计算机科学
COMPUTER SCIENCE

融合多层次信息的海关同义词识别方法

刘大为, 车超, 魏小鹏

引用本文

刘大为, 车超, 魏小鹏.融合多层次信息的海关同义词识别方法[J]. 计算机科学, 2022, 49(11A): 210800197-5.

LIU Da-wei, CHE Chao, WEI Xiao-peng. Customs Synonym Recognition Fusing Multi-level Information [J]. Computer Science, 2022, 49(11A): 210800197-5.

计算机科学 COMPUTER SCIENCE
http://www.jsjkx.com
DOI:10.11896/jsjkx.210800197

融合多层次信息的海关同义词识别方法

刘大为[1]　车　超[1]　魏小鹏[1,2]
1 大连大学先进设计与智能计算省部共建教育部重点实验室　大连 116622
2 大连理工大学计算机科学与技术学院　大连 116081
(772311056@qq.com)

摘　要　在海关进出口商品文本信息中,往往会用不同的词语描述同一商品的特征,识别这些商品的特征同义词能更好地进行观点汇总,进而对同一类特征的商品进行涉税风险的防控。针对海关申报要素短语的特点,提出一种融合多层次信息的卷积神经网络模型,构建并训练了一个基于孪生和三级网络结构的 Sentence-BERT,其对相近的要素短语具有更好的语义表示,弥补了 word2vec 短文本词嵌入特征离散稀疏的不足。利用多尺寸卷积核提取要素短语的不同特征。通过 BiLSTM 神经网络学习要素短语的语序信息,并利用注意力机制分配关键词权重。获得的全连接融合同义词语义特征和关键词特征,通过 softmax 层进行预测。实验证明,融合多层次信息的卷积模型比其他模型有更好的表现。

关键词:海关商品;同义词识别;要素短语;多层次信息;卷积神经网络
中图法分类号　TP391

Customs Synonym Recognition Fusing Multi-level Information

LIU Da-wei[1],CHE Chao[1] and WEI Xiao-peng[1,2]
1 Key Laboratory of Advanced Design and Intelligent Computing,Ministry of Education,Dalian University,Dalian 116622,China
2 School of Computer Science and Technology,Dalian University of Technology,Dalian 116081,China

Abstract　In the text information of customs import and export commodities,different words are often used to describe the characteristics of the same commodity. Recognizing the characteristic synonyms of these commodities can better summarize opinions, and then prevent and control the tax-related risks for commodities with the same characteristics. According to the characteristics of phrases of customs declaration elements,a convolution neural network model fusing multi-level information is proposed,and a Sentence-BERT based on twin and three-level network structure is constructed and trained,which has a better semantic representation of similar element phrases,and makes up for the shortage of discrete and sparse embedded features of short text words in Word2Vec. Multi-size convolution kernel is used to extract different features of keywords. The BiLSTM neural network is used to learn the word order information of element phrases,and the attention mechanism is used to assign the weight of keywords. The full connection layer integrates semantic features of synonyms and keyword features,and is predicted by SoftMax layer. Experiments show that the convolution model fusing multi-level information has better performance than other models.

Keywords　Customs commodity,Synonym recognition,Element phrases,Multi-level information,Convolution neural network

1　引言

近年来,跨境电商的迅速发展,对海关涉税风险的防控手段和作业方式都提出了更高的要求。传统的监管方式无法高效地甄别高风险数据,而源源不断的海关进出口商品信息为智能化处理涉税文本提供了条件。海关进出口相关企业依照规范申报标准填写海关进出口商品信息,商品信息“规格型号”一项中包含了材质、种类、用途、成分等商品的要素短语,是商品文本信息的重要载体。对海关商品要素短语进行同义词识别分类,建立海关同义词库,将具有同义词关系的进出口商品纳入到同类商品的风险布控中,可以帮助海关提高风险布控手段,使商品的监管范围锁定在要素粒度上,对企业进出口商品信息的规范申报具有重要意义。

同章节的海关要素短语在实体标注上可以划分的类别是固定的,海关同义词识别可以建模为将具有相同属性特征的,概念上具有相同关系的要素短语归类到一起,作为短文本分类问题,在自然语言处理领域有极广的应用场景。

