计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2020,29(10):44−52 [doi: 10.15888/j.cnki.csa.007671]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于网格 LSTM 混合算法的地质领域用户意图识别①

贺金龙[1,2], 付立军[1,3], 姚　郑[1,2], 吕鹏飞[4], 黄徐胜[1,3]

[1](中国科学院大学, 北京 100049)
[2](中国科学院 网络信息中心, 北京 100049)
[3](中国科学院沈阳计算技术研究所, 沈阳 110168)
[4](中国地质图书馆, 北京 100083)
通讯作者: 付立军, E-mail: fu_lijun@ucas.ac.cn

摘　要: 针对传统基于模板匹配、关键词共现、人工特征集合等方法的问答机器人存在用户意图识别耗时、费力且扩展性不强的问题, 本文结合地质领域文献中结构化知识问答的复杂特点, 使用了基于网格记忆网络 (LSTM+CRF+Lattice) 与基于卷积神经网络 (CNN) 融合的优化模型. 该模型将用户询问意图识别看作分类问题, 首先使用网格记忆网络进行文本信息的命名实体识别及关系抽取, 然后使用卷积神经网络将用户输入的其他文本信息进行属性分类, 接着将分类结果转化为满足知识图谱查询的结构化方式, 最终实现地质知识属性映射的用户询问意图识别. 实验证明, 在考虑地质知识特征的处理中, 对于准确率的提升起到了极大帮助.

关键词: 知识结构化; 询问意图; 实体识别; 属性映射

引用格式: 贺金龙,付立军,姚郑,吕鹏飞,黄徐胜.基于网格 LSTM 混合算法的地质领域用户意图识别.计算机系统应用,2020,29(10):44–52. http://www.c-s-a.org.cn/1003-3254/7671.html

User Intention Recognition in Geological Field Based on LSTM-CC Hybrid Algorithm

HE Jin-Jong[1,2], FU Li-Jun[1,3], YAO Zheng[1,2], LYU Peng-Fei[4], HUANG Xu-Sheng[1,3]

[1](University of Chinese Academy of Sciences, Beijing 100049, China)
[2](Network Information Center, University of Chinese Academy of Sciences, Beijing 100049, China)
[3](Shenyang Institute of Computing Technology, Chinese Academy of Sciences, Shenyang 110168, China)
[4](National Geological Library of China, Beijing 100083, China)

Abstract: Aiming at the time-consuming, laborious, and weak expansibility of user intention recognition in question answering robots based on template matching, keyword co-occurrence or artificial feature set, this study proposes a model based on the combination of grid memory network (LSTM+CRF+Lattice) and Convolutional Neural Network (CNN) combined with the characteristics of geological literature question answering. In this hybrid model, users' query intention recognition is regarded as a classification problem. Firstly, the grid memory network is used to identify the named entity and extract the relationship of the text information, then the CNN is used to classify the attributes of other text information input by users, and then the classification results are transformed into a structured way to meet the query of knowledge graph, and finally realizes the attribute mapping of user intention recognition. Experiments show that it is very helpful to improve the accuracy rate when considering the characteristics of geological knowledge.

Key words: knowledge structure; inquiry intention; entity recognition; attribute mapping

① 基金项目: 国家重点研发计划 (2018YFC1505501); 国土资源部大数据科研专项 (201511079-3)
Foundation item: National Key Research and Development Program of China (2018YFC1505501); Special Fund for Big Data Research of Ministry of Land and Resources of the People's Republic of China (201511079-3)
收稿时间: 2020-03-25; 修改时间: 2020-04-21; 采用时间: 2020-05-13; csa 在线出版时间: 2020-09-30

1 引言

近年来, 随着人工智能的蓬勃发展, 不同行业服务质量逐步提升, 其中最为耀眼的问答机器人得到了行业领域的充分应用, 例如微软小娜、阿里小蜜、京东 JIMI 等. 本文研究的主要内容是在地质领域问答服务中的用户意图识别, 用户意图是指用户为满足地质知识探索关联发现的需要, 通过文本表达出对相关知识的探索意愿. 在问答服务过程中, 用户会产生大量数据, 如何利用这些数据本身的特性去判别用户倾向、增强用户体验、使得问答机器人更加智能是当下研究的重要难点之一[1].

对于知识检索探求、结构化推荐、表示学习推理以及专家建议与决策, 准确识别响应用户意图尤为重要[2]. 在互联网技术的蓬勃发展过程中, 关于用户询问理解识别的研究如下: 基于 Luence、Elasticsearch 的树状分类方法来识别用户搜索内容的归属类别[3]、基于人工构建类别的正则匹配规则与图的方法来抽取和泛化用户意图[4] 及考虑到用户意图语料匮乏的跨领域迁移学习方法[5] 等. 这些基于人工构建匹配规则查询和引入路径优化探索的方法在应用中都存在一定的局限性, 前者是通过挖掘用户询问语句是否与预先设定的方式模板相匹配, 得到匹配度满足阈值的知识, 后者是通过文本的二元、三元、及多元特征作为分类特征, 使用集成学习的方式在多个特征分类器中训练得到最佳的意图判断. 上述方法都存在限定的泛化能力, 没有很好地理解文本的深层语义信息的问题, 从而导致识别用户的真实意图方面较弱.

