基于Transformer和BERT的名词隐喻识别*

张冬瑜[1] 崔紫娟[2] 李映夏[1] 张 伟[1] 林鸿飞[3]

[1](大连理工大学软件学院 大连 116620)

[2](大连理工大学国际合作与交流处 大连 116024)

[3](大连理工大学计算机科学与技术学院 大连 116023)

摘要:【目的】解决名词隐喻识别研究中语义信息利用不足和关系表征的问题,提高识别效果。【方法】使用BERT模型替代词向量,在语义表示中同时包含词与词之间的位置关系等信息,利用Transformer模型进行特征提取并通过神经网络分类器进行识别。【结果】本文模型在准确率(0.900 0)、精确率(0.896 4)、召回率(0.885 8)和F1值(0.891 0)4个指标上均表现最好,可以注意到多个关键点信息,提高名词隐喻的分类效果。【局限】对于中文文本中的冷僻词汇、成语古语以及干扰词汇等的判断比较困难。【结论】本文所提隐喻识别方法优于现有基于人工特征的分类模型及主流深度学习模型。

关键词: 隐喻识别 名词隐喻 语义理解 Transformer模型 BERT模型

分类号: TP391

DOI: 10.11925/infotech.2096-3467.2019.0896

引用本文: 张冬瑜, 崔紫娟, 李映夏等. 基于Transformer和BERT的名词隐喻识别[J]. 数据分析与知识发现, 2020, 4(4): 100-108. (Zhang Dongyu, Cui Zijuan, Li Yingxia, et al. Identifying Noun Metaphors with Transformer and BERT[J]. Data Analysis and Knowledge Discovery, 2020, 4(4): 100-108.)

1 引 言

隐喻普遍存在于人类语言中,在日常交流中使用频率非常高,每三句话中就可能出现一次,人一生平均使用大约2 100多万次隐喻[1-2]。隐喻是思维问题,不仅仅是修辞方式,更是一种认知手段,本质上是从具体的概念域向抽象的概念域的系统映射[1]。认知是指人们获得知识或应用知识的过程,是人类信息加工的基本过程[3]。随着社交媒体的迅速崛起以及跨语言文化交流的持续深入,隐喻出现在更加多元化的平台上,也得到了更加广泛的关注。

隐喻识别旨在赋予计算机像人类一样分析理解隐喻的能力,涉及计算机科学、认知科学等多个学科的交叉,具有极其重要的理论和实际应用价值[4]。如果隐喻识别问题不能很好地解决,将会影响语义的理解以及应用系统性能的提升。例如在情感分析、阅读理解、人机对话、文本摘要、文本生成和机器翻译等领域,隐喻理解直接影响到处理的精度。

按照句法构成特点,隐喻一般分为名词性隐喻、动词性隐喻、形容词性隐喻、副词性隐喻等类型。由于名词性隐喻在自然语言使用中所占比重较大,目前大量隐喻识别研究围绕名词隐喻开展。名词隐喻指自然语言表达中通过连接词表征的隐喻类型,其源域与目标域词汇通常以名词的形式出现在句子中,识别的难点主要在于如何准确定位源域词汇与目标域词汇(也称本体和喻体),进而推断具有隐喻义的喻底。常见形式是通过“是、像、宛如”等相连,

通讯作者:林鸿飞,ORCID:0000-0003-0872-7688,E-mail:hflin@dlut.edu.cn。

*本文系教育部人文社会科学基金项目“基于机器学习的情感隐喻识别研究”(项目编号:16YJCZH141)、国家自然科学基金项目“基于语义资源和深度学习的情感隐喻识别方法研究”(项目编号:61602079)和国家自然科学基金重点项目“社交媒体中文本情感语义计算理论和方法”(项目编号:61632011)的研究成果之一。

如“她像一朵鲜花”为名词隐喻，喻词为“像”，本体为“她”，喻体为“鲜花”，本体属于人类，喻体属于植物，是不同领域之间的映射。

隐喻识别及其应用研究刚刚起步，吸引了大量计算机领域的研究者开展相关研究，研究成果发表于ACL、EMNLP、COLING和NAACL等顶级国际会议。名词隐喻研究主要采用简单的特征工程或者深度神经网络模型进行识别任务，着重强调对隐喻的识别。但是目前对名词隐喻的语义表示不够充分，无法提供更为精细的信息对字面义和隐喻义加以区分。因此，如何从现有的语义资源和上下文神经语言环境中提炼出较为丰富的语义信息，从局部和全局的角度，提供细粒度的名词语义和属性的表示，提高名词隐喻的识别效果是亟待解决的问题。