不同于长文本,海关要素短语文本长度短,数据稀疏。海关同义词具有如下特点:1)同义词依赖高频次的关键词信息,如减压阀的用途描述中,“压力”“减少”是高频词,如表 1 所列;2)同义词多是由短语拼接而成,具有一定的语序关系;3)同义词具有聚类特征,在关于材质的两组同义词中,“表面材质(塑料、纺织物等)”描述的是轻工业相关的材质,“外壳

基金项目:国家自然科学基金面上项目(61877008,62076045)
This work was supported by the National Natural Science Foundation of China(61877008,62076045).
通信作者:魏小鹏(adtcwxp@126.com)

材质(金属制等)"描述的是金属成分的材质。

表 1 商品要素短语

Table 1 Commodity element phrase

减压阀规格型号(用途)
用于 LNG 车用瓶供气系统及其他气体管路压力的调节和稳定,调节管道压力;
用途:用于气控回路中,降低出口压力,保证出口压力稳定;用于启动装置需降压力的场合;
应用于液压系统,减小系统压力;
应用于 Arriel 2 型直升机发动机;
用途:安装在连接左右空气弹簧的连接管之中间部位;
用途:通过调节阀口开合度改变流体动能,从而达到减压的目的;
用途:减少管道内的压力;

对于海关申报要素短语来说,海关同义词在空间中具有相近的语义关系,非同义词则相距较远。Sentence-BERT 可以将同义词映射到相近的向量空间中,我们选取 Sentence-BERT 句向量提取要素短语的同义词信息,采用 word2vec 提取要素短语的词向量特征,分别构成了句子级和词汇级的向量表示。为了优化深度学习模型提取海关同义词特征的能力,本文提出了融合多层次信息的卷积神经网络模型 FMICNN(Convolutional Neural Network Fusing Multi-level Information)。我们利用双通道卷积提取同义词的词语和句子特征并进行拼接,再与 BiLSTM+Attention 并行提取要素短语的语序特征和关键词特征进行融合。FMICNN 解决了要素短语稀疏造成的同义词识别困难问题,提高了同义词识别的准确率。

2 相关工作

同义词识别的关键是学习有效的特征表示。传统的向量空间模型与同义词识别直接相关,然而基于词袋模型的文本挖掘方法在建模中遇到数据稀疏和歧义问题,忽略了词与词之间的语义关系。在基于神经网络方法中,Fei 等[1]扩展了 word2vec 的 skip-gram 模型,利用分层的多任务学习语义类型和语义关系,得到了大量新的医学术语同义词对。然而海关缺乏类似 UMLS 那样的外部知识库,无法获取更多的语义知识。近年来研究的注意力从词嵌入转到向量表征,自然语义处理领域中无法忽略的 BERT[2]在语义识别中有大量应用。由于 BERT 对单个句子的表示结果不佳,无法适用于大规模语义相似度检测任务[3],Reimers 等[4]提出了基于孪生网络结构的双编码器模型 Sentence-BERT。相较于 BERT,训练过的 Sentence-BERT 作为一种句向量编码方式,其单个句子在向量空间的表示中有更加丰富的同义词关系,在同义词识别、语义检测等任务中有更好的表现。

同义词识别任务可以转换为深度学习神经网络的多个范型。Hasan 等[5]使用字符的嵌入特征,将同义词任务转换为神经网络的机器翻译问题,在确定输入序列的情况下,通过双向 RNN 生成目标同义词。Zhao 等[6]分析并生成了船舶行业的同义词标记训练样本,通过训练 LSTM+CRF 将同义词抽取转化为序列标注问题,但标记同义词依赖第三方知识。Zhang 等[7]提出了一种可以组合多种不同的词嵌入的卷积网络模型 MGNC-CNN(Multi-Group Norm Constraint CNN),并对各自的词嵌入生成的特征权重施加不同的正则化惩罚项,但并没有分析不同的嵌入方式对具体任务的影响。

这种基于改进的多通道卷积模型在情感分类[8]等多个分类任务中取得了不错的效果,受其启发,我们将结合词向量与句向量的多通道卷积模型应用到同义词识别任务中。一些研究者引入了 CNN 和 RNN 的混合框架,然而,大多数混合框架平等地对待所有的词,而忽略了不同的词对文本语义有不同的贡献,对此在同义词模型中引入注意力机制;利用 BiLSTM 提取文本的语义特征,通过注意力机制学习单个字或词在整个同义词中的重要程度,捕获关键词信息。

