软件学报 ISSN 1000-9825, CODEN RUXUEW
Journal of Software, [doi: 10.13328/j.cnki.jos.006114]

E-mail: jos@iscas.ac.cn
http://www.jos.org.cn
Tel: +86-10-62562563

基于关联记忆网络的中文细粒度命名实体识别*

琚生根, 李天宁, 孙界平

(四川大学 计算机学院,四川 成都 610065)

通讯作者: 孙界平, E-mail: sunjieping@scu.edu.cn

摘 要: 细粒度命名实体识别是对文本中的实体进行定位,并将其分类至预定义的细粒度类别中.目前中文细粒度命名实体识别仅使用预训练语言模型对句子中的字符进行上下文编码,并没有考虑到类别的标签信息具有区分实体类别的能力.由于预测句子不带有实体标签,本文使用关联记忆网络来捕获训练集句子的实体标签信息,并将标签信息融入预测句子的字符表示中.该方法将训练集中带实体标签的句子作为记忆单元,利用预训练语言模型获取原句子和记忆单元句子的上下文表示,再通过注意力机制将记忆单元句子的标签信息与原句子的表示结合,从而提升识别效果.在 CLUENER 2020 中文细粒度命名实体识别任务上,本文方法对比基线方法获得了提升.

关键词: 中文细粒度命名实体识别;关联记忆网络;多头自注意力;预训练语言模型

中图法分类号: TP311

中文引用格式: 琚生根,李天宁,孙界平.基于关联记忆网络的中文细粒度命名实体识别.软件学报. http://www.jos.org.cn/1000-9825/6114.htm

英文引用格式: Ju SG, Li TN, Sun JP. Chinese Fine-grained name entity recognition based on associated memory networks. Ruan Jian Xue Bao/Journal of Software, (in Chinese). http://www.jos.org.cn/1000-9825/6114.htm

Chinese Fine-Grained Name Entity Recognition Based on Associated Memory Networks

JU Sheng-Gen, LI Tian-Ning, SUN Jie-Ping

(College of Computer Science, SiChuan University, Chengdu 610065, China)

Abstract: Fine-grained named entity recognition is to locate entities in text and classify them into predefined fine-grained categories. At present, Chinese fine-grained named entity recognition only uses pre-trained language models to encode characters in sentences and does not take into account that the category label information can distinguish entity categories. Since the predicted sentence does not have the entity label, the associated memory network is used to capture the entity label information of the sentences in the training set and incorporate label information into the representation of predicted sentences. In this method, sentences with entity labels in the training set are used as memory units, the pre-trained language model is used to obtain the sentence representations of the original sentence and the sentence in the memory unit. Then, the label information of the sentences in the memory unit is combined with the representation of the original sentence with the attention mechanism to improve the recognition effect. On the CLUENER 2020 Chinese fine-grained named entity recognition task, the method improves performance over the baseline methods.

Key words: chinese fine-grained name entity recognition;associated memory network;multi-headed self-attention;pre-trained language model

命名实体识别是自然语言处理中的信息抽取任务之一,其目的是对文本中特定类别的实体进行定位和分类.大多数命名实体识别任务只识别人名、组织、地点等实体类别,识别的实体类别少,并且类别划分的比较宽

* 基金项目: 国家自然基金(61972270);四川省新一代人工智能重大专项(2018GZDZX0039);四川省重点研发项目(2019YFG0521)

Foundation item: National Natural Science Foundation of China (61972270); Sichuan Province New Generation Artificial Intelligence Major Project (2018GZDZX0039); Sichuan Province Key Research & Development Project (2019YFG0521)

收稿时间: 2020-04-13; 修改时间: 2020-05-27, 2020-06-20; 采用时间: 2020-07-01; jos 在线出版时间: 2021-04-20

泛.然而,细粒度命名实体识别更符合现实世界的知识体系,在一些常见类别的基础上做了近一步的类别划分,需要识别的实体种类远多于一般的命名实体识别,这样从文本中抽取的实体就拥有了一个更详细的定义,为下游的知识图谱的构建和问答任务提供更有力的支撑.

在细粒度命名实体识别中,更细粒度的划分会造成各实体类别在语义上有更紧密的距离.模型对语义相近类别的实体进行分类时,容易发生混淆,这意味着细粒度实体类别的区分更具有挑战性.目前,中文公开的高质量细粒度命名实体识别的数据集很少,CLUENER2020[1]数据集包含 10 种不同的实体类别,并对一些常见类别进行了细粒度的划分,如从“地点”中分离出来了“景点”,从“组织”中分离出了“政府”和“公司”,这就造成“地点”和“景点”之间,“组织”、“政府”和“公司”之间的混淆程度较高.同时存在同一实体在不同语境下属于不同类别的情况,如“游戏”可以是一些“书籍”和“电影”的改编.如表 1 所示,实体“《黑暗之塔》”在第一个句子中属于“游戏”类别,在第二和第三个句子中属于“书籍”.在这种情况下,对实体类别的区分需要结合上下文语境,同时也给“游戏”、“书籍”和“电影”实体类别之间的区分造成混淆.Xu[1]等人使用 bilstm-crf[2]和预训练语言模型[3]在该数据集上进行实验,相同的模型在其他数据集上的 F1 值可以达到 95,然而在该数据集上最好的效果只能达到 80 左右,这是因为细粒度数据集的句子中经常存在多个类别的实体,模型在预测时会出现一些实体类别的丢失,同时模型对一些类别的区分能力也存在一定的限制,因此细粒度命名实体识别任务更具有挑战性.