针对以上问题, 本文采用地质领域文献数据知识关联特征与文本语义信息相结合的方式将用户意图识别看作文本分类问题, 使用了基于网格记忆网络 (LSTM+CRF+Lattice) 与基于卷积神经网络 (CNN) 融合的模型, 不仅很好的捕捉文本深层语义信息, 而且在文本问答过程中能快速识别用户意图. 该混合模型首先使用网格记忆网络进行用户文本信息的命名实体识别及关系抽取, 然后使用卷积神经网络将用户输入的除实体外其他文本信息进行属性分类, 再将分类结果转化为满足知识图谱查询的结构化方式, 最终将知识图谱的节点关联性通过结构化语言 Cypher 实现属性映射的用户意图识别.

实验结果表明, 在地质领域问答的用户意图识别任务中, 本文采用的网格 LSTM 与 CNN 的混合模型较传统的人工规则匹配与机器学习方法, 可以有效地识别用户问答过程中的意图.

2 相关工作

本文主要研究在地质领域中的用户检索意图识别. 针对地质文献中构建的关联知识, 用户以简洁的自然语句进行询问, 具体的用户询问形式如表 1 所示.

表 1 用户检索问题描述

序号	问题描述
1	青藏高原的简介是什么
2	青藏高原的别名是什么
3	青藏高原的区域范围是什么
4	青藏高原与火山机构存在什么关系
5	青藏高原与研究的知识推理是什么

这里针对用户提出的知识询问, 我们的意图识别处理如图 1 所示, 首先根据用户的自然语句进行语义解析, 其中包括两部分: 一部分对于语句中的命名实体识别, 一部分是对于语句信息的属性分类, 然后将分类结果映射到相应的用户意图类型中, 通过转化的结构化查询得到用户意图结果.

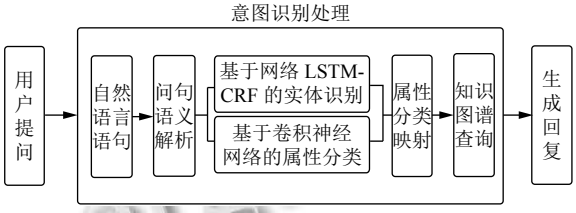

图 1 用户意图识别流程图

2.1 命名实体识别

命名实体识别 (Named Entity Recognition, NER) 是 Web 2.0 向 3.0 转化的一种深度学习技术工具, 是文本挖掘过程中基于句法分析理解的重要手段, 在大数据量的人工智能发展中, 基于数据的处理能力尤为最要[6].

英文文本词与词之间以空格分隔, 方便计算机识别, 但是中文以字为单位, 句子所有字连起来才能表达一个完整的意思. 如英文“China geological library builds big data intelligent platform”, 词与词之间有空格进行隔开, 而对应的中文“中国地质图书馆建设大数据智能化平台”, 句子中的词由多个独立的汉字组成并且字与字之间没有任何分割标记符, 同时还可能存在交集歧义、组合歧义、未登录词等特征, 所以中文的命名实体识别比英文的实体识别复杂的多. 目前, 随着技术的

不断革新, 中文命名实体识别也经历了 3 个阶段的研究. 第 1 阶段基于人工词典的规则匹配方法[7], 主要采用专家手工构建检索规则、模板, 以字符串相匹配为主要手段, 这也导致一定的局限性. 第 2 阶段基于二元、多元统计的方法, 利用人工标注数据作为训练基础学习文本特征, 对于机器学习不需要人工设定规则且线上识别可扩展性强. 这也是目前应用方式最多的技术, 如在满足已知约束的条件集合的概率学习时, 选择熵最大的模型[8]、在判断线性可分与否的感知器模型[9]、对于时序随机序列的状态转移概率计算的隐马尔可夫模型[10]、以及用于预测与输入标注序列相对应的模型等. 条件随机场 (Conditional Random Fields, CRF) 解决了句子特征参数选择优化与标记偏置问题, 是统计模型中应用最为广泛的一种模型[11]. 文献 [12] 提出基于 CRF 与地理词典规则结合的识别方法. 随着深度学习的兴起, 研究者不断注重对于时序数据上下文信息的捕捉, 提出循环神经网络 (Recurrent Neural Network, RNN), 利用当前数据的输出作为到下一个神经元的输入捕捉隐藏层特征信息[13]. 但当进行长距离特征信息捕捉时会出现梯度消失或者爆炸的问题, 基于此问题, 提出了通过“门”结构的网络细胞单元进行控制信息流转的输入、更新与删除的长短时记忆网络 (Long-Short Term Memory, LSTM). 另外, 文献 [14] 提出使用 4 种类型特征的 LSTM-CRF 模型, 分别是拼写特征, 内容特征, 词语向量和词典特征, 其实验表明这些额外的特征可以提高标签的准确率. 考虑 LSTM 与 CRF 在实体识别中存在模型互补的优势, 将二者相结合的训练模型不断出现[15,16].

