ISSN 1004-9037,CODEN SCYCE4
Journal of Data Acquisition and Processing Vol. 35,No. 5,Sep. 2020,pp. 930－941
DOI:10. 16337/j. 1004-9037. 2020. 05. 014

http://sjcj. nuaa. edu. cn
E-mail:sjcj@nuaa. edu. cn
Tel/Fax:＋86-025-84892742

问答系统命名实体识别改进方法研究

鲍静益[1]，于佳卉[2]，徐　宁[2,3]，姚　潇[2,3]，刘小峰[2,3]

(1. 常州工学院电气信息工程学院，常州，213022；2. 河海大学物联网工程学院，常州，213022；3. 江苏省特种机器人与智能技术重点实验室，常州，213022)

摘　要：问答系统是一种以准确且自然的语言来回答用户提问的系统。本文对其中涉及的“命名实体识别”环节尝试了一些改进措施：针对传统单向模板匹配耗时耗力的问题，提出一种双向格子结构的长短时记忆网络(Lattice Bi-LSTM)，解决了命名实体识别中对句子处理不当和对分词结果具有依赖性的两大问题。与单向结构相比，双向结构能更好地利用句子信息，使输出结果更具鲁棒性，从而更准确地捕获语义信息；针对传统方法未考虑实体间相似度的非线性耦合性问题，提出一种利用周期性核函数将“相似”实体准确链接到知识库中去的方法。对提出的两种改进方法进行了实验验证，结果表明：与经典方法相比，所用方法具有显著的改进效果。

关键词：问答系统；命名实体识别；双向格子长短时记忆模型；周期核函数；相似度评判

中图分类号：TN912.3　　**文献标志码**：A

Research on the Improved Method of Named Entity Recognition in Q & A System

BAO Jingyi[1], YU Jiahui[2], XU Ning[2,3], YAO Xiao[2,3], LIU Xiaofeng[2,3]

(1. School of Electrical and Information Engineering, Changzhou Institute of Technology, Changzhou, 213022, China; 2. School of IoT Engineering, Hohai University, Changzhou, 213022, China; 3. Changzhou Key Laboratory of Robotics and Intelligent Technology, Changzhou, 213022, China)

Abstract: The Q & A system is a kind of system which can answer user's questions with accurate and natural language. Some improvement measures have been tried for "named entity recognition". Aiming at the time- and labor-consuming problem of traditional one-way template matching, this paper proposes a lattice bi-directional structure of long short-term memory (Lattice Bi-LSTM) network, which solves the problems of improper sentence processing and dependence on the result of word segmentation in named entity recognition. Compared with the unidirectional structure, the bi-directional structure can make better use of sentence information and make the output more robust, thus capturing semantic information more accurately. To solve the problem of non-linear coupling of similarity between entities in traditional methods, a method is proposed to link "similar" entities to the knowledge base accurately by using periodic kernel function. The two improved methods are verified by experiments, whose results show that they have significant improvement effects compared with the classical method.

Key words: Q & A system; named entity recognition; lattice bi-directional long short memory model;

基金项目：江苏省重点研发计划(BK20192004, BE2018004-04)资助项目；中央高校科研基本业务费(B200202205)资助项目；飞行交通管理与技术重点实验室开放课题(SKLATM201901)资助项目。
收稿日期：2020-07-07；**修订日期**：2020-08-25

periodic kernel function; similarity evaluation

引　言

问答系统起源于图灵测试,若计算机能使用自然语言回答问题,则认为该计算机具有人工智能[1]。作为自然语言处理领域的主要研究方向之一,问答系统被应用在多个领域。如MIT大学的Boris Katz与其同伴研究出世界上第一个基于web的问答系统——Start系统,可完成查天气、设闹钟和搜信息等一系列生活服务[2];日常生活中人们经常用到的苹果语音助手siri也是问答系统的一种典型应用[3]。

一套完整的问答系统一般包含4项基本任务,即词性标注、句子情感分析、分类任务以及命名实体识别(Named entity recognition, NER)。NER中的传统方法主要有两类,一类基于规则和模板[4],即人工根据知识集或者词典搭建模板,选用一些关键字或者位置词作为特征,利用字符串匹配的方法将关键词和模板进行匹配;另一类是基于传统机器学习的方法,主要包括条件随机场(Conditional random fields, CRF)[6]、隐马尔可夫模型(Hidden Markov model, HMM)[7]、支持向量机(Support vector machine, SVM)[8]、最大熵(Maximum entropy, ME)[9]4种方法。CRF方法提供了一个灵活提取特征参数的框架,但该方法所需训练时间比较长;HMM模型训练时虽然所需时间较少、识别速度较快,但准确率不高;SVM模型用于NER中时,准确率通常比HMM要高,但一般仅用于分类子任务而不是完整的NER,作用域有限;ME模型准确率一般来说比HMM高,但其训练的时间复杂度较高,且需要进行归一化计算,损失值较大。

近年来,随着神经网络领域研究的蓬勃发展,传统NER方法用的越来越少,而基于神经网络的方法开始占据主要地位,被有效地应用在自然语言处理的各个领域。例如Zhang等[10]提出了一种格子结构的长短时记忆网络(Lattice long short-term memory, Lattice LSTM)模型,能够不受分词效果的影响,也不破坏原句的语义。神经网络方法的优点在于:对数据集的依赖程度没有前两种传统方法大。但神经网络中的模型种类较多,因此模型受自身定义的参数影响比较大。除此之外,该方法还有个弊端,即进行标签预测时,每次的预测过程是一个互相独立的分类,对于已预测好的标签,无法直接进行利用。另一方面,完成NER之后,识别出的实体需要与知识库中存在的实体进行相似度计算,以便找到相似度最高的一类特征,从而实现在知识图谱中搜索答案的目的。传统的计算相似度方法,如余弦相似度、编辑距离和马氏距离等,由于未考虑中文语言之间的相关性,故而计算所得的相似度评分一般偏低。

