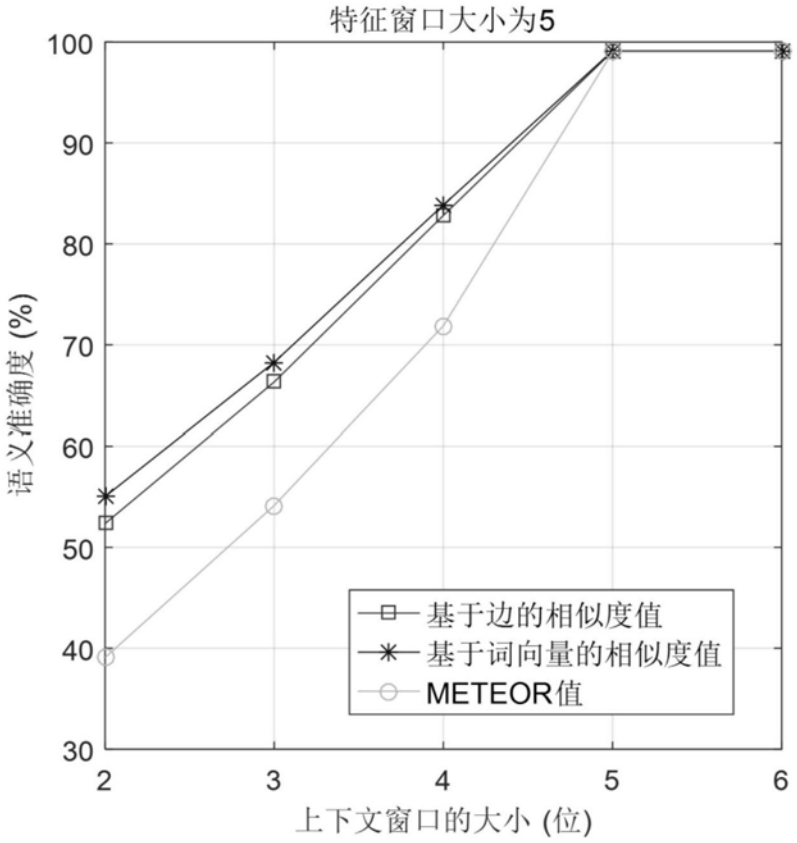

特征窗口大小为5
语义准确度 (%)
上下文窗口的大小 (位)
基于边的相似度值
基于词向量的相似度值
METEOR值
100
90
80
70
60
50
40
30
2
3
4
5
6

图8

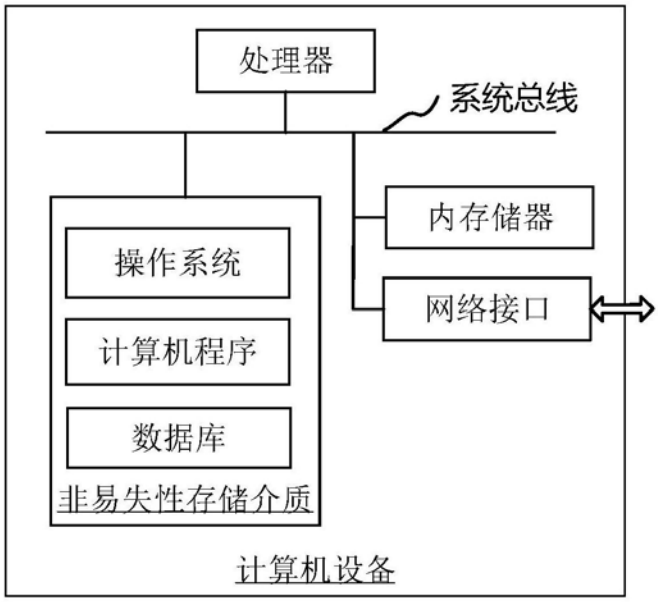

处理器
系统总线
内存储器
网络接口
操作系统
计算机程序
数据库
非易失性存储介质
计算机设备

图9