(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 108536754 A
(43)申请公布日 2018.09.14

(21)申请号 201810207151.2

(22)申请日 2018.03.14

(71)申请人 四川大学
地址 610064 四川省成都市武侯区望江路29号

(72)发明人 李智 杨金山 李健

(51)Int.Cl.
G06F 17/30(2006.01)
G06F 17/27(2006.01)
G06N 3/04(2006.01)
G06N 3/08(2006.01)

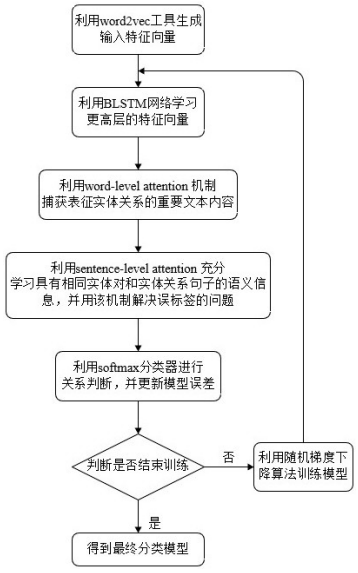
权利要求书4页 说明书4页 附图2页

(54)发明名称

基于BLSTM和注意力机制的电子病历实体关系抽取方法

(57)摘要

本发明提出一种基于BLSTM和注意力机制的电子病历实体关系抽取方法。该方法首先通过word2vec工具包将电子病历自然语句映射成为基本特征向量,然后利用BLSTM将基本特征向量编码成上层特征向量,接着利用基于词和句子级别的注意力机制捕获表征实体关系的重要文本内容以形成更高层次的特征向量,最后将得到的特征向量输入到softmax分类器中,抽取该语句中所有实体对之间的实体关系。另外,本方法没有利用任何依赖于任何知识库和专业词典来生成基本特征,降低了模型对人工特征工程的依赖性,为自动学习电子病历信息提供了技术途径。



CN 108536754 A

1.基于BLSTM和注意力机制的电子病历实体关系抽取方法，其特征在于：利用双向LSTM神经网络自动生成特征向量，降低现存电子病历实体关系抽取模型对于手工特征工程质量的依赖性，引入注意力机制提高模型正确识别实体关系的性能，步骤如下：

步骤1、得到输入基本特征向量表示

该基本特征向量主要由输入语句词本身(W)、每个词到实体对的相对距离和词类型3个部分连接构成

1)词本身(W)特征：

对于给定的具有n个单词的句子 $S=\{x_1,x_2,\cdots,x_n\}$，我们首先利用word2vec工具包将每个单词转换成低维度的实数向量，单词表示是通过嵌入矩阵 $W\in R^{d^w\times|V|}$ 中的列向量编码的，其中 V 是一个固定大小的词典，d^w 是嵌入矩阵的大小

2)每个词到实体对的相对距离特征：

我们用 $D\in\mathbb{R}^{d^d\times|v^d|}$ 矩阵来表征每个单词到实体对的距离，其中 d^d 是每个相对距离映射为实数向量后的维度，是一个可供用户调整的超参数，v^d 是固定大小的词典，即相对距离的范围大小，其具体定义是当前单词到头部或者尾部实体的相对距离

3)词类型特征：

在本方法中我们采用BIO标记法对所有单词进行标记，并将其作为基本特征之一，类似的，我们用 $T\in\mathbb{R}^{d^t\times|v^t|}$ 矩阵来表示它，其中 d^t 是单词所属类别映射为向量后的维度，v^t 是词类型特征矩阵的大小，即单词所属类别种类数量

最后，我们将上述3种基本特征拼接起来形成总的输入特征向量序列 $w=\{w_1,w_2,\ldots,w_n\}$，其中 $\mathbf{w}_i\in\mathbb{R}^d(d=d^w+2\times d^d+d^t)$

步骤2、利用BLSTM网络得到上层特征向量，其具体计算过程如下：

1)本方法中我们利用循环神经网络来学习长距离语义信息形成上层特征向量，其单个神经元结构如图2所示，具体而言，该LSTM模型主要涉及到遗忘门、更新门以及输出门3个组成部分，其中遗忘门的计算过程如下：

$$r_t=\sigma([\mathrm{W}_r\mathrm{x}]+[\mathrm{U}_r\mathrm{h}_{\langle t-1\rangle}]) \quad (1)$$