针对上述问题,本文首先提出了双向格子结构的长短时记忆网络(Lattice bi-directional LSTM, Lattice Bi-LSTM)模型,在原模型的基础上,添加了一层长短时记忆网络,使原模型中的LSTM层从单向变为双向,使得LSTM在处理信息时,能够同时进行前向传播和后向传播,从而在处理某些长句时,同时获取过去和未来两个状态的信息并对其进行综合性考虑,使其输出信息更具完整性和鲁棒性;其次,本文提出一种基于周期性核函数的相似度计算新方法,该方法充分考虑了长句之间的周期性重复词语出现的频率特征,对两个待评判的实体进行核函数向量计算,以实现对时间轴关系上的非线性耦合性特征的有效建模。

1　经典方法

1.1　经典命名实体识别模型

最常用的NER中的经典模型有基于字向量的模型和基于词向量的模型。这两种模型具有一定的

限制性，前者对句子处理不当，后者对分词结果具有依赖性。

基于字向量的模型结构如图1所示。可以看出，该模型是对“宁波市长江小学”单字分开，变成“宁/波/市/长/江/小/学”进行处理。假设图1中的模型有 a 个字序列通过，分别为 $c_1, c_2, \cdots, c_a$，其中第 i 个字 c_i 输入时，被表示为输入向量 $\boldsymbol{x}_i^c = \boldsymbol{e}^c(c_i)$，其中 $\boldsymbol{e}^c$ 表示权重矩阵，是在处理字的embedding层进行表示出来的。在基于字符向量模型中，用到的是一个双向LSTM，因此需要对每一个输入向量所对应的隐藏状态进行拼接，即 $x_1, x_2, \cdots, x_a$ 等分别对应了一个正方向的隐含层状态 $\overrightarrow{h_1^c}, \overrightarrow{h_2^c}, \cdots, \overrightarrow{h_a^c}$ 和一个反方向的隐含层状态 $\overleftarrow{h_1^c}, \overleftarrow{h_2^c}, \cdots, \overleftarrow{h_a^c}$。则输入的第 i 个字的总隐藏层状态输出就可以表示为：$h_i^c = \left[\overrightarrow{h_i^c}, \overleftarrow{h_i^c}\right]$，即总隐藏层状态需要将两个方向的隐藏层状态进行拼接后表示。

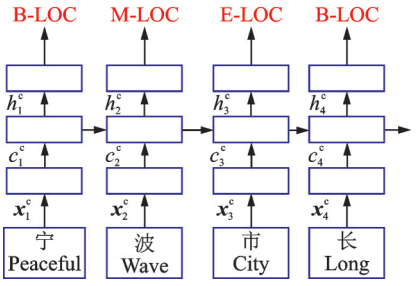

图1 基于字向量的处理模型

Fig.1 Processing model based on character vector

基于词向量的模型结构如图2所示，该模型是对“宁波市长江小学”中的词语进行处理拆分，变成“宁波/市/长江/小学”进行处理。其原理如下，假设图2中的模型有 n 个词语序列通过，分别为 $w_1, w_2, \cdots, w_n$，其中第 j 个词 w_j 输入时，被表示为输入向量 $\boldsymbol{x}_j^w = \boldsymbol{e}^w(w_j)$，其中 $\boldsymbol{e}^w$ 表示处理词语的embedding层定义的权重矩阵。其隐藏层状态是否进行拼接取决于是否使用双向的长短时记忆神经网络，一般采用单向的话可以直接得出其隐藏层状态 h_j^w。

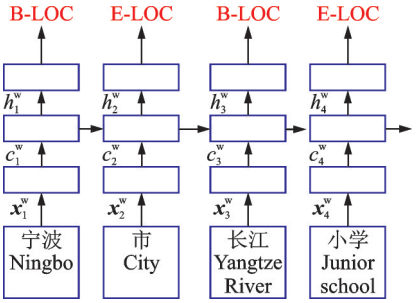

图2 基于词向量的处理模型

Fig.2 Processing model based on word vector

本文对比系统，即文献[10]提出的经典NER模型的基本思想是在基于字向量模型的基础上，对该模型增加一个栅格结构，该结构包含所有从词典里学到的词，其示意图如图3所示。可以看出，除了本身拆分的字之外，使用的栅格结构可得到整个句子中所有词典里学到的词，比如图3中的“宁波”“市长”“长江”“小学”“宁波市”“长江小学”，如果原句按照字向量进行划分，还可以组成“宁波/市长/江小学”的格式，但由于设定了栅格，栅格中不包括“江小学”这个词，就避免了原句划分所造成的歧义问题。

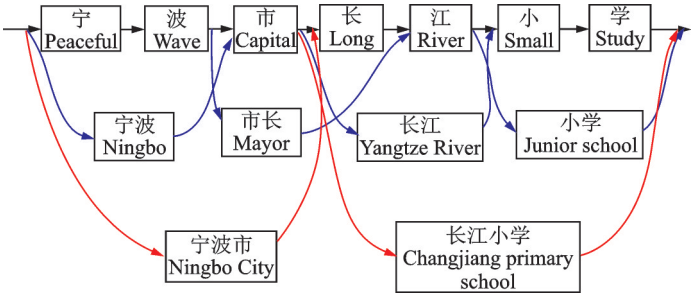

图3 格子结构的长短时记忆网络模型示意图

Fig.3 Structure of Lattice LSTM

1.2 经典相似度计算方法

最常用的文本相似度的计算方法有4种，分别是编辑距离、马氏距离、余弦相似度和皮尔逊相关系

数,前两种是通过计算文本之间的距离进行比较,距离越大,相似度越小;后两种直接计算相似度,数值越大,则相似度越大。

(1)编辑距离(Edit distance,ED),亦被称作Levenshtein distance。编辑距离实际上是一个计算序列间相似度的度量标准,通常用在检查英语单词拼写正误上,是指在两个字符串$<A,B>$之间,从A到B所需要进行的最少的编辑操作次数。其可进行的操作有且仅有3种,分别是:插入(Insert)、修改(Delete)和替换(Replace)。