3 FMICNN 模型

FMICNN 总体模型如图 1 所示,分为向量表示层、通道卷积层和多层次信息融合层 3 部分。

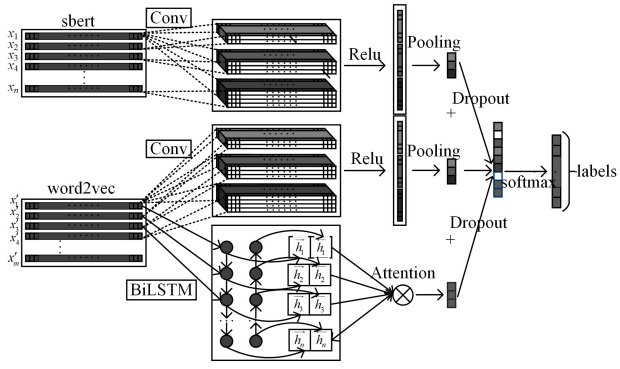

图 1 FMICNN 网络模型结构

Fig. 1 FMICNN network architecture

3.1 向量表示层

向量表示层包含两个不同大小的向量矩阵$\{x_1, x_2, \cdots, x_i, \cdots, x_n\}$和$\{x_1', x_2', \cdots, x_3', \cdots, x_m'\}$,其分别表示 Sentence-BERT 和 word2vec 训练的句向量和词向量,简称 sbert 和 w2v。其中,x_i 和 x_i'的长度是 d 和 d'。sbert 针对每个要素短语生成的是一维的句向量,为了适应多个尺寸的 2D 卷积滤波器,我们将 sbert 向量转换成二维矩阵作为单个通道的输入。sbert 句向量矩阵的维度大小为 $d \times n$,word2vec 词嵌入矩阵维度大小为 $d' \times m$。d'是 w2v 单个词的维度,m 是 w2v 海关短文本词数(补位 padding 填 0)。模型的输入层是两个预训练向量,不需要更新参数,它们不被计入模型总的训练时间复杂度中。

3.1.1 word2vec

受神经网络语言模型 NNLM 启发,Mikolov 于 2013 年提出了一款向量计算工具 Word2vec[9]。本文采用 Skip-gram 进行从目标词到上下文的自监督学习,使用 word2vec 进行词向量表示,很好地保留了文本语义信息,但未能解决一词多义问题,原始空间特征维度较大,忽略了语序关系。此外,word2vec 是以词为单位的向量表示方法,对文本的分词结果直接影响词向量表示的准确性,而高密度语义信息的句向量表示是对其较好的补充。

3.1.2 Sentence-BERT

sbert 生成有相近语义关系的句子嵌入,用于大规模语义相似度比较。其整体框架如图 2 所示,图 2(a)分别将表示向量 u, v 及其差异 $|u-v|$ 串联并传递给 softmax 分类器,如式(1)所示:

$$o=softmax(W_t[u\oplus v\oplus|u-v|]) \tag{1}$$

其中，$\oplus$表示向量拼接，W_t 表示可训练权值。文本采用图 2(b)表示计算余弦相似度的回归模型，计算两个句子 u 和 v 之间的余弦相似度，采用均方误差的损失函数，通过梯度下降更新参数最小化损失函数。

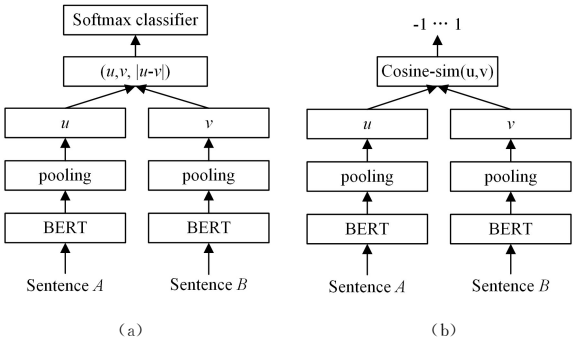

图 2　Sentence-BERT 模型结构

Fig. 2　Sentence-BERT network architecture

本文构建的 sbert 采用谷歌中文预训练模型作为编码器模型，最大句长 256 维，超过部分将被截断，采用 mean-pooling 平均池化方式，加入了一个损失函数为 Tanh 的 512 维全连接层作为句向量的最终表示。训练 sbert 输入两个句子，并根据两个句子的实际接近程度输出两个句向量表示。训练过程中，模型的编码器部分是权重共享的，这也是“孪生”网络的体现之处。相比于 BERT 这种交叉编码器，SBERT 可以帮助我们生成具有空间聚类信息的句向量。