其中，σ 是logisticsigmoidfunction，$\mathbf{x}$ 和 h_{t-1} 分别代表输入和先前隐藏状态，W_r 和 U_r 是将要学习的权重矩阵

同样的，更新门 z_t 计算方法为：

$$z_t=\sigma([\mathrm{W}_z\mathrm{x}]+[\mathrm{U}_z\mathrm{h}_{\langle t-1\rangle}]) \quad (2)$$

输出门 $\mathbf{h}_t$ 的激活函数计算方法：

$$h_{\langle t\rangle} = (1 - z_t) \odot h_{\langle t-1\rangle} + z_t \odot \bar{h}_{\langle t\rangle} \quad (3)$$

其中，

$$\bar{h}_{\langle t\rangle} = \tanh([Wx] + [U(r_t \odot h_{\langle t-1\rangle})]) \quad (4)$$

2)进一步地，我们采用BLSTM去学习过去和未来文本语义信息，其结构如图1的BLSTM encoder layer所示，所以上层特征向量$H = [h_1, h_2, \cdots, h_n]$通过下式计算：

$$h_j = [\vec{h}_j \oplus \overleftarrow{h}_j] \quad (5)$$

其中，$H \in \mathbb{R}^{d^a \times n}$， d^a表示每个上层特征向量的维度，n表示句子长度

步骤3、利用基于词别的注意力机制(word-level attention)捕获表征实体关系的重要文本内容，其具体计算过程如下：

1)word-levelattention的核心思想是在形成更高层次特征表示时需要为每个单词设置一个可学习的权重向量$w_{word-att} \in \mathbb{R}^{d^a}$，其计算过程如下：

$$m = \tanh(H^T) \quad (6)$$

$$g = softmax(mw_{word-att}) \quad (7)$$

则该层网络的输出，即一个句子的表示x可通过下式得出：

$$x = g^T m \quad (8)$$

步骤4、利用基于句子级别的注意力机制(sentence-level attention)来充分学习具有相同实体对和实体关系句子的语义信息，并用该机制解决误标签的问题，得到网络的最终输出

1)假设一个包含m个句子(具有同样实体对，同样实体关系)的集合$S = \{x_1, x_2, \cdots, x_m\}$，对于(sentence-level attention，我们为集合中的每一个句子赋予一个可学习的权重α_i，然后将这些句子编码成一个实数向量s，α_i和s的计算方法如下：

$$e = (p^T \odot w^T{}_{sen-att}) v_r \quad (9)$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{m} \exp(e_i)} \quad (10)$$

$$s = \sum_i \alpha_i x_i \quad (11)$$

其中e_i是用来评价每个句子和关系r_i的匹配程度，$\odot$表示点乘，$w_{sen-att} \in \mathbb{R}^{d^a}$

是注意力权值矩阵，$v_r \in \mathbb{R}^{d^a}$ 是一个询问向量

2)然后利用向量 S 预测最终的关系 r_i，其计算过程如下

$$p = \tanh(S) \tag{12}$$

$$s = \sum \alpha^T p^T \tag{13}$$

$$o = w_r s^T \tag{14}$$

其中，$w_r \in \mathbb{R}^{n_r \times d^a}$ 是关系表示矩阵，n_r 是关系类型总数，o 是网络的最终输出

步骤5、关系判断

1)这里，我们利用步骤4中的输出 O 来判断实体对所属关系类型，我们定义条件概率 $p(r \mid S, \theta)$ 来预测句子集合 S 所属类别 $\hat{y}$，计算过程如下：

$$p(y \mid S, \theta) = \frac{\exp(o_k)}{\sum_{k=1}^{n_r} \exp(o_k)} \tag{15}$$

$$\hat{y} = \arg\max_y p(y \mid S, \theta) \tag{16}$$

2)代价函数定义如下：

$$J(\theta) = -\sum_{i=1}^{n_r} y_i \log(p_i) + \lambda \left\| \theta \right\|^2 \tag{17}$$