本文提出一种多维语义表示机制，采用自注意力机制学习隐喻的潜在语言特征，充分利用已有隐喻语义资源，结合上下文语言模型的预训练，着重解决名词隐喻语义表示不充分的问题，构建基于深度语义知识的名词隐喻识别模型。具体而言，本文提出基于BERT（Bidirectional Encoder Representation from Transformers）与Transformer模型的名词隐喻识别方法，使用BERT模型替代词向量，在语义表示中同时包含词与词之间的位置关系等信息，利用Transformer模型进行特征提取并通过神经网络分类器进行识别。通过使用经过预训练的BERT模型，可以将文本转化为计算机能够处理的数学形式，其中包含文本所含有的深度信息。相比于传统的文本表示方法，BERT模型能够对句子中词的位置关系进行处理，并在形成的向量和矩阵中体现。而Transformer模型是一种新的深度神经网络模型，相比于传统的深度神经网络，如CNN、RNN等，Transformer抛弃了传统的神经网络结构而仅使用注意力机制构建网络，具有注意到多个关键点信息的能力，被广泛应用于文本分类模型中。

本文所提出的模型通过预训练的BERT模型提取文本所蕴含的深度信息。其中不仅包括单词之间的相对位置关系，而且蕴含名词隐喻中本体和喻体之间的关联性和差异性。随后通过训练Transformer模型，对文本中的深度信息进行挖掘和提取，能够有效建立本体和喻体之间的映射关系，从而提高名词隐喻的识别效果。

2　相关研究

随着深度学习技术的不断发展，越来越多的深度学习模型被提出，尤其是Transformer和BERT模型，在NLP领域中具有广泛的应用。Brunner等[5]证明了Transformer模型中自注意力机制的有效性，为提高Transformer模型的可解释性做出贡献。Khandelwal等[6]通过使用单独的预训练Transformer模型，构建一个文本摘要生成系统，在降低系统开销的同时实现出色的效果。Liu等[7]通过Transformer模型与RNN模型结合，搭建文章结构解析模型，对文章结构进行解析。Yang等[8]利用经过微调的BERT模型，在问答领域较之前的最优方法达到接近30%的提升效果。Alberti等[9]使用BERT模型，通过检索和答案判断的方法，较之前最优方法在短答案类型数据上效果提升50%以上，在长答案类型上提升30%以上。Xu等[10]利用BERT模型的深层语言处理能力，改进现有的排序模型，在不同数据集上效果提升30%-40%。

目前针对名词隐喻识别的研究工作对特征层面关注较多，早期研究中对语义知识规则和特征工程构建的方法较多。Fass[11]是最早对隐喻文本进行识别和理解的学者之一，通过语义规则识别隐喻文本。而Wilks等[12]在Fass的研究基础之上，使用VerbNet和WordNet两种语义资源获取语义规则，用以自动识别隐喻。Jang等[13]利用WordNet和FrameNet提供的分类信息以及LDA模型提供的主题分布信息，提出全局上下文特征和局部上下文特征，对隐喻进行识别。Rai等[14]构建一种图模型检测上下文中的不一致性，在VUAMC和TroFi数据集上进行名词隐喻识别，准确率提升近6%。上述研究方法均需要构建大量语义规则和特征，需要耗费极大的人力。

随着深度学习的飞速发展，基于深度学习的方法在名词隐喻识别任务中得到较快发展。Do Dinh等[15]在词向量的基础之上，提出一种神经网络模型识别隐喻，减少了对外部资源的依赖。Sun等[16]充分考虑隐喻的序列特性和上下文依赖特性，将文本输入Bi-LSTM模型抽取特征从而识别隐喻。汪梦翔等[17]将隐喻知识看作是本体和喻体的特征和属性

之间的关联，通过惯用语导入和句法模式识别机制获取名词的隐喻知识。Bizzoni 等[18]构建词级别的 Bi-LSTM 模型，然后构建基于结构的循环图结构模型用以识别隐喻。Gao 等[19]在 Bi-LSTM 模型的基础之上，提出一种端到端的上下文建模神经网络模型用于识别隐喻。

上述研究仅采用简单的特征工程或者深度神经网络模型进行名词隐喻的识别任务，其语义表示信息不够丰富，不足以分辨出名词隐喻的意义，需要根据不同的知识来源，结合上下文语言环境，从不同的维度获取语义表示信息，对名词隐喻的表示学习进行合理建模。研究证明，名词隐喻的识别和应用具有一定研究基础和较强的应用需求。但由于名词隐喻表达的复杂性和含蓄性，如何准确识别自然语言表达中的名词隐喻是一项亟待研究的课题。