3.2　通道卷积层

通道卷积层如图 3 所示，采用 3 种不同尺寸的卷积核来提取 n-gram 词袋特征，重点是捕捉短距离词袋信息。

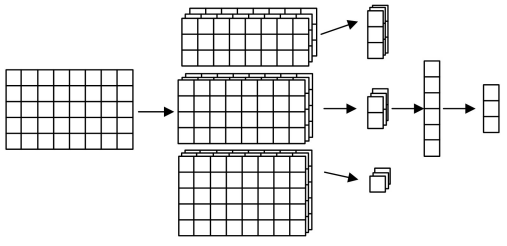

图 3　TextCNN 模型结构

Fig. 3　TextCNN network architecture

类似图像多通道卷积中的 R，G，B 三元组，FMICNN 模型用两个预训练向量作为双通道，由于两个输入的预训练向量矩阵大小不同，所以分别采用不同卷积核完成卷积和池化操作。每个 batch 大小为 64，单个通道包含了 3 组不同大小的卷积核，不同通道的卷积核长度为 d 和 d'，3 组卷积核的宽度分别为 3，4，5。以 sbert 句向量为例，矩阵在完成卷积后 3 组输出的向量为 $64\times(n-2)$，$64\times(n-3)$，$64\times(n-4)$，每组特征向量经过最大池化的选择后拼接在一起，成为 64×3 的隐藏层向量，最终得到 sbert 和 w2v 两个通道的特征表示。

3.3　多层次信息融合层

3.3.1　BiLSTM+attention

同义词 w2v 的词嵌入向量作为 BiLSTM+attention 的输入，模型示意图如图 4 所示，图中输入层向量在经过 BiLSTM 编码后得到输出隐藏层，经过字/词级别的 attention 机制相加求和输出 Y。BiLSTM 模型 t 时刻的计算如下：

$$h_t=[\overleftarrow{h_t},\overrightarrow{h_t}] \tag{2}$$

其中，$\overrightarrow{h_t}$，$\overleftarrow{h_t}$ 分别表示前向 LSTM 和后向 LSTM 在 t 时刻的输出，h_t 表示两个向量的拼接。

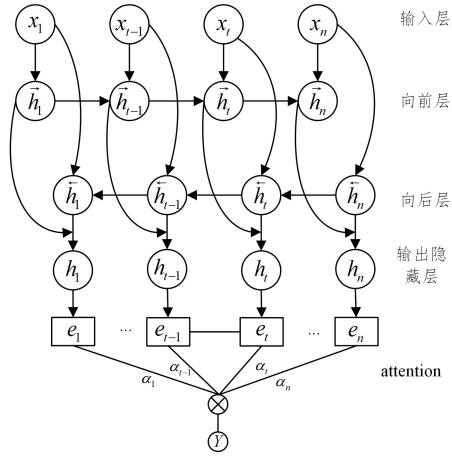

图 4　BiLST+attention 模型结构

Fig. 4　BiLSTM+attention network architecture

Attention 机制在 NLP 领域应用十分广泛，即在上下文中分配关键词的权重，从而捕捉到重要的特征。将 h_t 输入到注意力机制中得到初始状态向量 e_t，而后与归一化的权重系数 a_t 对应相乘并累加求和得到最终输出的向量 Y。计算公式如下：

$$e_t=\tanh(Wh_t+b_t) \tag{3}$$

$$a_t=softmax(e_t) \tag{4}$$

$$Y=\sum_{t=1}^{n}a_th_t \tag{5}$$

根据式(4)可以实现由输入初始状态到含有关键词特征的转换，之后通过式(5)得到最终输出的状态向量 Y。

3.3.2　特征融合

将 w2v 送入 BiLSTM+attention 中，要素短语的语序具有规格型号的排列顺序。BiLSTM 提取同义词的语序特征，输出的隐藏层向量经过 attention 层提取关键词特征，输出后的向量 w_3 与双通道卷积处理后的 w2v 和 sbert 的隐藏层向量 w_1 和 w_2 进行拼接，送入到 softmax 分类器之前，经过 dropout 层随机遮盖 10% 的数值防止过拟合，得到 FMICNN 最终融合 3 组特征向量的向量 D，具体如式(6)所示：

$$\begin{cases}w_1=\text{TextCNN(w2v)}\\ w_2=\text{TextCNN(sbert)}\\ w_3=\text{Attention(BiLSTM(w2v))}\\ D=\text{Concat}[w_1,w_2,w_3]\end{cases} \tag{6}$$