其中 $y_i \in \{0,1\}$ 是标签 i 的真实值，p_i 是每个类别的估计概率，λ 是一个L2 正则化参数。

2.根据权利要求1所述的基于BLSTM和注意力机制的电子病历实体关系抽取方法，其特征在于：步骤1中采用词本身、词和头尾实体的相对距离、词本身类型构成网络初始向量特征。

3.根据权利要求1所述的基于BLSTM和注意力机制的电子病历实体关系抽取方法，其特征在于：步骤2中采用LSTM网络学习步骤1中产生的基本特征向量形成更高层的特征向量，另外，我们利用BLSTM网络学习过去和未来文本语义信息，其计算方式如下：

$$r_t = \sigma([W_r x] + [U_r h_{\langle t-1 \rangle}])$$

$$z_t = \sigma([W_z x] + [U_z h_{\langle t-1 \rangle}])$$

$$h_{\langle t \rangle} = (1 - z_t) \odot h_{\langle t-1 \rangle} + z_t \odot \bar{h}_{\langle t \rangle}$$

$$\bar{h}_{\langle t \rangle} = \tanh([Wx] + [U(r_t \odot h_{\langle t-1 \rangle})])$$

$\mathrm{h}_j = [\vec{\mathrm{h}}_j \oplus \overleftarrow{\mathrm{h}}_j]$。

4.根据权利要求1所述的基于BLSTM和注意力机制的电子病历实体关系抽取方法，其特征在于：步骤3中利用基于词别的注意力机制(word-level attention)捕获步骤2中表征实体关系的重要文本内容，其具体计算过程如下：

$$m = \tanh(H^T)$$

$$g = softmax\left(mv_{word-att}\right)$$

$$x = g^T m。$$

5.根据权利要求1所述的基于BLSTM和注意力机制的电子病历实体关系抽取方法，其特征在于：步骤4中利用基于句子级别的注意力机制(sentence-level attention)来充分学习步骤3中具有相同实体对和实体关系句子的语义信息，并用该机制解决误标签的问题，得到网络的最终输出，其就算过程如下：

$$e = (p^T \odot w^T{}_{sen-att})v_r$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{m}\exp(e_i)}$$

$$s = \sum_i \alpha_i x_i$$

$$p = \tanh(S)$$

$$s = \sum \alpha^T p^T$$

$$o = w_r s^T 。$$

6.根据权利要求1所述的基于BLSTM和注意力机制的电子病历实体关系抽取方法，其特征在于：步骤5中利用步骤4中的输出 O 来判断实体对所属关系类型，我们定义条件概率 $p(r \mid S,\theta)$ 来预测句子集合 S 所属类别 $\hat{y}$，并利用随机梯度下降算法训练模型，计算过程如下：

$$p(y \mid S,\theta) = \frac{\exp(o_k)}{\sum_{k=1}^{n_r}\exp(o_k)}$$

$$\hat{y} = \arg\max_y p(y \mid S,\theta)$$

$$J(\theta) = -\sum_{i=1}^{n_r} y_i \log(p_i) + \lambda\left\|\theta\right\|^2 。$$

基于BLSTM和注意力机制的电子病历实体关系抽取方法

技术领域

[0001] 本发明属于自然语言处理领域，用于自动抽取电子病历中实体对之间的实体关系。

背景技术

[0002] 随着信息时代的到来，各领域数据呈爆炸式增长。具体到医疗领域中，积累了大量包含着医疗健康领域知识的电子病历文本。在这种背景下，从非结构化的电子病历中抽取相关信息成为了获取医疗知识的关键，具有重要的应用价值。电子病历实体对之间的关系抽取是其核心任务之一。

[0003] 目前，电子病历的实体关系抽取主要是采用有监督的机器学习，该方法首先对候选实体进行特征选择，加入医疗知识作为辅助分析，并将抽取得到的特征转化为特征向量，在向量空间模型中进行有监督学习的分类判别，由此而得到实体对的关系。具体又主要分为基于规则、基于特征向量2个研究方向：基于规则方法根据待处理语料涉及领域的不同，通过人工总结归纳出相应的规则或模板，然后采用模板匹配的方法进行实体关系抽取。基于特征向量的方法主要思想是从句子中提取词法、语法信息来构造特征向量，通过计算特征向量的相似度来训练实体关系抽取模型。但是，这些方法存在一些明显的缺点：