Table1 Examples of entities belonging to different categories in different contexts

表 1 实体在不同的语境属于不同类别的例子

句子	标签
《黑暗之塔》改编游戏将在 2013 年 5 月随电影版同步上市.	游戏:《黑暗之塔》
斯蒂芬金的《黑暗之塔》小说共有七卷本,这个系列是斯蒂芬金最负盛名的小说,	书籍:《黑暗之塔》 姓名:斯蒂芬金
蒂芬金《黑暗之塔》将改编成游戏	书籍:《黑暗之塔》 姓名:蒂芬金

Xu 等人使用 bilstm-crf、预训练语言模型作为 CLUENER2020 数据集的基线[1],其中预训练语言模型在该数据集上取得了最好的效果.Xu 等人使用预训练语言模型对句子中的字符进行上下文编码,在一定程度上解决了同一实体在不同句子中属于不同类别的问题,但模型还是存在实体类别丢失和分区类别能力不足的情况.本文受到 Wang[4]等人的启发,考虑到细粒度命名实体识别数据集中实体类别多,类别区分难的问题,通过在句子的字符表示中融入类别的标签信息,使句子字符的上下文表示更加接近类别的标签嵌入,来提高识别效果.

为了利用标签信息,本文通过关联记忆网络[5]的方式,使用训练集中带标签的句子,通过捕获训练集中相关句子的标签信息,并使用注意力机制将类别的标签信息融入句子的字符表示中.同时结合预训练语言模型和多头自注意力,提升模型的识别效果.最后,通过消融实验分别证明了关联记忆网络、多头自注意力和预训练语言模型 3 个部分在 CLUENER 2020 细粒度命名实体识别任务上的有效性.

本文的贡献主要包括:(1)提出了一种结合预训练语言模型和关联记忆网络的方法,利用标签类别信息辅助细粒度实体类别的区分,并通过实验证明了模型方法的有效性,同时证明实体的标签类别信息对细粒度命名实体识别有促进作用;(2)针对命名实体识别任务,本文提出了一种实体类别距离的记忆句子选择方式,在实体类别距离的选择方式上进行实验,验证了细粒度命名实体识别的挑战在于实体类别的区分,正确的实体类别的标签可以大幅度提升模型的识别效果.

本文剩余部分的结构如下:第一节介绍关于中文细粒度命名实体识别的相关工作;第二节详细地描述了本文提出的模型;第三节使用本文模型在 CLUENER2020 数据集上和其他模型进行对比实验,验证本文模型的有效性.第四节总结全文并提出未来的发展方向.

1 相关工作

命名实体识别任务主要的方法有 3 种：基于规则[6]、基于传统机器学习[7]和基于深度学习的方法,其中,基于深度学习的方法由于可以自动地捕获输入句子的特征,实现端到端的命名实体识别,已经成为了现在的研究热点.

近年来,基于深度学习的方法在命名实体识别任务上获得了很好的应用.Huang 等人[2],peng 等人[8]使用了双向长短期记忆网络(BiLSTM)和条件随机场网络进行命名实体标记,但 BiLSTM 编码长序列的能力有限,并且计算速度慢.Strubell 等人[9]将卷积神经网络(CNN)用于命名实体识别,相比循环神经网络,卷积神经网络具有更快的计算速度,但是,卷积神经网络更多捕获的是局部信息,会造成全局信息的大量丢失.

中文命名实体识别区别于英文,由于句子中的词没有天然的边界,更具有挑战性.Yang[10]等人使用分词工具对句子序列进行分词,然后对单词序列进行标注.然而,分词工具不可避免地会出现单词的错误划分,造成实体边界的错误识别.因此,一些工作(Liu 等人[11],Lu 等人[12])表明字符级别的命名实体识别的效果比单词级别的更好.基于字符的命名实体识别存一个很明显的缺点就是没有充分地利用单词信息.所以,中文命名识别的研究热点是将词典信息充分融入字符模型中.Zhang 和 Yang[13]提出了 Lattice-LSTM 模型,通过长短期神经网络的门控机制自动地匹配句子中每个字符对应的单词,将词典中与句子语义最匹配的单词信息融入句子表示中,从而提升了模型识别的能力.除了字典信息可以提高命名实体识别的效果,Xu 等人[14]表示中文字符部首蕴含的特征信息能够帮助命名实体的识别.他在模型中同时使用了字嵌入、词嵌入和部首嵌入来丰富句子中的字符表示,并验证了部首信息的有效性.然而,这些研究都没有关注到实体类别的标签信息可能帮助命名实体的识别.Wang[4]等人连接词嵌入和标签嵌入进行文本分类,引入标签注意力机制为句子中的每个单词分配权重来生成文本表示,通过训练使词嵌入接近于它们对应的标签嵌入.Luo[15]等人使用标签注意力机制生成全局的句子表示,为句子中的每个位置补充全局信息.Li[16]等人基于阅读理解的框架进行命名实体识别,通过编码每个类别标签的注释构建问题,然后通过问题在文本中匹配相应的类别的实体.Guan[5]等人提出了关联记忆网络的方法将单词级别的标签信息融入单词表示中进行语义角色标记,但是其通过双向长短期神经网络对单词表示进行上下文编码,由于双向长短期神经网络的编码能力有限,并且无法并行.所以,本文抛弃了长短期神经网络,使用 BERT 对句子中的字符进行上下文编码,从而获得更好的编码表示.

Xu 等人[1]构建了 CLUENER2020 中文细粒度命名实体数据集,该数据集包含大量的训练样例,同时各实体类别在数据集中分布平均.Xu 等人分别使用双向长短期神经网络、预训练语言模型作为该数据的基线,并在细粒度数据集和普通数据上进行对比试验,结果表明相同的方法在细粒度数据集上的效果要低于普通数据集.所有的基线方法都只是对句子中的字符进行了简单的上下文编码,然后通过条件随机场层学习实体标记之间的约束.所以本文通过关联记忆网络连接句子中字符的上下文表示和实体类别的标签嵌入,使上下文表示更接近于实体类别的标签嵌入,来提升细粒度命名实体识别的效果.