在以上研究的基础上, 本文尝试使用序列标注中 BIOES 标签与改进的基于网格的双向 LSTM 相结合的方式进行地质领域中命名实体识别, 包括以下 14 种类型, 也是我们的创新点应用, 如图表 2 所示.

2.2 文本信息分类

自 20 世纪 80 年代初起, 文本分类在经历基于词匹配研究、基于知识工程研究后, 由于大量数字电子化数据驱使, 使得分类向机器自动学习靠拢, 使得分类技术成为数据处理的重要分支[17]. 它是按照预先定义的规则和体系, 将文本实现自动归类的过程[18], 其结构化形式定义如下:

$$c_{ij}=\begin{cases}1, & \text{文档}i\text{属于类别}j\\ 0, & \text{文档}i\text{不属于类别}j\end{cases} \quad (1)$$

表 2 特定领域数据标签类型

序号	类型	缩写
1	地质区域	GARE
2	地质化学	GEHE
3	地质作用	GEFF
4	地质实体	GENT
5	地质位置	GLOC
6	地质方法	GMET
7	地质科学	GSCI
8	生物实体	BENT
9	生物细菌	BFUN
10	生物方法	BMET
11	古生物	BPAL
12	生物植物学	BBOT
13	生物昆虫	BINS
14	实体关系	RELA

在梳理时首先对文本进行去停用词、无效符号, 接着使用分词工具对其进行文本切分, 然后使用 TextRank 等技术进行关键特征提取, 最后使用分类器等集成学习方式归类.

在数据挖掘中, 可以分为两种: 二分类器和多分类器. 本文根据地质领域数据特征及问答环节涉及的用户知识将数据划分为 9 种类型进行验证, 采用了基于字符的深度学习的知识表示进行多分类. 如表 3 所示.

表 3 领域数据问答知识分类

序号	类型	描述	英文标识
1	定义型	的定义, 定义是什么, 简介, 简介是什么	definition
2	属性标签型	青藏高原有哪些属性	label
3	类别型	类型, 哪一类, 哪一类别, 属于哪一类	type
4	关系型	图谱关系, 存在什么关系, 可能有什么关系	relation
5	实体型	相关实体, 下一节点, 实体, 下一节点是什么	entity
6	知识推理型	知识推理是什么, 知识推理节点是什么	reasoning
7	实体-实体型	青藏高原与火山机构的关系	End2End
8	实体-关系型	细菌与浸出的知识推理是什么	End2Rel
9	三元组型	青藏高原的实体对, 三元组, 实体组	triple

3 基于网格 LSTM 混合算法的意图识别模型

算法模型主要依据 LSTM 模型长期记忆特性与基于字符向量的融合构建. 在模型构建方面, 本文主要在实体识别和属性分类上引入了自定义地质知识的改进, 同时将二者联合进行研究实现.

3.1 基于网格 LSTM+CRF 命名实体识别

1) 数据标注策略

为了数据标注任务的便利性和统一标准, 本文采用中文字符作为 token, 采用最常用的 BIOES 标注规范[19] 结合类别进行字符序列标注.

2) 网格模型

在实体检测中先使用 n-gram 问题词搜索与问题具有公共子字符串的实体[20], 后使用神经网络与句法指标进行捕捉问题和实体名称之间的相似匹配. 例如, 文献 [21] 使用字符级别的 LSTM 来编码问题和实体名称; 文献 [22] 使用字符级别 CNN 来编码问题和实体; 文献 [23] 同时使用单词级别和字符级别来编码问题.

使用的 LSTM 基本模型结构如图 2 所示, 其中包含遗忘门、输入更新门、输出门 3 个门结构.

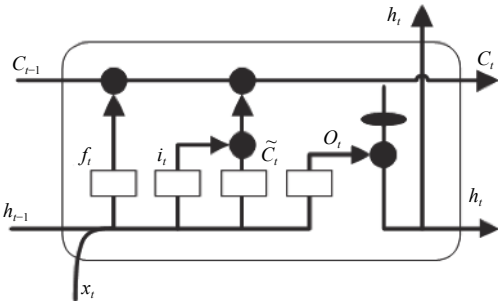

图 2 长短时记忆网络结构图

x_t 为 t 时刻输入, h_{t-1} 为 $t-1$ 时刻的输出, C_{t-1} 为 $t-1$ 时刻的细胞输出, C_t 为 t 时刻的细胞输出, f_t 为输入到 C_{t-1} 的值, i_t 为输入门向量、$\widetilde{C}_t$ 为新的候选值向量, o_t 为输出向量.