(1)模型的性能极大程度依赖于手工特征工程的质量而导致其泛化性能较差，而且十分耗时

(2)规则制定依赖于专家知识和人工归纳

(3)模型过度依赖于知识库和其他NLP系统

发明内容

[0004] 本发明为了降低现存电子病历实体关系抽取模型对于手工特征工程质量的依赖性和提高模型正确识别实体关系的性能，提出了基于BLSTM和注意力机制的电子病历实体关系抽取方法。为了实现上述目的，该方法首先通过word2vec工具包将电子病历自然语句映射成为基本特征向量，然后利用BLSTM自动将基本特征向量编码成上层特征向量，接着利用基于词和句子级别的注意力机制捕获表征实体关系的重要文本内容以形成更高层次的特征向量，最后将得到的特征向量输入到softmax分类器中，抽取该语句中所有实体对之间的实体关系。另外，本方法没有利用任何依赖于任何知识库和专业词典来生成基本特征，降低了模型对人工特征工程的依赖性

附图说明

[0005] 图1是该电子病历实体关系抽取模型系统框架图。

[0006] 图2是循环神经网络(LSTM)单个神经元示意图。

[0007] 图3是本发明中提出的电子病历实体关系抽取方法流程图。

具体实施方式

[0008] 下面结合具体实施方式对本发明做进一步的详细说明：

1.得到输入基本特征向量表示

该基本特征向量主要由输入语句词本身(W)、每个词到实体对的相对距离和词类型3个部分连接构成

1)词本身(W)特征：

对于给定的具有n个单词的句子S＝{x_1,x_2,…,x_n},我们首先利用word2vec工具包将每个单词转换成低维度的实数向量。单词表示是通过嵌入矩阵$W \in R^{d^w \times |V|}$中的列向量编码的,其中V是一个固定大小的词典,d^w是嵌入矩阵的大小

2)每个词到实体对的相对距离特征：

我们用$D \in \mathbb{R}^{d^d \times |v^d|}$矩阵来表征每个单词到实体对的距离,其中$d^d$是每个相对距离映射为实数向量后的维度,是一个可供用户调整的超参数,v^d是固定大小的词典,即相对距离的范围大小。其具体定义是当前单词到头部或者尾部实体的相对距离

3)词类型特征：

在本方法中我们采用BIO标记法对所有单词进行标记,并将其作为基本特征之一。类似的,我们用$T \in \mathbb{R}^{d^t \times |v^t|}$矩阵来表示它,其中$d^t$是单词所属类别映射为向量后的维度,$v^t$是词类型特征矩阵的大小,即单词所属类别种类数量

最后,我们将上述3种基本特征拼接起来形成总的输入特征向量序列 w＝{w_1,w_2,...,w_n},其中$\mathbf{w}_i \in \mathbb{R}^d (d = d^w + 2 \times d^d + d^t)$

2.利用BLSTM网络得到上层特征向量,其具体计算过程如下:

1)本方法中我们利用循环神经网络来学习长距离语义信息形成上层特征向量。其单个神经元结构如图2所示,具体而言,该LSTM模型主要涉及到遗忘门、更新门以及输出门3个组成部分。其中遗忘门的计算过程如下：

r_t＝σ([W_rX]+[$U_r h_{<t-1>}$]) (1)

其中,σ是logistic sigmoid function,X和h_{t-1}分别代表输入和先前隐藏状态,W_r和U_r是将要学习的权重矩阵

同样的,更新门Z_t计算方法为：

Z_t＝σ([W_zx]+[$U_z h_{<t-1>}$]) (2)

输出门h_t的激活函数计算方法：

$$h_{\langle t \rangle} = (1 - z_t) \odot h_{\langle t-1 \rangle} + z_t \odot \bar{h}_{\langle t \rangle} \quad (3)$$