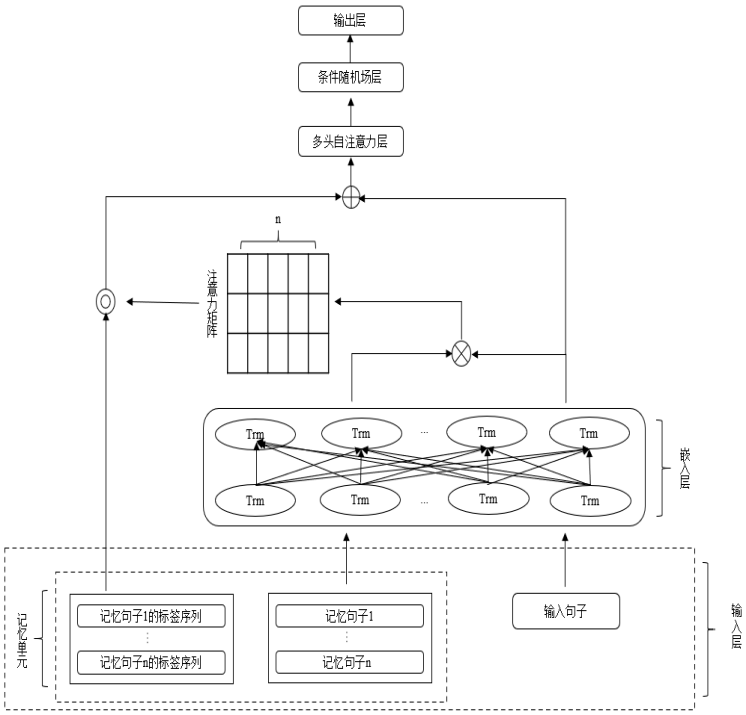

Fig.1 Model overall framework diagram

图 1 模型整体框架图

2 本文方法

给定输入句子 s,将其表示为字符序列 $s=[x_i]_{i=1}^{l_s}$,其中 l_s 表示句子的长度,x_i 表示句子中的第 i 个字符.本文将细粒度命名实体识别看成序列标注任务,将句子 s 作为输入,对 s 中的每一个字符进行标记,生成标签序列 $y=[y_i]_{i=1}^{l_s}$,其中 y_i 表示句子中第 i 个字符的标签.

本文为了将实体类别的标签信息融入输入句子 s 的字符表示中,提出了一种结合预训练语言模型和关联记忆网络的方法,将训练集中相关句子的正确实体标签信息融入输入句子的字符表示中,框架如图 1 所示.

该方法的模型自底向上包括输入层、嵌入层、关联记忆网络、多头自注意层、条件随机场层和输出层.其中,输入层进行记忆句子的选择,计算输入句子和训练集中句子的距离,将训练集中与输入句子距离最近的几个句子作为记忆句子;嵌入层对输入句子和记忆句子中的字符进行上下文编码,并将记忆句子的实体标签转换成标签序列,进行标签嵌入;关联记忆网络计算输入句子中每个字符和记忆句子中每个字符的注意力矩阵,并与记忆句子对应的标签嵌入相乘,将标签信息融入序列表示中;多头自注意力层结合句子任意位置之间的相互关注,对融入了标签信息的序列表示进行重新编码;条件随机场层学习各实体标签之间的规则;输出层使用维特比算法输出概率最高的标签序列.

2.1 输入层

输入层的主要目的是构建记忆单元,记忆单元的最小组成部分是一个句子以及与该句子对应的 n 个记忆句子,所以记忆句子的选择是本文需要解决的主要问题.模型的计算时间和空间,随着记忆句子的增大而增大,但在引入记忆句子的同时,模型的效果也有所提升.

模型的输入包括:输入句子 s 、句子 s 在训练集中对应的 n 个记忆句子和这 n 个句子对应的标签序列.本文使用了两种句子距离的计算方式,第一种使用 Guan[5]等人提供的计算两个句子词性序列编辑距离的方法,计算句子 s 和训练集中所有句子的距离,选择前 n 个与句子 s 距离最近的句子和这 n 个句子对应的标签序列存入记忆单元中;第二种针对命名实体识别任务,本文提出一种计算实体类别距离的方法,计算两个句子包含的实体类别的差异.由于记忆句子需要包含输入句子中的相应的实体类别,本文首先通过文本多标签预测模型预测句子中可能包含的实体类别,然后将训练集中与输入句子包含的实体类别最相近的句子存储在记忆单元中.具体地,假如输入句子包含的实体类别集合为 $Y_s=\{y_1,y_2,...,y_k\},y_k\in C$,训练集中句子中包含的实体类别集合为 $Y_t=\{y_1,y_2,...,y_n\},y_n\in C$,其中,C 是所有实体类别的集合,k 和 s 分别表示输入句子和记忆句子包含实体类别的数量,该方法先计算集合 Y_s 和集合 Y_t 的差集,以两者差集包含的实体类别数量 $|Y_s-Y_t|$ 降序,再以集合 Y_t 包含的类别数量 $|Y_t|$ 降序,对训练集中的句子进行排序.让记忆句子包含输入句子实体类别的同时,记忆句子包含的实体类别最少.本文通过实验证明,当多标签文本分类模型预测句子中包含的实体类别越准确,模型的命名实体识别的效果越好.