输出层的表示如式(7)所示，W_d 表示状态层到输出层的训练权重，b 为对应的偏置，经过 softmax 后得到最终的分类结果。

$$y=softmax(W_dD+b) \tag{7}$$

4　实验

4.1　数据集

我们从海关进出口商品报关文本中提取了同种商品要素类别下的要素短语，经过海关专家的确认，确定了几个关键要素类别下的同义词。训练集、验证集和测试集的比例是 8∶1∶1，用于 4.4.2 节同义词分类任务，同义词数据集如表 2 所列。

表 2 海关要素同义词语料

Table 2 Custom elements words corpus

同义词类别	同义词	数量
牛肉部位	“部分牛前部位肉”“牛肋排 5 骨”“牛三角肉”	1 595
牛种(安格斯牛、和牛等)	“海伏特”“英国品种,主要是海福特牛”“赫里富”	539
加工方法(整头半头带骨去骨等)	“去骨 全身各部位”“去骨,腰臀部,未炼制,瘦肉约战 10%”	1 603
种类(舌、肝、心管、板筋等)	“猪后脚 短切 15～20 cm”“后蹄,20 cm 以上长切带趾”	1 206
材质构成(是否加强或与其他材料合制)	“硫化橡胶与尼龙合制”“硫化橡胶与尼龙纺织材料合制”“硫化橡胶与尼龙合制,三层橡胶和两层尼龙线”	1 331
表面材质(塑料、纺织物等)	“70%棉 10%牛皮 10%锦纶 10%金属镀膜纤维”“31% 棉 24% 锦纶 21% 聚酯纤维 14% 涤纶 10% 小牛皮”	1 471
材质(里、外)	“外:赤狐毛坯,里:百分之六十铜氨纤维,百分之四十棉”“100%袋鼠毛皮 MACROPUS RUFUS,里料:100%聚酯纤维”	1 546
外壳材质(金属制等)	“LTP-1391L-4AVDF 机械指示式电子手表”“不锈钢表壳,牛皮表带,电子驱动,指针指示”	1 341

4.2 训练向量

在 word2vec 的训练方面,为了模拟真实海关进出口商品信息不断录入更新的业务场景,我们随机抽取了训练集 30% 的要素短语,去除特殊字符、全角转半角正则化处理,并设置了停用词,对 word2vec 进行了训练。

为了让 Sentence-BERT 更好地掌握要素下的语义关系,我们根据要素类别的关键词信息建立了同义词打分规则。首先选取 8 个类别下共 1 000 条要素短语两两组合成约 45 万条同义词对,然后根据建立的基于关键词的打分规则进行批量打分操作。对 Sentence-BERT 预训练模型进行微调,计算余弦相似度,采用均方差损失函数(见式(8)),对 45 万条同义词对进行了 13 个小时的训练。

$$J_{\text{mse}}=\frac{1}{N}\sum_{i=1}^{n}(y_i-\tilde{y}_i)^2 \tag{8}$$

4.3 参数设置

FMICNN 模型采用 tensorflow-keras 框架,卷积核的宽度决定了提取海关要素短语特征的粒度,设置不同组的卷积核宽度,当组合卷积核宽度为 2,3,4 时模型取得了分类精度的极值。采用 Adam 优化器和交叉熵的损失函数,在 10 个 Epoch 内模型精度没有提升则训练提前停止,相关参数如表 3 所列。

表 3 实验模型参数

Table 3 Experimental model parameters

Parameter	Value
Epoch	30
Kernel_size	2,3,4
Kernelnumber	64
Loss	Cross Entropy
Batch_size	64
Word2vec dimension	150 * 300
Sbert dimension	16 * 32
BiLSTM hidden_size	64
Attention_size	32