其中,

$$\bar{h}_{\langle t \rangle} = \tanh([\mathbf{Wx}] + [\mathbf{U}(\mathbf{r}_t \odot \mathbf{h}_{\langle t-1 \rangle})]) \quad (4)$$

2)进一步地,我们采用BLSTM去学习过去和未来文本语义信息,其结构如图1的BLSTM encoder layer所示,所以上层特征向量H＝[h_1,h_2,…,h_n]通过下式计算：

$$h_j = [\vec{h}_j \oplus \overleftarrow{h}_j] \quad (5)$$

其中，$H \in \mathbb{R}^{d^a \times n}$，$d^a$表示每个上层特征向量的维度，n表示句子长度 3.利用基于词别的注意力机制(word-level attention)捕获表征实体关系的重要文本内容，其具体计算过程如下：

1)word-level attention的核心思想是在形成更高层次特征表示时需要为每个单词设置一个可学习的权重向量$w_{word-att} \in \mathbb{R}^{d^a}$，其计算过程如下：

$m = \tanh(H^T)$ (6)

$g = \mathrm{softmax}(mw_{word-att})$ (7)

则该层网络的输出，即一个句子的表示x可通过下式得出：

$x = g^T m$ (8)

4.利用基于句子级别的注意力机制(sentence-level attention)来充分学习具有相同实体对和实体关系句子的语义信息，并用该机制解决误标签的问题

1)假设一个包含m个句子(具有同样实体对，同样实体关系)的集合$S = \{x_1, x_2, \cdots, x_m\}$，对于sentence-level attention，我们为集合中的每一个句子赋予一个可学习的权重α_i，然后将这些句子编码成一个实数向量s。α_i和s的计算方法如下：

$$e = (p^T \odot w^T_{sen-att}) v_r \quad (9)$$

$$\alpha_i = \frac{\exp(e_i)}{\sum_{i=1}^{m} \exp(e_i)} \quad (10)$$

$$s = \sum_i \alpha_i x_i \quad (11)$$

其中e_i是用来评价每个句子和关系r_i的匹配程度，⊙表示点乘，$w_{sen-att} \in \mathbb{R}^{d^a}$注意力权值矩阵，$v_r \in \mathbb{R}^{d^a}$是一个询问向量

2)然后利用向量s预测最终的关系r_i，其计算过程如下

$p = \tan(S)$ (12)

$s = \sum \alpha^T p^T$ (13)

$o = w_r s^T$ (14)

其中，$w_r \in \mathbb{R}^{n_r \times d^a}$是关系表示矩阵，$n_r$是关系类型总数，o是网络的最终输出

5.关系判断

1)这里，我们利用步骤4中的输出o来判断实体对所属关系类型。我们定义条件概率 $p(r|S,\theta)$来预测句子集合S所属类别$\hat{y}$，计算过程如下：

$$p(y|S,\theta) = \frac{\exp(o_k)}{\sum_{k=1}^{n_y} \exp(o_k)} \quad (15)$$

$$\hat{y} = \arg\max_{y} p(y \mid S, \theta) \quad (16)$$

2)代价函数定义如下：

$$J(\theta) = -\sum_{i=1}^{n_r} y_i \log(p_i) + \lambda \left\| \theta \right\|^2 \quad (17)$$

其中yi∈{0,1}是标签i的真实值，p_i是每个类别的估计概率，λ是一个L2正则化参数。

[0009] 附图详细说明：

图1为基于BLSTM和注意力机制的电子病历实体关系抽取模型网络结构图，输入为由电子病历文本内容映射为实数的特征向量，经过BLSTM网络学习得到更高层的特征向量，然后利用注意力机制学习表征实体关系的重要文本内容，输出结果为实体对所属实体关系类别；

图2是LSTM网络的单个神经元结构图，具体而言，该LSTM模型主要涉及到遗忘门、更新门以及输出门3个组成部分；

图3为本发明的方法流程图，主要在于利用深度神经网络自动学习得到特征向量，降低模型对手动特征工程的依赖性；利用注意力机制学习表征实体关系的重要文本内容，提高模型正确识别实体关系的性能。