2.2 嵌入层

嵌入层是为了将句子的字符映射到同一个语义空间中, 根据上下文的语义将句子中的字符编码成向量. 本文选择 RoBERTa 语言模型[17]对句子中的字符进行编码,因为该模型是深度的神经网络模型,并且在大规模的语料上进行训练,可以更好的归纳自然语言文本中的语义和语法上的特性,但是由于模型参数量大,需要的计算空间也随之增大.

嵌入层包括两个部分：使用预训练语言模型对句子中的字符进行上下文嵌入;对记忆句子的标签序列进行标签嵌入.

本文使用预训练语言模型 RoBERTa 分别对输入句子 s 和 s 对应的 n 个记忆句子进行上下文嵌入,捕获每个字符在给定句子中的上下文信息.假设输入句子 $s=[x_i]_{i=1}^{l_s}$,其中 l_s 表示句子的长度, x_i 表示句子中的第 i 个字符.使用预训练语言模型对句子 s 中字符进行上下文编码,得到嵌入表示 s' ,嵌入公式(1)如下：

$$s'=[x_i']_{i=1}^{l_s}=RoBERTa([x_i]_{i=1}^{l_s}) \tag{1}$$

其中, x_i' 是字符 x_i 上下文编码向量,维度为 $\mathbb{R}^d$,其中 d 是预训练语言模型隐藏层的维度.记忆句子的上下文嵌入过程与输入句子相同.假设, n 个记忆句子为 $\mathrm{a}_j=[x_{j,k}]_{k=1}^{l_j}$, $j\in\{1,2,...,n\}$ 其中 l_j 表示第 j 个记忆句子的长度, $x_{j,k}$ 表示第 j 个记忆句子中的第 k 个字符.通过预训练语言模型对 n 个记忆句子进行上下文编码,得到嵌入表示 a_j' , $j\in\{1,2,...,n\}$,嵌入公式(2)如下,其中, $x_{j,k}'$ 是字符 $x_{j,k}$ 的上下文编码向量,维度为 $\mathbb{R}^d$.

$$a_j'=[x_{j,k}']_{k=1}^{l_j}=RoRERTa([x_{j,k}]_{k=1}^{l_j}) \tag{2}$$

对于记忆句子的标签嵌入,本文首先使用预训练的词向量对训练集中的各类实体进行词嵌入,如果出现未登陆的实体则进行字符嵌入,词嵌入和字符嵌入的维度为 300 维.然后,将各类实体嵌入表示的平均数作为标签嵌入矩阵的初始化权重.另外,由于本文采用 BIOS 的形式对实体进行标记,为了表明实体标签的位置信息,如图 2 所示,我们将 4 位 onehot 向量拼接在各实体标签嵌入的尾部.最后,标签嵌入的维度 $\mathbb{R}^l$ 为 304 维.标签嵌入矩阵在训练过程中更新,使句子字符上下文的表示更加接近实体类别的标签嵌入.本文使用预训练的词向量和字符向量对标签嵌入矩阵进行初始化,而不是选择随机初始化,是为了让初始化的标签矩阵就包含一些实体类别的

相关特征.

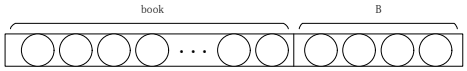

Fig.2 The from of label embedding

图 2 标签嵌入形式

2.3 关联记忆网络

一般的命名实体识别模型将经过上下文编码的输入序列表示通过 softmax 激活或者输入到条件随机场层进行标签序列的预测.本文通过引入关联记忆网络,在对输入序列表示进行标签预测之前,让输入句子的字符去记忆句子中匹配和其类别最相关的字符的标签信息,然后将匹配的标签信息融入输入句子的字符表示中进行预测.

关联记忆网络包含两部分：输入句子和记忆句子之间的注意力计算、标签序列的融入和合并.计算输入句子和记忆句子之间的注意力,是为了捕获输入句子的中的每一个字符对记忆句子中每一个字符在实体类别方面的相似度,如果两个字符拥有相同的实体标签,那么这两个字符之间就拥有较高的关联度.

在嵌入层得到句子 s 的向量表示 s' ,以及记忆句子的向量表示 a'_j , $j \in \{1,2,..,n\}$ 后,通过公式(3)计算句子 s' 中每个字符和记忆句子 a' 中每个字符的相似度,得到 n 个 $l_s \times l_j$ 的注意力矩阵 $M_j^{raw} = s'a_j'^T$, $j \in \{1,2,,...,n\}$,其中 n 是记忆句子的数量, l_s 是输入句子的长度, l_j 是第 j 个记忆句子的长度.

$$M_j^{raw} = s'a_j'^T \tag{3}$$

最后,通过公式(4)和(5)对未经过归一化的 M_j^{raw} 注意力矩阵按行进行归一化,得到归一化后的矩阵 M_j $j \in \{1,2,...,n\}$,其中 $a_{i,j}$ 是一个 l_j 维的向量,该向量的分量表示句子 s 中的第 i 个字符对记忆句子 a_j 中每个字符的注意力权重.

$$a_{i,j} = f([M_j^{raw}(i,1),...,M_j^{raw}(i,l_j)]) \tag{4}$$

$$M_j = [a_{1,j}, a_{2,j},...,a_{l_s,j}] \tag{5}$$

其中, $f(\cdot)$ 代表 softmax 函数, $a_{i,j}$ 是 M_j^{raw} 矩阵第 i 行的简化形式.