以单词"kitchen"和"situate"为例,要把"kitchen"转化为"situate"需要进行的编辑操作有:(1) kitchen变为sitchen(把"k"换成"s");(2)sitchen变成sit(把"chen"删除掉);(3)sit变成situate(把"sit"插入字符"uate")。因此,将"kitchen"变成"situate"需要3步编辑操作,则这两单词的编辑距离就是3。

(2)Mahalanobis distance方法简称马氏距离。马氏距离一般用来表示某个点和某个分布间的关系,可用来计算两不同样本数据集之间的相似性,并对于不同量纲也有所考虑,即顾虑两个不同维度之间向量的相关性。

假设有M个样本向量,分别为$\boldsymbol{x}_1,\boldsymbol{x}_2,\cdots,\boldsymbol{x}_M$,其均值用$\boldsymbol{\mu}$表示,其协方差矩阵用$\boldsymbol{S}$表示,则样本向量$\boldsymbol{x}$到均值$\boldsymbol{\mu}$的马氏距离计算公式为

$$D_{\mathrm{M}}(\boldsymbol{x})=\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}S^{-1}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{1}$$

若$\boldsymbol{S}$为单位矩阵,则马氏距离变成欧氏距离,即

$$D_{\mathrm{M}}(\boldsymbol{x})=\sqrt{(\boldsymbol{x}-\boldsymbol{\mu})^{\mathrm{T}}(\boldsymbol{x}-\boldsymbol{\mu})} \tag{2}$$

若$\boldsymbol{S}$是对角矩阵,则公式变成欧氏距离的标准化表示形式,即

$$D_{\mathrm{M}}(\boldsymbol{x})=\sqrt{\sum_{i=1}^{n}\frac{(x_i-\mu_i)^2}{\sigma_i^{\ 2}}} \tag{3}$$

如果需要计算两个点$\boldsymbol{x}$、$\boldsymbol{y}$之间的马氏距离,则其计算公式为

$$D_{\mathrm{M}}(\boldsymbol{x})=\sqrt{(x-y)^{\mathrm{T}}\boldsymbol{S}^{-1}(x-y)} \tag{4}$$

马氏距离可被看成特殊的欧氏距离,与其不同的是,马氏距离的计算必须建立在协方差矩阵存在的基础上,这就要求总体样本的数目必须比样本的维数要大,且总体的样本数对其影响较大;其次,由于协方差矩阵不太稳定,导致有时马氏距离无法正确计算得出,而且易对产生细微变化的变量进行夸大,导致影响整个计算过程。

(3)基于余弦相似度(Cosine similarity)的计算方法是指通过计算得出两向量间夹角的余弦值,从而计算其相似度的方法,又被称为余弦相似性。其向量间的夹角越小,余弦值就越大,则证明两个向量越相似。在计算相关文本及字符串的相似度之前,必须把两个文本数据或者字符串统一变成向量的形式,一般通过word2vec等方法进行处理。

若存在两个二维向量$\boldsymbol{C},\boldsymbol{D}$,向量$\boldsymbol{C}$为$(x_1,y_1)$,向量$\boldsymbol{D}$为$(x_2,y_2)$,则其夹角$\theta$的余弦值计算公式为

$$\cos\theta=\frac{\boldsymbol{C}\cdot\boldsymbol{D}}{\|\boldsymbol{C}\|\times\|\boldsymbol{D}\|}=\frac{(x_1,y_1)\cdot(x_2,y_2)}{\sqrt{x_1^2+y_1^2}\times\sqrt{x_2^2+y_2^2}} \tag{5}$$

若$\boldsymbol{C},\boldsymbol{D}$为$n$维向量,则其夹角$\theta$的余弦值计算公式为

$$\cos\theta=\frac{\sum_{i=1}^{n}(x_i\cdot y_i)}{\sqrt{\sum_{i=1}^{n}x_i^2}\times\sqrt{\sum_{i=1}^{n}y_1^2}} \tag{6}$$

(4)基于皮尔逊相关系数(Pearson correlation)的方法可看做是余弦相似度的一个特例,取值范围是[-1,1]。

该方法是用来表示向量间相关性的强弱程度的,通过将其中心化,即减去向量的平均值后,再计算余弦相似度。该方法的计算是通过对分布中样本点的标准分数进行均值估计,使用$p(X,Y)$用来表示皮尔逊相关系数,公式为

$$p(X,Y)=\frac{\operatorname{cov}(X,Y)}{\sigma_X\sigma_Y}=\frac{E[(X-\mu_X)(Y-\mu_Y)]}{\sigma_X\sigma_Y}=\frac{\sum_{i=1}^{n}(X_i-\overline{X})(Y_i-\overline{Y})}{(n-1)\sigma_X\sigma_Y}=\frac{1}{n-1}\sum_{i=1}^{n}\frac{(X_i-\overline{X})}{\sigma_X}\frac{(Y_i-\overline{Y})}{\sigma_Y} \tag{7}$$