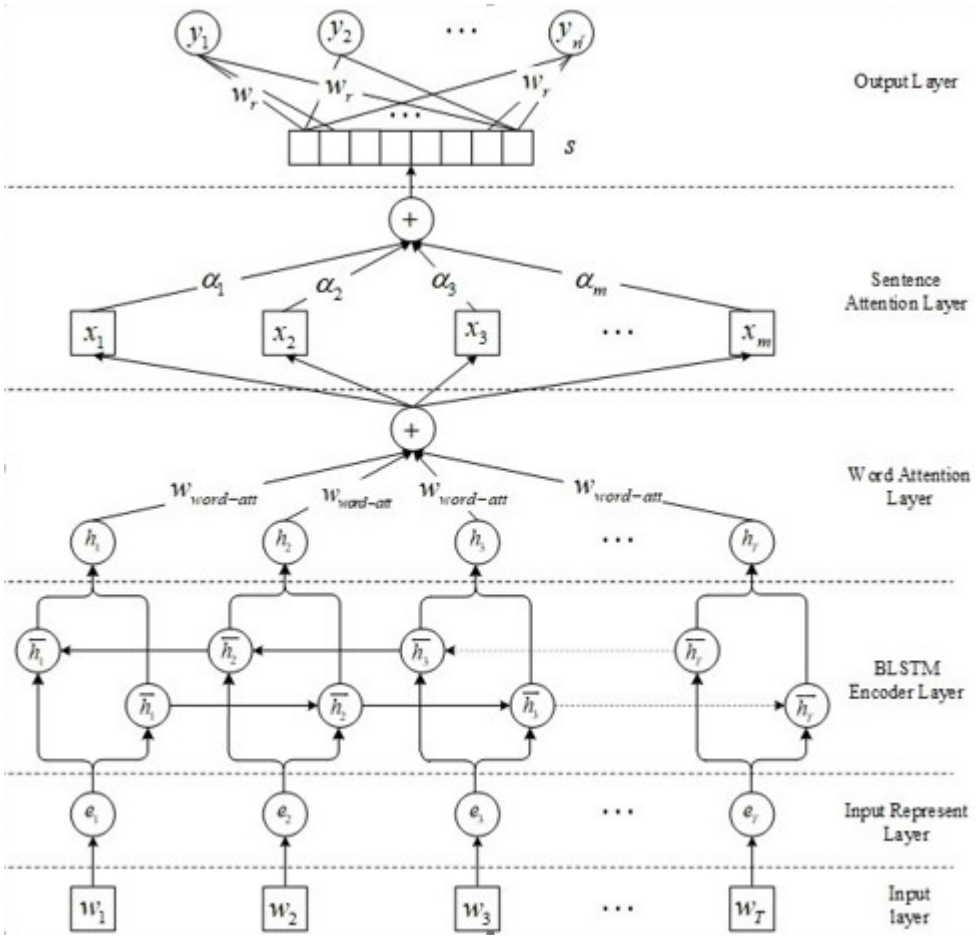

Output Layer
Sentence Attention Layer
Word Attention Layer
BLSTM Encoder Layer
Input Represent Layer
Input layer