标签序列的融入和合并,如公式(6)所示,是将归一化后的注意力矩阵 M_j 与记忆句子对应的标签嵌入序列 L_j 相乘后,得到融入了标签信息的序列 L'_j ,它根据输入句子 s 中的每个字符对记忆句子中每个字符标签的关注程度,来计算输入句子中每个字符对应的标签类别信息.

$$L'_j = [a_{i,j} \cdot L_j^T]_{i=1}^{l_s}, \quad j \in \{1,2,...,n\} \tag{6}$$

最后,将 n 个融入了标签信息的序列 L'_j 进行平均,并与句子 s 的上下文向量 s' 拼接,得到最后的输入句子表

示 e ,如公式(7)所示.其中 e 的表示维度为 $\mathbb{R}^{d+l}$, $mean(\cdot)$ 是平均函数.图 3 是嵌入层和关联记忆网络层的向量间的形状转换图.

$$e=[s',mean(L'_j)] \tag{7}$$

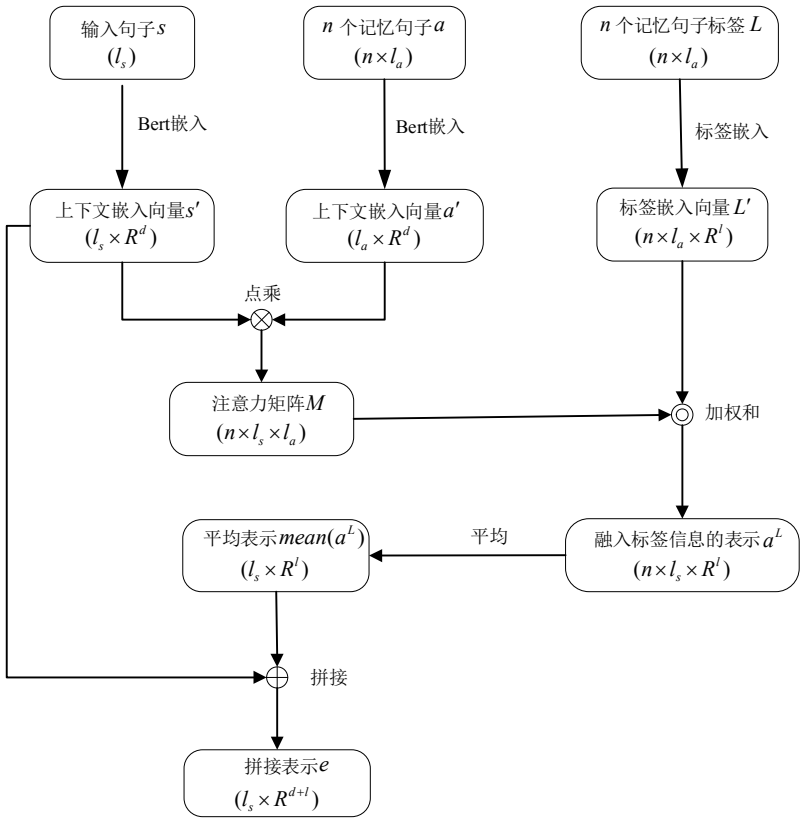

Fig.3 Shape transition diagram of vector in each layer

图 3 各层向量形状转换

2.4 多头自注意力层

多头自注意力层的主要作用是结合序列各个位置的相关度,对融合了标签信息的字符表示进行重新编码,使用自注意力机制对句子序列进行编码,避免了 LSTM 不能并行的缺点,同时可以更好的捕获全局信息.

多头自注意力层将最后的句子表示 e 作为输入,通过多头自注意力机制从多个角度计算输入序列任意位置之间的相关度,突出序列每个位置实体类别的最重要信息,图 4 为多头自注意力的计算机制.

如公式(8)所示,多头自注意力机制通过不同的线性映射将输入向量映射成 query、key 和 value 的形式,并映射到不同的子空间中,每个子空间反映不同的隐藏特征.其中, $score_i(e)$ 表示第 i 个自注意力头的输出, W_i^Q , W_i^K , W_i^V 表示映射到第 i 个子空间对应的参数,各参数的维度大小为 $W_i^Q\in\mathbb{R}^{d\times d_Q}$, $W_i^K=\mathbb{R}^{d\times d_v}$, $W_i^V\in\mathbb{R}^{d\times d_v}$,其中, d 表示多头自注意力层输入向量的维度, d_Q , d_V , d_K 分别表示 query、key 和 value 的映射维度.

$$score_i(e) = attention(eW_i^Q, eW_i^K, eW_i^V) \tag{8}$$

然后,如公式(9)所示,计算输入序列中某个位置的 query 和所有位置的 key 的相似度,得到注意力矩阵.这个注意力矩阵表示了句子中两两位置之间的关注度,将注意力矩阵和该位置 value 相乘,捕获句子中任意位置之间的关系.

$$attention(Q,K,V) = softmax(\frac{QK^T}{\sqrt{d}})V \tag{9}$$

最后,如公式(10)所示,将各个子空间的计算结果进行拼接,经过线性映射,得到最终的输出.其中,$W^O \in \mathbb{R}^{nd_v \times d}$,$n$ 是子空间的数量.

$$m(e) = concat(score_1(e), score_2(e), ..., score_n(e))W^O \tag{10}$$

这样最后得到的输入序列就包含了每个自注意力头学习到的语义和语法特征.