4.4 对比实验

我们对 FMICNN 同义词识别模型从向量表示和模型结构两方面进行了实验评估。

4.4.1 向量同义词识别

我们用不同的嵌入方式训练同义词向量,以验证本文采用的词向量和句向量的表示质量。Glove[10] 采用和 word2vec 相同训练集;bert-as-service 是 BERT 的句向量表示服务,采用和 Sentence-BERT 相同的语料微调。我们从海关同义词集“牛肉部位”中进一步细分出 10 组不同牛肉部位的同义词(牛前后部位肉、牛四分体肉、牛外脊肉、牛眼肉、牛上脑、肩部肉、牛腿肉、牛腱带肉、牛椎骨、牛柳),给定“牛肩肉”作为目标词,“牛肉部位”词集作为候选词,任务的目标是从“牛肉部位”同义词集中找出与“牛肩肉”最相似的成员。向量间的相似度计算公式如下所示:

$$Sim(w,v)=\text{cosine}(\sum_{i=1}^{n}w_i,\sum_{j=1}^{n}v_j) \tag{9}$$

其中,n 是 word2vec 和 Glove 词向量的句长,w 和 v 是按维度相加再进行余弦计算。我们根据“牛肩肉”计算每个候选词的余弦相似度来比较其排名的前 k 个词的精度。

表 4 列出了目标词在候选词中相似度计算的余弦相似度,其中 $Precision=tp/k$,tp 是相似度最高的前 k 个候选词中是“牛肩肉”同义词的数量。

表 4 $k=3,5,7$ 时同义词识别的精度

Table 4 Accuracy of synonym recognition when $k=3,5,7$

Model	*Precision*		
	$k=3$	$k=5$	$k=7$
Glove	0	0.14	0.29
Word2vec	**0.33**	**0.20**	**0.29**
bert-as-service	0.67	0.60	0.43
sentence-BERT	**1.00**	**0.80**	**0.71**

从表 4 中可以发现 Glove 和 word2vec 的精度较差,这说明词向量表示依赖分词结果,且同义词短语稀疏,全局特征的表示能力较弱,Sentence-BERT 的同义词向量表示在同义词识别中得到了最好的效果。

4.4.2 模型同义词识别

为了验证文本提出的 FMICNN 模型,我们将其与主流的短文本分类模型 TextRCNN[11],BiLSTM + Attention[12] 和 BERT[2] 在海关数据集上进行了性能对比,实验结果如表 5 所列。结果显示,FMICNN 在同义词识别上的精确率(Precision)、召回率(Recall)和 F 值(F-Score)取得了最佳的表现,相比 TextRCNN 和 BiLSTM+attention 传统短文本分类模型有明显提升,与 BERT 相比精度提高了 2%。

表 5 FMICNN 与其他模型的性能对比

Table 5 Performance comparison between FMICNN and other models

model	*Precision*	*Recall*	*F-Score*
TextRCNN	0.876	0.900	0.888
BiLSTM+att	0.862	0.881	0.871
BERT	0.946	0.961	0.953
FMICNN	**0.960**	**0.963**	**0.961**

4.5 消融实验

对 FMICNN 模型的各部分子模型进行单独的验证,结果如图 5、图 6 所示。w/o w2v 是去除 w2v 词向量通道保留 sbert 句向量通道的 textCNN 模型;w/o sbert 是去除 sbert 句向量通道的 FMICNN 模型;w/o bilstm + att 是 FMICNN

模型去除了 BiLSTM＋attention 层，以 w2v 和 sbert 作为向量的双通道 textCNN 模型；FMICNN 是本文提出的融合多层次信息的卷积神经网络。它们在 30 个 epoch 总共 5 000 次训练次数的 Loss 曲线如图 5 所示，可以看到，采用字词结合的双通道卷积收敛效果要明显好于单个通道的表现。

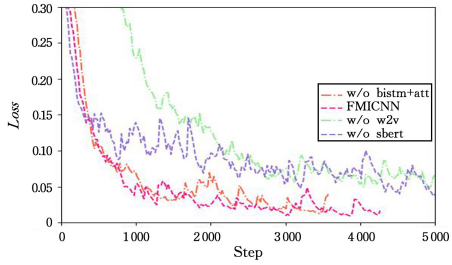

图 5　FMICNN 子模型的 Loss 曲线

Fig. 5　Loss curve of FMICNN submodel

4 个子模型的准确率如图 6 所示，其中 w/o sbert 上升趋势平缓，w/o w2v 起点低，随后上升趋势明显，说明了 sbert 对准确率的结果影响较大。双通道模型 w/o bilstm＋att 的准确率超过了 w/o w2v 和 w/o sbert 的最高表现，最终在 3 000 次训练后没有提升而训练停止，本文提出的 FMICNN 取得了最高的准确率。