图1

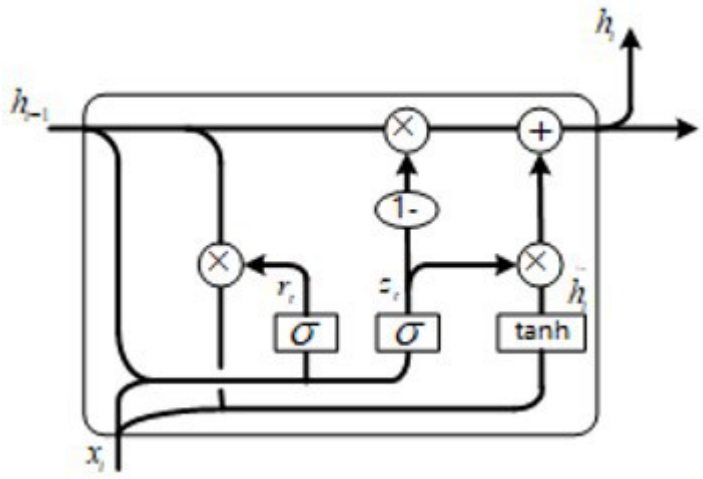

1-
tanh

图2

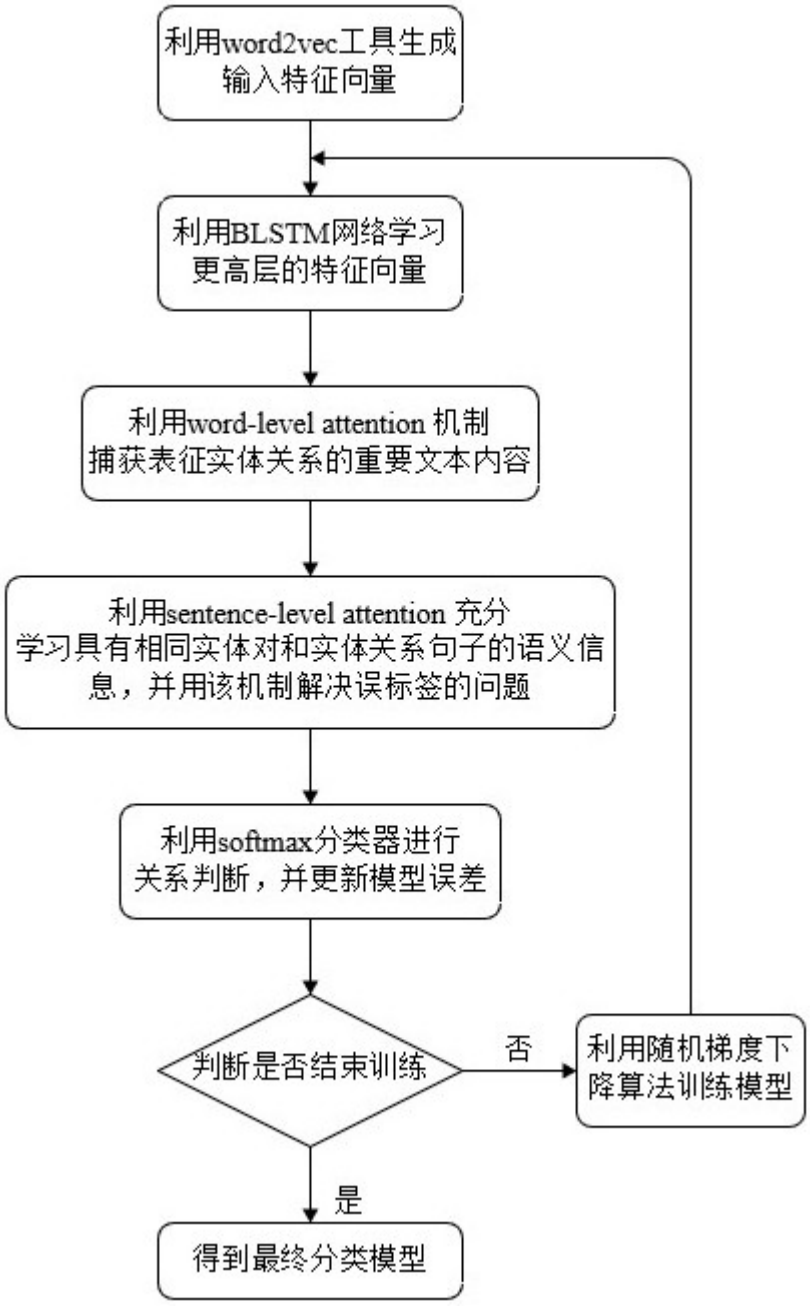

利用word2vec工具生成
输入特征向量
利用BLSTM网络学习
更高层的特征向量
利用word-level attention 机制
捕获表征实体关系的重要文本内容
利用sentence-level attention 充分
学习具有相同实体对和实体关系句子的语义信
息，并用该机制解决误标签的问题
利用softmax分类器进行
关系判断，并更新模型误差
判断是否结束训练
否
利用随机梯度下
降算法训练模型
是
得到最终分类模型

图3