遗忘门负责决定剔除多少信息. 主要考虑 t 时刻的输入、$t-1$ 刻的输出, 其中输出值为 1 表示“完全保留该部分信息”, 输出值为 0 表示“删除这部分信息”, 计算公式如下:

$$f_t = \sigma\left(W_f * [h_{t-1}, x_t] + b_f\right) \tag{2}$$

其中, f_t 是输出到 C_{t-1} 的值, W_f, b_f 分别为遗忘门的权值与偏置.

输入与更新门决定哪些信息被存储到细胞状态中. 主要考虑两部分, 首先是针对 $t-1$ 刻的输出与 t 时刻的输入信息中的哪些信息被更新, 然后再对其转换作为更新输入, 再加入到 cell 中, 如下:

$$i_t = \sigma(W_i * [h_{t-1}, x_t] + b_t) \tag{3}$$

$$\tilde{C}_t = \tanh(W_C * [h_{t-1}, x_t] + b_C) \tag{4}$$

其中, W_i, b_t, W_C, b_C 分别为输入门与更新门的权值与偏置.

经过遗忘门、输入与更新门之后, 需要将 $t-1$ 刻的细胞状态更新到 t 时刻的 C_t 上, 主要将 C_{t-1} 细胞状态剔除遗忘信息, 再加上输入门与更新门的更新信息, 其计算公式如下:

$$C_t = f_t * C_{t-1} + i_t * \tilde{C}_t \tag{5}$$

输出门决定当前 t 时刻细胞的输出. 首先需要经过 Sigmoid 层决定 $t-1$ 刻输出与 t 时刻输入的信息哪些被输出; 然后将当前的细胞状态 C_t 送入 tanh 激活函数, 将数值范围变为 -1 到 1 之间; 最后将以上两步的输出相乘得到最终的输出, 计算公式如下:

$$o_t = \sigma(W_o [h_{t-1}, x_t] + b_o) \tag{6}$$

$$h_t = o_t * \tanh(C_t) \tag{7}$$

其中, W_o, b_o 为输出门的权值与偏置.

在此基础上, 本文提出的基于字符的网格模型与基于词的模型相比能够很好地避免分词错误带来的影响, 受启发于 LSTM+CRF 模型改进[24], 主要原因之一是采用基于字向量的模型, 其二是将领域词加入到模型中充分利用显性的词和词序信息 (比如“青藏高原”这个词如果拆成字向量就成了“青”、“藏”、“高”、“原”, 这 4 个词的单独含义明显与其组合的词的含义大相径庭).

首先, 定义一个输入句子 s, 以字为基本单位:

$$s = c_1, c_2, \cdots, c_m \tag{8}$$

其中, c_j 为 s 的第 j 个字, s 表示为:

$$s = w_1, w_2, \cdots, w_n \tag{9}$$

其中, w_i 为 s 的第 i 个词, 设 $t(i,k)$ 为句子的第 i 个词的第 k 个字在句子中的位置, 比如“青藏高原, 火山机构”这句话中的“山”字, 我们就有 $t(2,2) = 7$.

如图 3 所示, 是一个基于字序列 $c_1, c_2, \cdots, c_m$ 的模型, 其中每一个字被表示为:

$$x_j^c = e^c\left(c_j\right) \tag{10}$$

其中, e^c 为权重矩阵, 输入 $x_1, x_2, \cdots, x_m$ 都会有一个隐含状态, 即 $\overrightarrow{h}_1^c, \overrightarrow{h}_2^c, \cdots, \overrightarrow{h}_m^c$ 和 $\overleftarrow{h}_1^c, \overleftarrow{h}_2^c, \cdots, \overleftarrow{h}_m^c$, 那么隐藏层的总输出可以表示为:

$$h_j^c = \left[\overrightarrow{h}_j^c, \overleftarrow{h}_j^c\right] \tag{11}$$

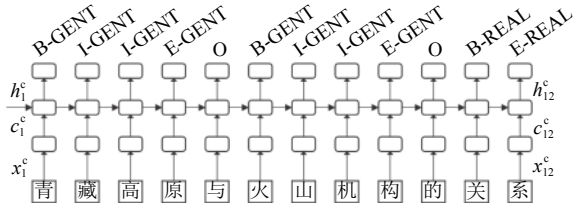

图 3 基于字符 BIOES 的序列标注图

从基于字符的模型可以看出单个字组成正确的一句话需要考虑所有路径的组合, 而路径的个数随字符个数的增长呈指数增长, 为解决这个问题, 我们引入了构建词典中的词语信息, 如图 4 中黑色圆形阴影部分, 这样就可以控制信息的始终导向, 进而提升模型效率.

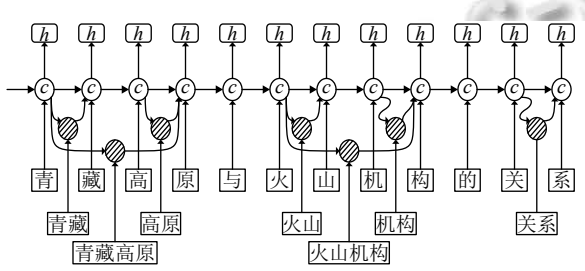

图 4 基于 LSTM 与领域知识融合图

模型的主干部分采用基于字符的双向 LSTM-CRF, 与普通 LSTM 不一样的地方在于, 模型中具有一些句子中潜在词汇的细胞信息, 同主干 LSTM 的 cell 细胞状态信息连接起来就构成了基于词的网格模型, 例如“青藏”、“高原”、“青藏高原”这三者之间的考虑. 如图 5 所示.