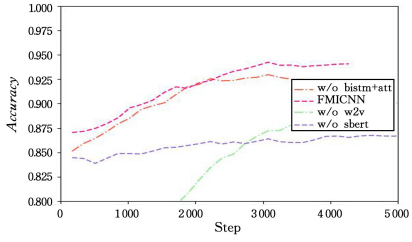

图 6　FMICNN 子模型的同义词识别准确率

Fig. 6　Synonym recognition accuracy of FMICNN submodel

此外，我们发现 attention 机制加快了模型训练的收敛速度，进行 10 次重复实验后发现，模型平均在 20 个 epoch 左右收敛，迭代次数低于 w/o attention，得到了最高的准确率。

表 6 列出了各项子模型在预测集上的 F 值，反映了各个子模型对模型整体效果的贡献程度。

表 6　FMICNN 子模型在预测集上的 F 值

Table 6　F-Score of FMICNN submodel prediction set

sub-model	*F*-score
w/o w2v	0.8135
w/osbert	0.8593
w/obilstm＋attention	0.9642
FMICNN	**0.9679**

结束语　本文提出的融合多层次信息的卷积神经网络模型 FMICNN，利用词句结合的向量表示方式，结构上融合了海关同义词的语料特征，解决了海关同义词识别困难的问题。同时我们也看到，训练句向量需要对同义词对进行标注，模型训练需要大量同义词类别标记。因此，对要素短语进行无监督聚类，是海关同义词识别下一步的研究方向。

参 考 文 献

[1] FEI H, TAN S, LI P. Hierarchical multi-task word embedding learning for synonym prediction[C]// Proceedings of the 25th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining. 2019: 834-842.

[2] DEVLIN J, CHANG M W, LEE K, et al. Bert: Pre-Training of Deep Bidirectional Transformers for Language Understanding [C]// Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics. Minneapolis, Minnesota, 2019: 4171-4186.

[3] CHANG W C, YU F X, CHANG Y W, et al. Pre-training tasks for embedding－based large－scale retrieval [C]// Proceedings of the 8th International Conference on Learning Representations (ICLR). 2020.

[4] REIMERS N, GUREVYCH I. Sentence-Bert: Sentence Embeddings Using Siamese Bert-Networks [C]// Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing. 2019: 3973-3983.

[5] HASAN S A, LIU B, LIU J, et al. Neural clinical paraphrase generation with attention[C]// Proceedings of the Clinical Natural Language Processing Workshop (ClinicalNLP). 2016: 42-53.

[6] ZHAO Y, LIU Q, HU F, et al. Synonym Extraction in Shipping Industry with Distant Supervision and Deep Neural Network [J]. Journal of Coastal Research, 2019, 94(sp1): 455-459.

[7] ZHANG Y, ROLLER S, WALLACE B. MGNC-CNN: A simple approach to exploiting multiple word embeddings for sentence classification [C]// Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. 2016: 1522-1527

[8] YU W, ZHENNI Z, JIE Y, et al. Weibo Sentiment Classification Based on Two Channels Text Convolution Neural Network with Multi-Feature [C]// 2020 International Conference on Cyber-Enabled Distributed Computing and Knowledge Discovery (CyberC). 2020: 152-60.

[9] MIKOLOV T, CHEN K, CORRADO G, et al. Efficient Estimation of Word Representations in VectorSpace[C]// Proceedings of the 1st International Conference on Learning Representations. Scottsdale, USA, 2013.

[10] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). 2014: 1532-1543.

[11] LAI S, XU L, LIU K, et al. Recurrent convolutional neural networks for text classification[C]// Proceedings of the Twenty-ninth AAAI Conference on Artificial Intelligence. 2015: 2267-2273

[12] ZHOU P, SHI W, TIAN J, et al. Attention-based bidirectional long short-term memory networks for relation classification [C]// Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. 2016: 207-212.

LIU Da-wei, born in 1993, postgraduate. His main research interests include synonym recognition and so on.

WEI Xiao-peng, Ph. D, professor. His main research interests include medical and health informatics, computer animation, computer vision, and intelligent CAD.