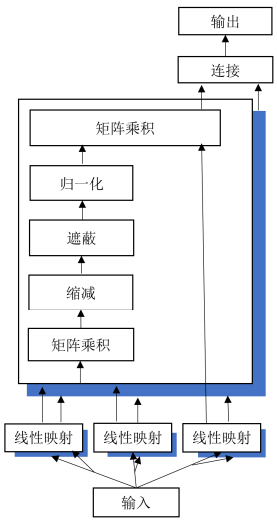

Fig.4 Multi-head self-attention layer

图 4 多头自注意力层

2.5 条件随机场层和输出层

条件随机场的作用是约束标签序列的预测,通过公式(11),计算预测标签序列 $y = l_1, l_2, ..., l_s$ 的概率 $P(y|s)$ 为:

$$P(y|s) = \frac{\exp(\sum_i (W_{CRF}^{l_i} h_i + b_{CRF}^{(l_{i-1}, l_i)}))}{\sum_{y'} \exp(W_{CRF}^{l_i'} h_i + b_{CRF}^{(l_{i-1}', l_i')})} \tag{11}$$

其中,W_{CRF} 和 b_{CRF} 是条件随机场层的权重和偏置,反映的是各标签之间的转换分数,通过训练对参数进行更新.

给定带标签的训练集 $\{(s_i, s_j)\}|_{i-1}^N$,训练的损失函数为句子级别的对数似然损失,如公式(12)所示:

$$L = -\sum_{i-1}^N \log(P(y_i|s_i)) \tag{12}$$

我们的模型使用随机梯度下降进行端到端训练,通过最小化句子级别的负对数似然来训练模型的参数.

在训练过程中,由于微调预训练语言模型的学习率满足不了条件随机场层参数的训练.本文在模型训练的过程中,增大了除预训练语言模型层之外其他层的学习率,来优化模型参数.在预测阶段,输出层使用维特比算法找到分数最高的标签预测序列进行输出.

3 实验与结果

3.1 数据集与评价指标

本文采用 CLUENER 2020 数据集进行实验,该数据集的实体类别分为 10 种：address, book,company, game,government, movie,name, organization,position,scene.该数据集只提供训练集和验证集的标注,不提供测试集的标注.数据集的详细信息如表 2 所示.本文采用 CLUENER 2020 提供的线上测评网站*,以 F1 值对实验结果进行评价.

Table2 The description of CLUNER2020 dataset

表 2 CLUNER2020 数据集描述

Dataset	Train	Dev	Test	Avg length	Max length	Classes
CLUENER2020	10748	1343	1345	37.4	50	10

3.2 参数设置

本实验使用 Colab pro p100 16g 内存.由于内存限制,在嵌入层使用中文预训练语言模型 RoBERTa 的 base 版本,该模型是包含 12 层的 Transformer.

本文模型中使用的参数取值如表 3 所示.通过实验证明,增大其他层的学习率,包括自注意力层和条件随机场层的学习率后,模型的效果有所提升,学习出来的条件随机场层的参数也符合真实情况.

Table3 Parameter value

表 3 参数取值

参数	取值
Epoch	20
Batchsize	16
最大句子长度	64
记忆句子数	4
标签嵌入维度	304
预训练模型层数	12
预训练模型学习率	3e-5
其他层学习率	2e-4
Dropout	0.1
自注意力层维度	512

3.3 实验结果

本文模型在 CLUNER2020 数据集上的训练过程如图 5 所示,图 5 是模型在验证集上的 F1 值曲线图.从图中可以看出,模型训练的前期 F1 值提升较快,然后不断的波动寻找局部最优值,最后趋于平稳.

* 评测网站 https://www.cluebenchmarks.com/ner.html

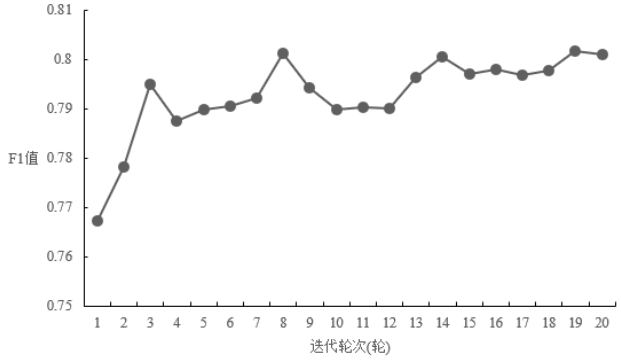

Fig.5 F1 value graph on the validation set

图 5 验证集上的 F1 值曲线图

(1)实体类别距离选择方式分析

为了证明文本提出的实体类别距离的有效性,文本使用训练集和验证集中的句子包含的 gold 的实体类别构建记忆单元,并在验证集上进行实验.如表 4 所示,引入正确的类别标签信息,可以大幅度促进命名实体的识别.表 4 中第一行和第三行对比了编辑距离方法和实体类别距离在验证集上的效果,表中的 pred 表示句子包含的实体类别是通过基于 bert 的文本多标签预测模型得到.从结果可以看出,本文提出的实体类别距离在验证集上的效果要高于编辑距离.通过第二行和第三行的对比,句子包含的实体类别预测越准确,模型的效果越高.从实验结果可以看出,细粒度命名实体识别任务还有很大的提升空间,该任务的挑战在于对实体类别的预测.

Table4 Ablations about distance on development set

表 4 验证集上距离计算方法的消融实验

计算方法	P	R	F1
编辑距离	78.97	81.41	80.17
实体类别距离(gold)	86.07	87.11	86.59
实体类别距离(pred)	79.33	81.58	80.44

(2)各实体类别评价指标分析

RoBERTa-Base 模型和本文模型在验证集上,各实体类别精确率,召回率和 F1 的对比如表 5 所示.从表中可以得出,在所有类别总体的准确率和召回率上,本文模型都有所提升.在各类别的召回率上,本文模型都较高,说明本文模型能多识别更多的命名实体.从表 5 可以看出模型对“地址”和“景点”的类别实体的类别的实体效果差,模型的效果取决于对这两种类别实体的识别和区分.