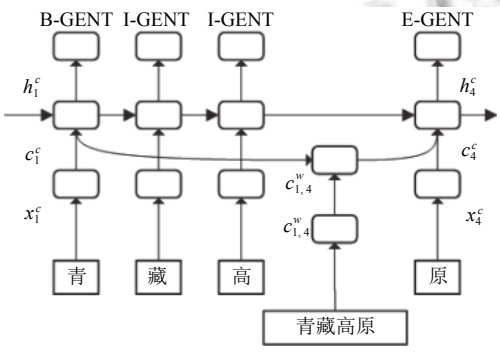

图 5 基于网格模型序列选择策略图

主干部分 LSTM 的记忆细胞内部算法如下:

$$x_j^c = e^c\left(c_j\right) \tag{12}$$

$$\begin{bmatrix} i_j^c \\ o_j^c \\ f_j^c \\ \tilde{C}_j^c \end{bmatrix} = \begin{bmatrix} \delta \\ \delta \\ \delta \\ \tanh \end{bmatrix}\left(W^{c^{\mathrm{T}}}\left[x_j^c\right] + b^c\right) \tag{13}$$

$$C_j^c = f_j^c * c_{j-1}^c + i_j^c * \tilde{C}_j^C \tag{14}$$

$$h_j^c = o_j^c * \tanh\left(C_j^c\right) \tag{15}$$

对于词汇的语义信息算法如下:

$$x_{b,e}^w = e^w\left(w_{b,e}^d\right) \tag{16}$$

$$\begin{bmatrix} i_{b,e}^w \\ f_{b,e}^w \\ \tilde{C}_{b,e}^w \end{bmatrix} = \begin{bmatrix} \delta \\ \delta \\ \tanh \end{bmatrix}\left(W^{w^{\mathrm{T}}}\left[x_{b,e}^w\right] + b^w\right) \tag{17}$$

$$C_{b,e}^w = f_{b,e}^w * C_b^c + i_{b,e}^w * \tilde{C}_{b,e}^w \tag{18}$$

其中, e^w为 embedding 矩阵, $w_{b,e}^d$中 b, e 表示为词汇的首尾字符索引.

有了词格$C_{b,e}^w$后, 并不是所有的词汇信息都需要传入当前词汇细胞, 要利用逻辑门单元 cell 来计算当前字符与历史信息的权重, 进而选取最有用的词汇.

$$i_{b,e}^c = \delta\left(W^{l^{\mathrm{T}}}\begin{bmatrix} x_e^c \\ C_{b,e}^w \end{bmatrix} + b^l\right) \tag{19}$$

$$C_j^c = \sum_{b\in\{b'|w_{b',j}^d\in D\}} \alpha_{b,j}^c * c_{b,j}^w + \alpha_j^c * \tilde{C}_j^c \tag{20}$$

最终结合主干部分, 通过当前字符状态得到中间层的输出, 再通过 CRF 做标签序列的实体识别:

$$C_j^c = \sum_{b\in\{b'|w_{b',j}^d\in D\}} \alpha_{b,j}^c * c_{b,j}^w + \alpha_j^c * \tilde{C}_j^c \tag{21}$$

$$h_j^c = o_j^c * \tanh\left(C_j^c\right) \tag{22}$$

$$P(y|s) = \frac{\exp\left(\sum_i W_{\mathrm{CRF}}^{li} hi + b_{\mathrm{CRF}}^{l_{i-1},li}\right)}{\sum_{y'}\exp\left(\sum_i W_{\mathrm{CRF}}^{li} hi + b_{\mathrm{CRF}}^{l_{i-1},li}\right)} \tag{23}$$

其中, $\alpha_{b,j}^c$, α_j^c当前字符词与字符输入的权重, $y = l_1$, $l_2,\cdots,l_T$为预测标签序列, y'为任意标签序列, W_{CRF}^{li}为针对每个 l_i 的参数, $b_{\mathrm{CRF}}^{(l_{i-1},l_i)}$为$l_{i-1}$到$l_i$的偏置.

通过以上方法最终找到概率最大的序列, 即得到最终的实体识别输出.

3.2 基于字符编码的 CNN 问句属性分类

1) 数据集构建

通过网格模型正确提出用户语句中的地质实体之后, 还需要理解用户的意图, 其具体表现为地质实体具备的知识属性, 即需要将用户询问意图与知识图谱属性进行映射, 为满足用户的询问需求, 标注了围绕地质知识自身特性及关联的结构化特征具备一般性原则的语句描述方式标签, 标签一共包括定义型、别名型、海拔型、大小型、种类型、区域范围型、地质构造型、基本组成型、关系型等 9 大类别.