式中:X_i,Y_i分别代表两个样本;$\overline{X},\overline{Y}$分别代表两个样本的平均值;$\sigma_X,\sigma_Y$分别代表两个样本的标准差;$\frac{X_i-\overline{X}}{\sigma_X},\frac{Y_i-\overline{Y}}{\sigma_Y}$分别代表两个样本的标准分数。

2 改进方法

2.1 Lattice Bi-LSTM模型

文献[10]中提出的模型是在字向量的基础上同时考虑字粒度和词粒 度,进而来处理输入的数据,但是该模型只能单向的对句子进行处理,无法考虑整个句子的含义,对于某些需要同时考虑前后文关系的问题,无法给出正确答案。针对这个问题,本文对模型进行改进,采用了双向的长短时记忆神经网络,使得LSTM在处置信息时,能够同时进行前向传播和后向传播,使得在处理某些长句时,同时获取过去和未来两个状态的信息并对其进行通盘考虑,从而输出更具完整性和更具准确性的信息,对于应该正确输出的信息更具鲁棒性。应用于NER领域时,其效果则体现在对于实体的标签预测更具准确性和稳定性,从而对于命名实体的识别将具有更好的效果,其模型图如图4所示。

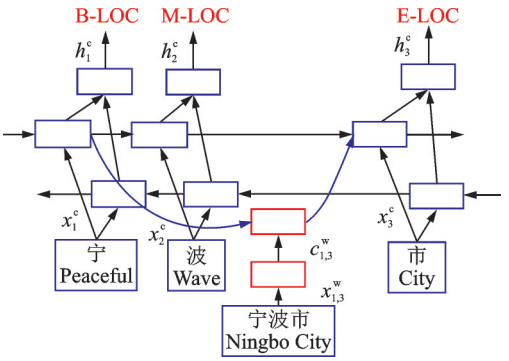

图4 双向格子结构的长短时记忆示意模型图

Fig.4 Structure of Lattice Bi-LSTM

该模型在处理字和词时的内部结构略有不同,处理单个字符时的模型如图5所示。假设需要处理一个字符序列$c_1,c_2,c_3,\cdots,c_a$,通过$\boldsymbol{x}_i^c=\boldsymbol{e}^c(c_i)$可以得到每个字符的字符向量$\boldsymbol{x}_j^c$,即输入向量。字符部分的计算公式为

$$\begin{bmatrix}i_i^c\\o_i^c\\f_i^c\\\tilde{c}_i^c\end{bmatrix}=\begin{bmatrix}\sigma\\\sigma\\\sigma\\\tanh\end{bmatrix}\left(W^{c^{\mathrm{T}}}\begin{bmatrix}x_i^c\\h_{i-1}^c\end{bmatrix}+b^c\right) \tag{8}$$

$$c_i^c=f_i^c\odot c_{i-1}^c+i_i^c\odot\tilde{c}_i^c \tag{9}$$

$$\overrightarrow{h}_i^{\mathrm{c}}=\overleftarrow{h}_i^{\mathrm{c}}=o_i^{\mathrm{c}}\odot\tanh(c_i^{\mathrm{c}}) \tag{10}$$

$$h_i^{\mathrm{c}}=\left[\overrightarrow{h}_i^{\mathrm{c}};\overleftarrow{h}_i^{\mathrm{c}}\right] \tag{11}$$

式中:h_{i-1}^{c}表示前上一个字LSTM cell的隐藏层状态输出;$\overrightarrow{h}_i^{\mathrm{c}}$,$\overleftarrow{h}_i^{\mathrm{c}}$表示两个方向的输出,$h_i^{\mathrm{c}}$为结合两个方向的最后的输出,此处的隐藏状态是对于两个方向的考虑,即为本文提出双向模型的部分体现;c_{i-1}^{c}表示从前一个字和该字相关的词传过来的细胞状态;i_i^{c},o_i^{c},f_i^{c}分别表示这个LSTM单元中的输入门、输出门和遗忘门;σ,tanh分别表示激活函数sigmoid函数和tanh函数;$\odot$表示矩阵点积。

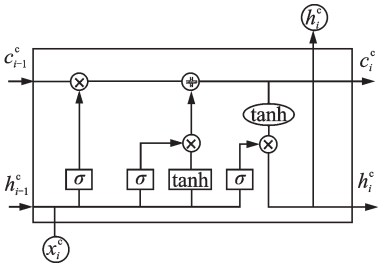

图5 基于字向量的处理模型内部结构图

Fig.5 Internal structure diagram of processing model based on character vector

处理词的模型如图6所示,将序列S和单词查找树进行匹配,可得到这个序列的词集合,表示为$w_{b,e}^{d}$,则其计算公式为

$$x_{b,e}^{\mathrm{w}}=e^{\mathrm{w}}(w_{b,e}^{\mathrm{d}}) \tag{12}$$

$$\begin{bmatrix}i_{b,e}^{\mathrm{w}}\\f_{b,w}^{\mathrm{w}}\\\tilde{c}_{b,e}^{\mathrm{w}}\end{bmatrix}=\begin{bmatrix}\sigma\\\sigma\\\tanh\end{bmatrix}\left(W^{\mathrm{w}^{\mathrm{T}}}\begin{bmatrix}x_{b,e}^{\mathrm{w}}\\h_b^{\mathrm{c}}\end{bmatrix}+b^{\mathrm{w}}\right) \tag{13}$$

$$c_{b,e}^{\mathrm{w}}=f_{b,e}^{\mathrm{w}}\odot c_b^{\mathrm{c}}+i_{b,e}^{\mathrm{w}}\odot\tilde{c}_{b,e}^{\mathrm{w}} \tag{14}$$