Table5 The comparison of models on validation set

表 5 验证集上的模型对比

实体	RoBERTa-base			RoBERTa-base +关联记忆网络 +多头自注意力		
	P	R	F1	P	R	F1
人名	86.75	90.11	88.4	86.25	89.03	87.62
组织	80.93	80.93	80.93	80.05	**83.11**	81.55
职位	81.26	83.14	82.19	80.73	82.22	81.46
公司	84.2	81.75	82.95	80.25	**83.86**	82.02
地址	59.76	66.49	62.94	63.07	**67.29**	65.11
游戏	81.21	86.44	83.74	82.19	**89.15**	85.53
政府	79.1	85.83	82.33	79.93	**87.04**	83.33
景点	69.12	71.77	70.42	72.02	66.51	69.15
书名	82.0	79.87	80.92	81.82	**81.82**	81.82
电影	85.19	76.16	80.42	84.56	76.16	80.14
Overall @Macro	78.76	80.99	79.86	78.97	81.41	80.17

(3)消融实验和基线模型对比

为了分析模型不同模块对实体识别效果的影响程度,本文在 RoBERTa-Base 模型+关联注意力网络的基础上,分别做了 2 组消融实验,分别去除了关联记忆网络、去除了预训练语言模型并使用 BiLSTM+字符嵌入进行上下文编码.实验结果如表 6 所示,分别验证了预训练语言模型,关联记忆网络对实验结果的影响.

Table7 Comparison of F1 values of each model

表 7 各模型 F1 值对比

模型	线上效果 F1(%)
LSTM+CRF[1]	70.00
LSTM+CRF+char	70.16
LSTM+CRF+char +关联记忆网络	71.20
BERT-Base[1]	78.82
RoBERTa-wwm-base-ext	79.16
RoBERTa-wwm-base-ext +关联记忆网络	79.98

表 7 将本文模型结果与对应的基线模型进行对比,表 7 中,LSTM+CRF 和 BERT-Base 语言模型是 Xu 等人提出的 2 个基线模型,本文使用 RoBERTa-wwm-base-ext 语言模型作为基线. RoBERTa 是 BERT 语言模型的升级版本,wwm 表示该语言模型在训练过程中使用了完整的单词遮蔽,ext 表示使用了更大规模的扩展语料, base 表示模型使用 12 层的 Transformer.

表 7 第二行和第三行的对比中,可以看出在基于 LSTM+CRF 的模型结构上加入关联记忆网络,模型的识别效果有所提升.从第五行和第 6 行可以看出,在使用预训练语言模型的基础上,加上关联记忆网络之后,借助正确的实体类别信息,模型的效果也有明显的提升.

(4)关联注意力矩阵分析

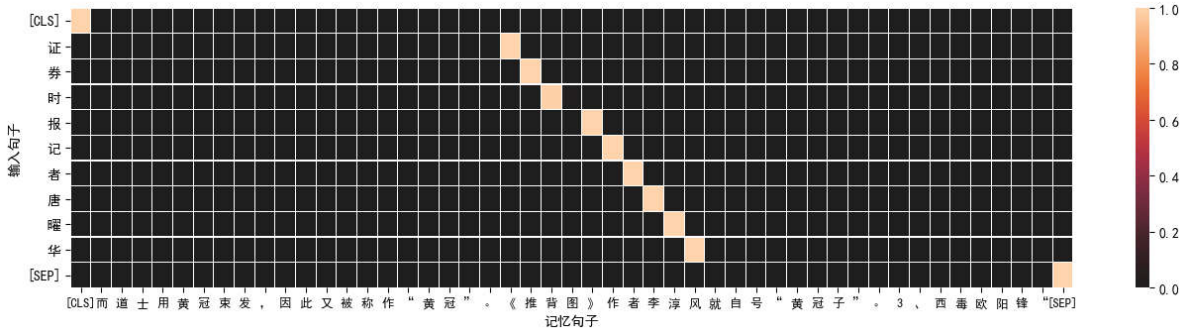

Fig.6 Heat map of associated attention based on entity category distance

图 6 基于实体类别距离的关联注意力热力图

我们通过分析各关联句子间的注意力矩阵来验证关联记忆网络的有效性.图 6 是基于实体类别距离的一对关联句子之间注意力矩阵的热力图,图中方格的亮度代表字符之间的相关性.从图中可以看出,对于“书籍”实体,输入句子中的“证券时报”与记忆句子中的“《推背图》”存在强关联;对于“名字”实体,实体“唐曜华”和实体“李淳风”存在强关联;对于“职位”实体,实体“记者”和“作者”存在强关联.这说明相同实体种类的上下文向量更加接近,通过捕获记忆句子中强关联实体的真实标签信息,就可以提升输入句子中实体的类别预测.

图 7 是基于编辑距离的一对关联句子之间的注意力矩阵的热力图,从图中发现当记忆句子中不包含输入

句子中的一些实体类别时,记忆句子的一些实体会关联到句子的其他位置。如图 7 所示,因为记忆句子中不包含“书籍”的实体类别,导致“书籍”实体“证券时报”关联到了“职位”实体“处长”,但是最终模型还是对“证券时报”做出了正确的预测。从实验中发现,基于编辑距离的记忆句子选择方式出现上述情况的可能性要大于基于实体类别距离的记忆句子选择方式,但模型通过训练能很好的对错误融入的标签信息进行处理,所以在对句子包含的实体类别预测效果不佳的情况下,基于编辑距离的记忆句子选择方式要好于第二种。

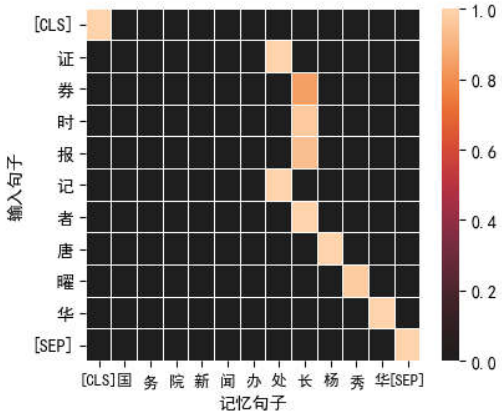

Fig.7 Heat map of associated attention based on edit distance

图 7 基于编辑距离的关联注意力热力图

由于,基于编辑距离的模型具有一定的利用正确实体类别信息和处理错误实体类别信息的能力,本文针对模型识别“地址”类别实体能力差的特点,尝试将所有句子的类别标签信息用“地址”类别的标签嵌入进行替换,如表 8 所示,发现模型的整体的识别效果得到了大幅度提升.