2) 分类算法

针对用户询问的短文本特征, 以及 n-gram 语言模型可知, CNN 模型[25] 对于自然语言的局部语义特征提取存在优势, 因此常被用于表示句子级别的信息和短文本分类. 本文使用基于字符的 CNN 模型对用户询问的除命名实体识别外的语句进行语义表示并进行属性分类, 映射为知识图谱中的属性关系标签, 进而实现用户询问意图. 结构图如图 6. 属性分类具体方法如算法 1.

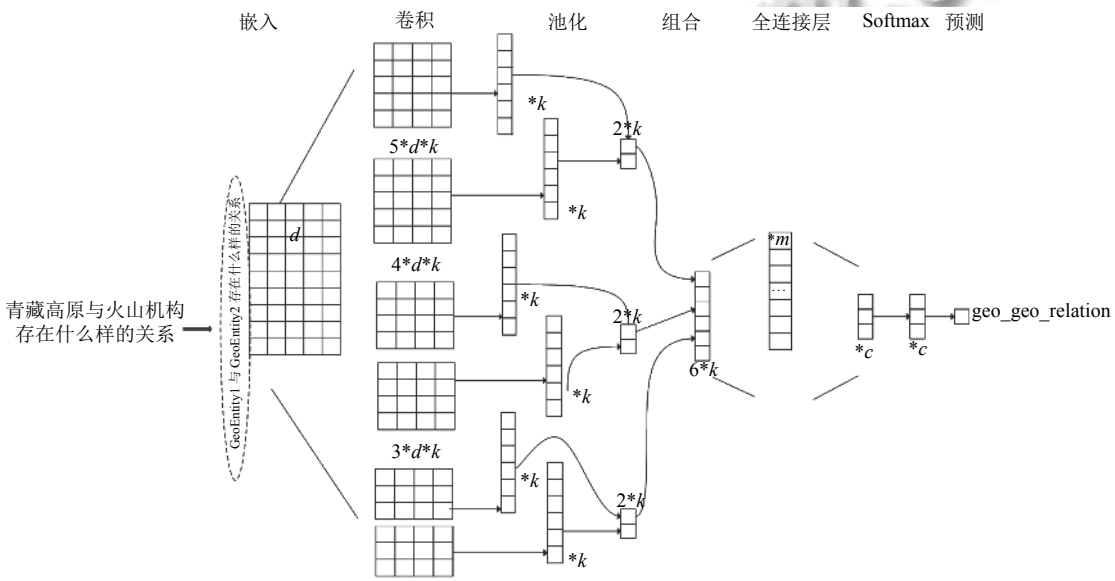

图 6 基于字符的 CNN 分类模型结构图

算法 1. 基于字符 CNN 的语句属性分类算法

1) 用户输入问句 q, 问句训练集$Q=[q_1,q_2,\cdots,q_n]$, $L=[l_1,l_2,\cdots,l_n]$;
2) 加载配置初始化 CNN 网络;
3) 将问句训练集构建词汇表, 使用 char 的表示, 进而将词汇表转化为{词: id}的表示;
4) 将分类目录固定, 转化为类别{类别: id}表示;
5) 将训练集、验证集数据划分 Epoch 批次数据;
6) 将输入的语句转化为句子向量v_q;
7) 采用 CNN 网络的不同大小卷积核对问句进行特征提取;
8) 计算输入语句向量v_q的卷积结果;
9) 采用梯度下降方法更新 CNN 网络模型参数;
10) 将所有的卷积特征结果使用最大池化处理后拼接成一个向量;
11) 经过全连接层, 使用 Softmax 对最大池化输出做分类计算, 得到用户语句对应知识类别l_x;
12) 输出用户输入语句 q 相近的类别标签l_x.

3.3 用户意图映射检索

当问句信息中的实体识别与属性映射分类完成后将二者相结合, 使用集束搜索转化为满足知识图谱查询的结构化语言 Cypher 进行问答机器人检索. 如“青藏高原与火山机构具有怎样的关系”, 通过实体识别模块将“青藏高原”“火山机构”进行识别出, 归类为地质实体 geoEntity, 然后通过问句属性分类, 将“具备怎样的关系”归属为 relation 类, 接着将二者结合转化为 Neo4j 图形数据库的结构化语句, 如:

“Match p=(n1: entity1)-[r: rel]->(n2: entity2) where n1.name=‘{0}’ and n2.name=‘{1}’ return distinct r.rel”.

其中 name1、name2 为实体名称, rel 相当于两个实体之间的关系.

4 实验过程与结果评估

4.1 数据集与评价标准

本文进行了实验来研究网格 LSTM-CC 优化算法

在不同领域的有效性. 首先使用 SimpleQuestion 数据集与地质领域 300 篇文献标注数据进行对基于字符的神经网络汉语 NER 进行实验识别. 同时我们使用两类数据集进行分类器验证训练, 一类是使用 THUCNews 数据, 每类 6500 条数据; 一类是使用实验室对于地质问答中用户常问问题及问答类型对应通用语句进行标注的数据, 每类平均大约 400 条, 共计 6500 条数据, 按照固定比例划分训练集、测试集、验证集.