式中:$w_{b,e}^{\mathrm{d}}$表示从b开始到e结束的词的子序列;h_b^{c}表示第b个字的LSTM cell的输出;$i_{b,e}^{\mathrm{w}}$,$f_{b,w}^{\mathrm{w}}$分别表示这个词的网络单元中的输入门和遗忘门;$\tilde{c}_{b,e}^{\mathrm{w}}$相当于一个中间信息状态,承载着经过变化后的细胞状态信息。

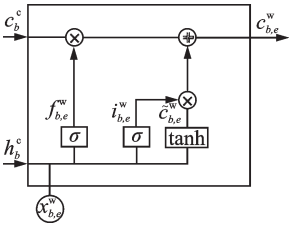

图6 基于词向量的处理模型内部结构图

Fig.6 Internal structure diagram of processing model based on word vector

从图6中可以看出,处理词的长短时记忆网络单元没有输出门,这是由于处理词的LSTM单元中的细胞状态,都传给了这个词最后一个字的字的LSTM单元。除此之外,字符LSTM单元的输入不仅来自于上一个字符的隐藏状态和字符向量,还包括前面多个词的LSTM单元输出的细胞状态。因此该模型的相关状态及其权重的计算公式为

$$c_j^{\mathrm{c}}=\sum_{b\in\left\{b'\mid w_{b'}^{\mathrm{d}}\in D\right\}}\alpha_{b,j}^{\mathrm{c}}\odot c_{b,j}^{\mathrm{w}}+\alpha_j^{\mathrm{c}}\odot\tilde{c}_j^{\mathrm{c}} \tag{15}$$

$$\alpha_{b,j}^{\mathrm{c}}=\frac{\exp(i_{b,j}^{\mathrm{c}})}{\exp(i_j^{\mathrm{c}})+\sum\limits_{b'\in\left\{b''\mid w_{b''}^{\mathrm{d}},j\in D\right\}}\exp(i_{b'},j)} \tag{16}$$

$$\alpha_j^{\mathrm{c}}=\frac{\exp(i_j^{\mathrm{c}})}{\exp(i_j^{\mathrm{c}})+\sum\limits_{b'\in\left\{b''\mid w_{b''}^{\mathrm{d}},j\in D\right\}}\exp(i_{b'}^{\mathrm{c}},j)} \tag{17}$$

式中:$c_{b,j}^{\mathrm{w}}$为上一个词的细胞状态,$\alpha_{b,j}^{\mathrm{c}}$为其状态的权重。

以句子“宁波市长江小学”为例,其中c_7^{c}“学”的细胞状态,输入量包含x_7^{c}(学)、$c_{6,7}^{\mathrm{c}}$(小学)、$c_{4,7}^{\mathrm{c}}$(长江小学)的信息,所以有

$$c_j^{\mathrm{c}}=\alpha_7^{\mathrm{c}}\odot\tilde{c}_{\mathrm{j}}^{\mathrm{c}}+\alpha_{6,7}^{\mathrm{c}}\odot c_{6,7}^{\mathrm{c}}+\alpha_{4,7}^{\mathrm{c}}\odot c_{4,7}^{\mathrm{c}} \tag{18}$$

$$\alpha_7^{\mathrm{c}}=\frac{\exp(i_7^{\mathrm{c}})}{\exp(i_7^{\mathrm{c}})+\exp(i_{6,7}^{\mathrm{c}})+\exp(i_{4,7}^{\mathrm{c}})} \tag{19}$$

$$\alpha_{4,7}^{\mathrm{c}}=\frac{\exp(i_{4,7}^{\mathrm{c}})}{\exp(i_7^{\mathrm{c}})+\exp(i_{6,7}^{\mathrm{c}})+\exp(i_{4,7}^{\mathrm{c}})} \tag{20}$$

$$\alpha_{6,7}^{\mathrm{c}}=\frac{\exp(i_{6,7}^{\mathrm{c}})}{\exp(i_7^{\mathrm{c}})+\exp(i_{6,7}^{\mathrm{c}})+\exp(i_{4,7}^{\mathrm{c}})} \tag{21}$$

对于本文提出的模型，一般需要在之后添加CRF层进行标签预测，对输入数据进行标注处理后，完成命名实体识别的识别任务，其具体流程如图7所示。

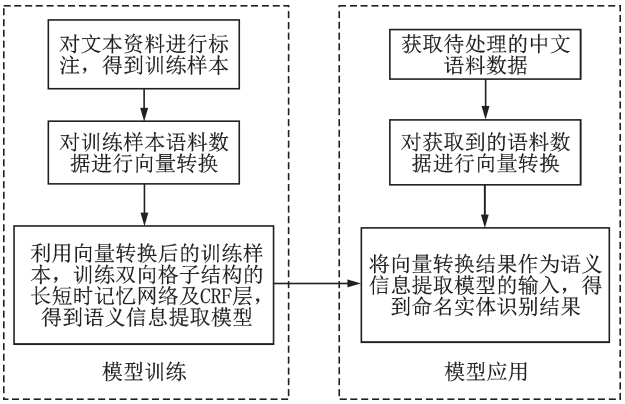

图7 模型应用流程图

Fig.7 Model application flow chart

2.2 基于核函数的相似度计算方法

使用核函数的原因如下：(1)不受非线性变换函数的形式影响；(2)改变核函数的不同形式和不同参数，能实现不同种类的核函数，实现不同的功能；(3)核函数还可以与其他算法结合，形成复合方法，实现更多功能；(4)解决了“维度灾难”的问题，对于高维度的输入能够高效处理，从而使得在使用核函数的方法时减少了计算量。