Table8 F1 value using enhanced address information

表 8 使用增强地址信息的模型 F1 值

模型	线上效果 F1(%)
RoBERTa-wwm-base-ext +关联记忆网络(地址信息增强)	80.62
RoBERTa-wwm-base-ext +关联记忆网络	79.98

4 总结

本文充分利用了预训练语言模型捕获了句子字符的上下文信息,同时利用关联记忆网络,使字符的上下文信息接近于实体类别的标签信息,并将类别的标签信息融入到序列的字符表示中,最后利用多头自注意网络高效的计算了句子任意位置间的关注度,对融入了标签信息的字符表示进行重新编码,增加了实体识别的效果.实验结果表明,本文模型在细粒度命名实体识别任务中取得了更好的效果.在未来的工作中,希望针对细粒度命名实体识别,设计多标签文本分类模型,来提高预测句子中包含的实体类别的效果,结合本文提出的实体类别距离的计算方法,来提高模型的识别效果.

References:

[1] Xu L,Tong Y,Dong QQ,Liao YX,Yu C,Tian Y,Liu WT,Li L,Liu CQ,Zhang XW.CLUENER2020: Fine-grained named entity recognition dataset and benchmark for chinese.2020.arXiv: Computation and Language, Available: https://arxiv.org/abs/20-01.04351.

[2] Huang Z, Xu W, Yu K. Bidirectional LSTM-CRF models for sequence tagging. arXiv: Computation and Language,2015. Available: https://arxiv.org/abs/1508.01991.

[3] Devlin J, Chang MW, Lee K, Toutanova K. Bert: pre-training of deep bidirectional transformers for language understanding.arXiv: Computation and Language,2018.Available: https://arxiv.org/abs/1810.04805.

[4] Wang G, Li C, Wang W, Zhang YZ, Shen DH,Zhang XY,Henao R,Carin L. Joint embedding of words and labels for text classification.In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics,2018: 2321-2331.

[5] Guan CY, Cheng YH, Zhao H. Semantic role labeling with associated memory network.In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), 2019,3361-3371.

[6] Xiang XW, Shi XD, Zeng HL. Chinese named entity recognition system using statistics-based and rules-based method.Computer Applications, 2005, 25(10): 2404-2406.

[7] Zhang CY, Hong XG, Peng ZH. Extracting web entity activities based on SVM and extended conditional random fields.Journal of Software, 2012, 23(10): 2612-2627.

[8] Peng JY, Fang Y, Huang C, Liu L, Jiang ZW. Cyber security named entity recognition based on deep active learning.Journal of Sichuan University: Natural Science Edition, 2019, 56(3): 457-462.

[9] Strubell E, Verga P, Belanger D, McCallum A. Fast and accurate entity recognition with iterated dilated convolutions.In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing,2017,2670-2680.

[10] Yang J, Teng ZY, Zhang MS, Zhang Y. Combining discrete and neural features for sequence labeling. International Conference on Intelligent Text Processing and Computational Linguistics,2016:140-154.

[11] Liu ZX, Zhu CH, Zhao TJ. Chinese named entity recognition with a sequence labeling approach: based on characters, or based on words? International Conference on Intelligent Computing,2010: 634-640.

[12] Lu YN, Zhang Y, Ji DH. Multi-prototype chinese character embedding.In: Proceedings of the Tenth International Conference on Language Resources and Evaluation,2016,855-859.

[13] Zhang Y,Yang J. Chinese ner using lattice lstm.In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (ACL), 2018,1554-1564.

[14] Xu CW, Wang FY, Han JL, Li CL. Exploiting multiple embeddings for chinese named entity recognition.In: Proceedings of the 28th ACM International Conference on Information and Knowledge Management,2019: 2269-2272.

[15] Luo Y, Xiao FS, Zhao H. Hierarchical contextualized representation for named entity recognition.arXiv: Computation and Language,2019.Available: https://arxiv.org/abs/1911.02257.

[16] Li XY, Feng JR, Meng YX, Han QH, Wu F, Li JW. A unified MRC framework for named entity recognition. arXiv:Computation and Language,2019.Availabel:https://arxiv.org/abs/1910.11476.

[17] Liu YH,Ott M,Goyal N,Du JF,Joshi M,Chen DQ,Levy O,Lewis M,Zettlemoyer L,Stoyanov V. RoBERTa: A Robustly Optimized BERT Pretraining Approach. arXiv:Computation and Language,2019.Availabel: https://arxiv.org/abs/1907.11692.

附中文参考文献:

[6] 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.计算机应用,2005, 025(010):2404-2406.

[7] 张传岩,洪晓光,彭朝晖,李庆忠.基于 SVM 和扩展条件随机场的 Web 实体活动抽取.软件学报,2012,23(10):2612-2627.

[8] 彭嘉毅,方勇,黄诚,刘亮,姜政伟. 基于深度主动学习的信息安全领域命名实体识别研究.四川大学学报:自然科学版,2019,56: 457-462.