实验中使用精确度、召回率和 $F1$ 作为验证评价指标, 对于整体多分类结果使用混淆矩阵.

$$\text{精确率}: P = \frac{TP}{TP+FP} \tag{24}$$

$$\text{召回率}: R = \frac{TP}{TP+FN} \tag{25}$$

$$F1\text{值}: F1 = \frac{2*P*R}{P+R} \tag{26}$$

混淆矩阵:

$$P_{\text{macro}} = \frac{p_1+p_2+\cdots+p_3}{n} \tag{27}$$

$$R_{\text{macro}} = \frac{R_1+R_2+\cdots+R_n}{n} \tag{28}$$

4.2 实验过程

实验过程中, 使用 CPU 对实体识别与属性分类进行了训练. 实体识别部分针对双向神经网络使用字符嵌入大小为 100, 单词批量大小为 60, LSTM 单元为 100, 剪枝大小为 5.0, 训练学习速率为 0.001, 与防止过拟合的 dropout 大小为 0.5, 训练内容包括 97 万带 BIOES 标签标注的文本信息, 迭代次数循环 64 次, 直至损失变化幅度稳定结束. 属性分类过程中采用卷积核分别为 3、4、5、 256 个卷积核, 词向量维度为 64, 序列长度为 600, 全连接层为 128 个神经元, 词汇表大小为 500, 迭代总轮次为 10 轮, 每批训练大小 64, 学习率为 0.001, 及 dropout 大小 0.5. 实验结果采用精确率、召回率、$F1$ 值求算数平均值, 作为最后结果.

4.3 结果分析

在实体识别中, 使用地质标注数据集与进行验证, 使用基于模板匹配和基于网格的 LSTM+CRF 的神经网络验证得到结果如表 4.

表 4 基于网格 LSTM+CRF 命名实体识别结果

方法	精确率	召回率	$F1$值
基于模板匹配	0.76	0.68	0.64
基于LSTM+CRF	0.84	0.87	0.86

在用户属性分类中, 使用 THUCNews 数据集对其 10 个类别, 每类 6500 条数据采用基于字符的 CNN、RNN 模型实验结果如表 5、表 6 所示, 通过训练可以发现基于 CNN 的模型较基于 RNN 模型用时较短, 如表 7 所示.

表 5 基于 THUCNews 数据集的字符 RNN 分类模型训练结果

类别	精确率	召回率	$F1$值
体育	0.97	0.99	0.98
财经	0.96	0.98	0.97
房产	0.99	0.99	0.99
家居	0.96	0.82	0.88
教育	0.91	0.94	0.92
科技	0.93	0.98	0.95
时尚	0.94	0.94	0.94
时政	0.96	0.91	0.93
游戏	0.97	0.95	0.96
娱乐	0.92	0.97	0.94

表 6 基于 THUCNews 数据集的字符 CNN 分类模型训练结果

类别	精确率	召回率	$F1$值
体育	1.00	0.99	0.99
财经	0.96	0.98	0.97
房产	1.00	1.00	1.00
家居	0.98	0.84	0.91
教育	0.94	0.98	0.96
科技	0.93	0.98	0.95
时尚	0.93	0.98	0.96
时政	0.95	0.94	0.95
游戏	0.98	0.98	0.98
娱乐	0.98	0.97	0.98

表 7 基于 THUCNews 数据集的字符 CNN 与 RNN 模型对比

方法	平均精确率	运行耗时
基于字符CNN	0.9633	11分15秒
基于字符RNN	0.9545	2时10分37秒

在 THUCNews 的基础上我们可以知基于字符的 CNN 模型不仅运行时间为基于字符的 RNN 模型的 1/13, 且在数据集上得到 96.3% 的精确率, 由此我们使用基于字符的 CNN 模型在我们针对用户一般询问语句人工标注的地质问答数据得到如表 8 所示, 平均精确率达到 96.9%, 使得应用效果超过基线模型.

表 8 基于地质标注数据集的字符 CNN 分类模型结果

方法	平均精确率	运行耗时
基于字符CNN	0.9691	3分44秒
基于字符RNN	0.9556	2时20分20秒

5 结论与展望

本文在地质领域用户意图识别中通过构建地质领域的实体字典, 来源包括地质百科大辞典、搜狗语料等, 在基于字符的网格神经网络上进行专家及用户的询问语句实体识别训练, 采用的是地质文献数据, 在验证集上验证, 采用 Adam 随机梯度下降时, 准确率达到 84.57%、召回率达到 87.12%, *F*1 值更是达到 86.18%, 超过了基于模板匹配与基于 RNN 的现有模型, 可有效地识别特定领域的实体及关系. 同时在短文本信息分类过程中借鉴卷积网络考虑语义信息的优势, 采用基于字符的分类模型达到 96.9% 的精确率, 对于分类结果使用知识图谱分类属性映射得到匹配的知识描述返回用户, 整体实现了在基于地质领域的问答过程中意图识别.