本文尝试利用几种不同的核函数来计算文本的相似度，分别是高斯核函数、马顿核函数、γ指数的核函数以及最终选用的周期核函数，下面将依次对这几类核函数进行介绍。高斯核函数沿径向对称，一般是指从输入样本到样本中心的径向距离，又被称为径向基函数(Radial basis function, RBF)，通过该函数可以将输入数据映射到无穷维，其表达式为

$$K(\boldsymbol{x},\boldsymbol{x}')=\exp\left(-\frac{\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2}{2\sigma^2}\right) \tag{22}$$

式中：exp指e的次方运算；$\|\boldsymbol{x}-\boldsymbol{x}'\|_2^2$为两个向量的二范数运算，可以看作两向量之间欧氏距离的平方；σ相当于一个宽度参数，可以自由调节，能够用来调整函数的作用范围。

径向基核函数有以下优点：(1)对于非线性函数能够将其映射到特征空间；(2)参数较少，训练时较简单，能节省训练时间；(3)计算更简单，能够减少计算量。

马顿核函数(Matérn kernel function)的内核是固定的，相当于径向基核函数的泛化表示，其原理公式为

$$K(r)=\frac{2^{1-v}}{\Gamma(v)}\left(\frac{\sqrt{2v}\,r}{l}\right)^{v}K_{v}\left(\frac{\sqrt{2v}\,r}{l}\right)\quad l,v>0 \tag{23}$$

式中:$r=\left\|\boldsymbol{x}_1-\boldsymbol{x}_2\right\|$;$l$、$v$为该核函数的超参数,$v$决定函数的可导性与平滑程度,并且当$v\to\infty$时,这个Matérn核函数就变成了使用$l$作为超参数的径向基核函数;$K_v$为修正后的贝塞尔函数,表示Matérn核函数由指数函数与多项式函数的乘积组合而成。

γ指数的核函数的基本公式为

$$K(r)=\exp\left(-\frac{r}{l}\right)^{\gamma}\quad 0<\gamma\leqslant 2 \tag{24}$$

当该γ指数核函数的指数取1时,则式(24)变为

$$K(r)=\exp\left(-\frac{r}{l}\right) \tag{25}$$

此时,式(25)就是前述的Matérn核函数中超参数v=0.5时的形式,这时运行的过程被称作Ornstein-Uhlenbeck过程,即是一个连续但不平滑的随机过程。

本文用到了一种周期性的随机函数,该函数由MacKay[11]提出,其公式为

$$K(x,x')=\exp\left(-\frac{2\sin^2\left(\dfrac{x-x'}{2}\right)}{l^2}\right) \tag{26}$$

由式(26)可以看出,该公式是关于正余弦函数的,其中,$2\sin^2\left(\dfrac{x-x'}{2}\right)$是通过公式$(\cos x-\cos x')^2+(\sin x-\sin x')^2=4\sin^2\left(\dfrac{x-x'}{2}\right)$演变得到的。而式(26)实际上也是一个指数型核函数,其中$\gamma=2$,在进行训练时,只需改变超参数l的大小进行对比测试。

3 实验结果分析

3.1 NER实验结果分析

本文数据集主要采用weiboNER、resumeNER、CMNER。weiboNER是从新浪微博上进行采集的,包含多类信息;resumeNER是新浪经济类的数据,包含中国上市公司高管的简历;CMNER是CCKS2017的中文医学命名实体识别数据集,包括多类实体,比如身体部位、症状体征、检查和疾病名等。

本文主要通过4项指标来测试模型的有效性,分别是准确率(Accuracy, acc)、精确率(Precision, pre)、召回率(Recall, rec)和F_1-Measure。准确率是指在所有测试的数据集中,正确识别出的语料除以所有语料总数之值;精确率是指在所有正确识别出的语料中,实际正确识别的语料除以所有正确识别的语料的数值;回归率是指在所有识别为正确的语料中,实际能够识别出正确语料的比例;F_1由精确率和召回率得到,计算公式为$F_1=2*P*R/(P+R)$,P代表精确率,R代表召回率。

不同模型在数据集resumeNER上的最佳表现如表1所示。训练模型主要包括4类,分别是LSTM+bigram、LSTM+unigram、本文模型(双向格子LSTM)+bigram和本文模型(双向格子LSTM)+unigram,其中,bigram和unigram代表两种分词方式,分别是二元分词(将句子每两个字切分一次)和一元分词(将句子每一个字切分一次)。从表1可以看出,应用该模型在该数据集上采用两种分

表1 不同模型在resumeNER数据集上的最佳表现

Table 1 The best performance of different models on the resumeNER dataset

指标	LSTM+ bigram	LSTM+ unigram	本文模型+ bigram	本文模型+ unigram
acc	0.960 187	0.955 436	0.962 347	0.961 915
pre	0.924 717	0.920 188	0.926 813	0.934 007
rec	0.927 188	0.916 5	0.930 528	0.926 52
F_1	0.925 951	0.918 34	0.928 667	0.930 248

词方式的表现均比使用LSTM的效果好,对于分别使用unigram和bigram分词方式时,与LSTM相比,应用该模型F_1分数分别提升了0.27%和2.60%,其余4类指标均得到了有效提升,并且可以看出,此时采用bigram分词方式时效果最好。总之,该模型在resumeNER数据集上的效果比LSTM模型好。

不同模型在数据集weiboNER上的最佳表现如表2所示。从表2可以看出,应用该模型在该数据集上采用两种分词方式的表现均比使用LSTM的效果好,对于分别使用unigram和bigram分词方式时,与LSTM相比,应用该模型F_1分数分别提升了8.6%和4.7%,其余3类指标也得到了有效提升,acc,pre和rec最高分别提升了0.4%、6.1%、10.7%。可以看出,该模型在weiboNER数据集上的效果十分突出。