在此基础上, 将来的工作更多的是将用户热点问题及知识意图推理进行深入探索, 通过接下来的实验, 将知识图谱中知识的构建环节引入知识阶层路径, 实现用户复杂文本信息意图的识别.

参考文献

1 罗成, 刘奕群, 张敏, 等. 基于用户意图识别的查询推荐研究. 中文信息学报, 2014, 28(1): 64–72. [doi: 10.3969/j.issn.1003-0077.2014.01.009]

2 赵乐, 张兴旺. 面向 LDA 主题模型的文本分类研究进展与趋势. 计算机系统应用, 2018, 27(8): 10–18. [doi: 10.15888/j.cnki.csa.006456]

3 Li X. Understanding the Semantic Structure of Noun Phrase Queries. Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA. 2010. 1337–1345.

4 Ramanand J, Bhavsar K, Pedanekar N. Wishful thinking: Finding suggestions and 'buy' wishes from product reviews. Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Stroudsburg, PA, USA. 2010. 54–61.

5 Song HJ, Park SB. Identifying intention posts in discussion forums using multi-instance learning and multiple sources transfer learning. Soft Computing, 2018, 22(24): 8107–8118. [doi: 10.1007/s00500-017-2755-8]

6 孙镇, 王惠临. 命名实体识别研究进展综述. 现代图书情报技术, 2010, (6): 42–47.

7 Florian R, Ittycheriah A, Jing HY, *et al*. Named entity recognition through classifier combination. Proceedings of the 7th Conference on Natural Language Learning at HLT-NAACL 2003. Stroudsburg, PA, USA. 2003. 168–171.

8 Borthwick A, Sterling J, Agichtein E, *et al*. NYU: Description of the MENE named entity system as used in MUC-7. Proceedings of the 7th Message Understanding Conference. Fairfax, VA, USA. 1998. 145–150.

9 Isozaki H, Kazawa H. Efficient support vector classifiers for named entity recognition. Proceedings of the 19th International Conference on Computational linguistics. Stroudsburg, PA, USA. 2002. 1–7.

10 Zhou GD, Su J. Named entity recognition using an HMM-based chunk tagger. Proceedings of the 40th Annual Meeting on Association for Computational Linguistics. Stroudsburg, PA, USA. 2002. 473–480.

11 张硕果, 汪成亮. 结合 CRFs 的词典分词法. 计算机系统应用, 2010, 19(11): 115–118. [doi: 10.3969/j.issn.1003-3254.2010.11.026]

12 何炎祥, 罗楚威, 胡彬尧. 基于 CRF 和规则相结合的地理命名实体识别方法. 计算机应用与软件, 2015, 32(1): 179–185, 202. [doi: 10.3969/j.issn.1000-386x.2015.01.046]

13 Hu ZT, Ma XZ, Liu Z, *et al*. Harnessing deep neural networks with logic rules. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin, Germany. 2016. 2410–2420.

14 Huang ZH, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv preprint arXiv: 1505.01991.

15 Lample G, Ballesteros M, Subramanian S, *et al*. Neural architectures for named entity recognition. Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. San Diego, CA, USA. 2016. 260–270.

16 杨文明, 褚伟杰. 在线医疗问答文本的命名实体识别. 计算机系统应用, 2019, 28(2): 8–14. [doi: 10.15888/j.cnki.csa.006760]

17 胡泽文, 王效岳, 白如江. 国内外文本分类研究计量分析与综述. 图书情报工作, 2011, 55(6): 78–81, 142.

18 薛春香, 张玉芳. 面向新闻领域的中文文本分类研究综述. 图书情报工作, 2013, 57(14): 134–139. [doi: 10.7536/j.issn.0252-3116.2013.14.022]

19 Yang J, Liang SL, Zhang Y. Design challenges and misconceptions in neural sequence labeling. Proceedings of the 27th International Conference on Computational Linguistics. Santa Fe, NM, USA. 2018. 3879–3889.

20 Bordes A, Usunier N, Chopra S, *et al*. Large-scale simple question answering with memory networks. arXiv preprint arXiv: 1506.02075.

21 Golub D, He XD. Character-level question answering with attention. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin, TX, USA. 2016. 1598–1607.

22 Yin WP, Yu M, Xiang B, *et al*. Simple question answering by attentive convolutional neural network. Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers. Osaka, Japan. 2016. 1746–1756.

23 Lukovnikov D, Fischer A, Lehmann J, *et al*. Neural network-based question answering over knowledge graphs on word and character level. Proceedings of the 26th International Conference on World Wide Web. Perth, Australia. 2017. 1211–1220.

24 Zhang Y, Yang J. Chinese NER using lattice LSTM. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne, VIC, Australia. 2018. 1554–1564.

25 Kim Y. Convolutional neural networks for sentence classification. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Doha, Qatar. 2014. 1746–1751.