表2 不同模型在数据集weiboNER上的表现

Table 2 The best performance of different models on the weiboNER dataset

指标	LSTM+ bigram	LSTM+ unigram	本文模型+ bigram	本文模型+ unigram
acc	0.957 771	0.955 209	0.958 671	0.959 086
pre	0.578 313	0.582 317	0.594 203	0.620 29
rec	0.493 573	0.491 003	0.526 992	0.550 129
F_1	0.532 594	0.532 775	0.558 583	0.583 106

不同模型在数据集CMNER上的最佳表现如表3所示。可以看出,在该数据集上的效果没有前两个数据集明显,分别使用bigram和unigram分词方式时,与经典模型相比,其精确率和F_1得分都略有下降,但准确率和召回率均为使用该模型时最高,分别提升了0.02%和0.50%。该模型在CMNER数据集效果不太明显,仅有两项指标效果有所提升,可能是因为该数据集中的实体多为类似“胸部正位DR片”等检查项目类的实体,名称比较复杂,难以辨认,导致建立的格子词典的作用没有发挥出来,因此格子结构没有取得更优异的效果。

表3 不同模型在数据集CMNER上的表现

Table 3 The best performance of different models on the CMNER dataset

指标	LSTM+ bigram	LSTM+ unigram	本文模型+ bigram	本文模型+ unigram
acc	0.990 078	0.983 278	0.990 299	0.982 465
pre	0.973 988	0.948 673	0.966 418	0.945 019
rec	0.963 916	0.951 595	0.968 757	0.945 435
F_1	0.968 926	0.950 132	0.967 586	0.945 227

3.2 相似度计算实验结果分析

首先,采用语义相同的两句话进行测试,分别是“嗓子疼怎么办”和“嗓子疼咋办”,分别使用基于高斯核函数、基于指数核函数和周期性随机函数的方法来计算文本的相似度,并与其他经典方法进行对比。此时的高斯核函数中的超参数$\sigma^2=1$,指数核函数中的超参数$\gamma=1,l=1$,相当于马顿核函数中的超参数$v=0.5,\gamma=1,l=1$。其相似度(距离)对比图如图8所示。

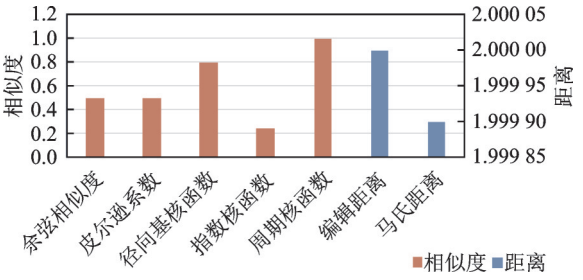

图8 结构近似句子相似度对比图

Fig.8 Comparison of sentence similarity of similar structure

从图8中可以看出,在计算两个结构相似语义相同的句子时,本文中使用的两种核函数(径向基核函数和周期核函数)所得的相似度值均较大,均大于经典方法计算出的相似度值,且均突破了0.75,其中本文用到的周期核函数效果最为显著,其相似度计算为1.0,是经典方法余弦相似度的两倍,说明其计算相似度值的准确率提高了一半,能够完美得出计算的两个句子语义相同的结论,应用在问答系统中时,相比其他方法,能够更好地得出用户提出问题的真正意图,从而更好地输出答案。

图9为几种方法的相似度对比图,从图中可以看出,在计算这两个长句的相似度时,皮尔逊系数、径向基核函数和指数函数计算的相似度均不理想,均未达到0.1,不能得出长句相似的结论。然而,本文采用的周期核函数所得相似度为0.891 4,远大于余弦相似度的0.517 1,能够完美得出这两长句语义相似的结论,因此能够得出同样的答案。

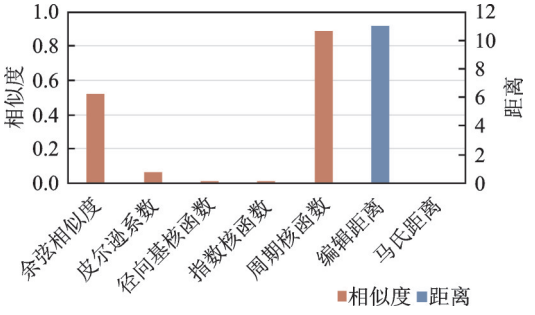

图9 长句相似度对比图

Fig.9 Comparison of long sentence's similarity

3.3 综合效果评测

上述两小节分别对两个创新环节进行了单独测试,本小节将创新工作合并形成一套完整的系统,对整体性能进行评估。分别以例句“我好像得了尘螨过敏性哮喘,有啥法子啊”和“最近一直在打喷嚏,怀疑是尘螨过敏性哮喘,应该咋办”进行实际效果展示,其结果如表4所示。对于前者,采用经典方法时,系统无法识别出“有啥法子啊”与“怎么办”的相似度,因此无法给出“尘螨过敏性哮喘”的治疗方法,只给出了几种疾病的治愈率,采用本文方法时,则给出了一系列关于该疾病的治疗方法,效果显著;对于后者,由其结果对比可知,采用基于核函数的方法可得出该句与前句语义相似,从而得出相同的答案,能够正确处理用户的询问信息。

对于经典方法与本文方法询问不同问题类型,其得到的结果如表5所示。可以看出,经典方法有时在回答某些问题,比如疾病的治疗方法、检查类型和种类时,其回答结果分别是治愈周期、某疾病的概率和治愈周期,并非问题的正确结果,而本文的核函数方法则可以回答出问题本该得出的结果,效果显著,能够正确回答出问题。

表4 应用于问答系统的实际效果

Table 4 Practical effect of application in Q & A system

问题	方法类型	机器回答效果
我好像得了尘螨过敏性哮喘,有啥法子啊	经典方法(余弦+ED)	疾病过敏的治愈率为:60%
我好像得了尘螨过敏性哮喘,有啥法子啊	本文方法	疾病尘螨过敏的治疗方法有:药物治疗、脱敏治疗;可用药品包括:氯雷他定糖浆
最近一直在打喷嚏,怀疑是尘螨过敏性哮喘,应该咋办	经典方法(余弦+ED)	疾病尘螨过敏的治愈率为:75%
最近一直在打喷嚏,怀疑是尘螨过敏性哮喘,应该咋办	本文方法	疾病尘螨过敏的治疗方法有:药物治疗、脱敏治疗;可用药品包括:氯雷他定糖浆

表5 不同方法结果判定

Table 5 Result judgment of different methods

方法类型	回答结果	问题类型	是否解答
经典方法	治愈周期/治愈率	疾病治疗方法	否
	为某疾病的概率	疾病检查类型	否
	治愈周期	疾病的种类	否
本文方法	治疗方法	疾病治疗方法	是
	疾病的检查项目	询问疾病检查类型	是
	某种疾病的概率	疾病的种类	是

4 结束语

本文针对问答系统中命名实体识别技术处理句子不完善的问题,提出了一种双向格子结构的长短时记忆神经网络(Lattice Bi-LSTM)模型,解决了NER中基于字向量模型所存在的对句子处理不当的问题,同时解决了基于词向量模型所具有的对分词效果依赖严重的问题;由于采用的是双向结构,与单向相比,能够更好地理解句子的含义,输出结果更具鲁棒性,并且能够增进对上下文内容的理解。通过在数据集上的测试,也表明该方法具有比单向结构更好的效果,能够对句子进行更好的处理。

问答系统在进行命名实体识别后,需要对识别出的实体与知识库中的实体进行相似度计算,本文提出一种将周期性核函数用于相似度计算的方法,并与其他经典方法进行了对比。结果显示,对于相同语义和相似语义的句子计算出的相似度比其他方法高,能够更好地识别出两个实体之间的相似度,使提出的问题能更准确地链接到知识库中的答案,从而提高了问答系统回答问题的准确率。

参考文献:

[1] 冯升. 聊天机器人问答系统现状与发展[J]. 机器人技术与应用, 2016 (4): 34-36.
FENG Sheng. Status and development of chat robot question answering system[J]. Robot Technology and Application, 2016 (4): 34-36.

[2] BORIS Katz. START natural language system[EB/OL]. http://start.csail.mit.edu/index.php. 1993.

[3] 北齐安. 苹果Siri:与众不同的"语音助手"[J]. 创新时代, 2012 (2): 63-64.
BEI Qian. Apple Siri: A different "voice assistant" [J]. The Age of Innovation, 2012(2): 63-64.

[4] LIN J, KATZ B. Question answering from the web using knowledge annotation and knowledge mining techniques[C]//

Proceedings of the International Conference on Information and Knowledge Management. New Orleans, USA: ACM, 2003: 116-123.

[5] 向晓雯,史晓东,曾华琳. 一个统计与规则相结合的中文命名实体识别系统[J]. 计算机应用, 2005, 25(10): 2404-2406.
XIANG Xiaowen, SHI Xiaodong, ZENG Hualin. Chinese named entity recognition system using statistics-based and rules-based method[J]. Journal of Computer Applications, 2005, 25(10): 2404-2406.

[6] AJEES A P, IDICULA S M. A named entity recognition system for Malayalam using conditional random fields[C]// Proceedings of International Conference on Data Science and Engineering. Kochi, India: [s.n], 2018: 1-5.

[7] GUO Y, GAO H. A Chinese Person name recognition system based on agent-based HMM position tagging model[C]// Proceedings of the 6th World Congress on Intelligent Control and Automation. Dalian, China: [s.n], 2006: 4069-4072.

[8] LIN X, PENG H, LIU B. Chinese named entity recognition using support vector machines[C]//Proceedings of International Conference on Machine Learning and Cybernetics. Dalian, China: [s.n], 2006: 4216-4220.

[9] EKBAL A, SAHA S, HASANUZZAMAN M. Multiobjective approach for feature selection in maximum entropy based named entity recognition[C]//Proceedings of the 22nd International Conference on Tools with Artificial Intelligence. Arras, France: [s.n], 2010: 323-326.

[10] ZHANG Y, YANG J. Chinese NER using lattice LSTM[J]. Computation and Language, 2018, 1(4): 1554-1564.

[11] MACKAY J C. Introduction to Gaussian processes[J]. NATO ASI Series F Computer and Systems Sciences, 1998, 168(1): 133-166.

作者简介:

鲍静益(1984-),女,讲师,研究方向:人工智能与模式识别,E-mail:baojy@czu.cn。

于佳卉(1996-),女,硕士研究生,研究方向:人工智能与模式识别,E-mail:815720839@qq.com。

徐宁(1981-),男,副教授,研究方向:人工智能与模式识别,E-mail:20101832@hhu.edu.cn。

姚潇(1982-),男,副教授,研究方向:人工智能与模式识别,E-mail:20141925@hhu.edu.cn。

刘小峰(1974-),男,教授,研究方向:人机交互与自适应导航,E-mail:20111842@hhu.edu.cn。

(编辑:王静)