3 研究思路与框架

3.1 基于 Transformer 和 BERT 的名词隐喻识别

本文提出一种基于 Transformer 和 BERT 的名词隐喻识别方法，主要过程如图 1 所示。

图 1 BERT+Transformer 的名词隐喻识别流程

Fig.1 Noun Metaphor Identification Process of BERT+Transformer Model

预训练的 BERT 模型用于构建计算机可以识别的句子矩阵表示，其中蕴含着本体和喻体之间的关联性与差异性。使用 Transformer 模型，以句子矩阵作为输入，经过训练集的训练，可以学习到本体和喻体之间的映射规律，最终输出句子的分类结果。对于输入的中文句子，经过谷歌提供的预训练的 BERT 模型，得到句子的向量表示，其中包含着句子文本的深层语义信息。然后经过 Transformer 模型对文本深层语义信息中包含的隐喻特征进行抽取，并利用标签进行训练和学习。最后通过 Sigmoid 函数激活，得到一个 0 到 1 的函数，大于 0.5 是名词隐喻，否则不是名词隐喻，完成对于名词隐喻任务的分类。

Transformer 中使用的 *Attention* 函数是基本的点乘方式，计算过程如公式（1）所示。

$$Attention(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Softmax(\frac{\boldsymbol{Q}\boldsymbol{K}^{\mathrm{T}}}{\sqrt{d_k}})\boldsymbol{V} \quad (1)$$

输入包括维度为 d_k 的序列以及键，还有 $\boldsymbol{V}$ 的维度值。计算序列和所有键的点乘，均除以 $\sqrt{d_k}$。利用 *Softmax* 函数获取值的权重。实际操作中，*Attention* 函数是在一些序列上同时进行的，将这些序列并在一起形成矩阵 $\boldsymbol{Q}$，同时键以及值形成矩阵 $\boldsymbol{K}$ 和 $\boldsymbol{V}$。多头 *Attention* 的计算过程如公式（2）所示。

$$MultiHead(\boldsymbol{Q},\boldsymbol{K},\boldsymbol{V}) = Concat(head_1,\cdots,head_n) \quad (2)$$

Concat 函数用于将不同的 *head* 连接起来形成新的向量或者矩阵，*head* 的计算过程如公式（3）所示。

$$head_i = Attention(\boldsymbol{Q}W_i^{Q},\boldsymbol{K}W_i^{K},\boldsymbol{V}W_i^{V}) \quad (3)$$

其中，W_i 为第 i 个单词，$\boldsymbol{Q}$、$\boldsymbol{K}$、$\boldsymbol{V}$ 为前文所介绍的矩阵。

对于最后的前馈网络层，本文也给出其定义，计算过程如公式（4）所示。

$$FFN(x) = \max(0, xW_1 + b_1)W_2 + b_2 \quad (4)$$

其中，W_1、b_1、W_2、b_2 分别为权重与偏差，由算法自动调整优化。

BERT 将在下游 NLP 任务中的操作转移到预训练词向量中，在获得 BERT 词向量后，最终只需在词向量上加简单的 MLP 或线性分类器即可。但笔者研究发现，BERT 与 Transformer 结合对于名词隐喻识别任务效果会更好。对于隐喻情感识别任务来说，使用 Attention 机制可以使模型更多地关注本体与喻体之间的差异性与相似性，能够更好地建立从本体到喻体的映射关系，提升名词隐喻识别的效果。BERT 与 Transformer 都是基于多头注意力机制的模型，BERT 抽取深层语义信息，Transformer 学习隐喻特征，二者相辅相成可以达到更好的识别效果。

3.2 BERT 预训练模型

BERT 是由 Devlin 等[20]提出的全新的预训练的语言模型表示。BERT 模型沿袭了 GPT（Generative Pre-Training）模型的结构，采用 Transformer 的编码

器作为主体模型结构。

GPT 利用 Transformer 的结构进行单向语言模型训练。所谓的语言模型其实是自然语言处理中的一种基础任务，其目标是给定一个序列文本，预测下一个位置上会出现的词。而 BERT 对 GPT 的第一个改进就是引入双向的语言模型任务。

BERT 是一种预训练语言表示的方法，意味着在大型文本语料库（如维基百科）上训练通用“语言理解”模型，然后将该模型用于下游的 NLP 任务，如问答、情感分析、文本聚类等。BERT 优于之前的方法，因为它是第一个用于预训练 NLP 的无监督（Unsupervised）且深度双向（Deeply Bidirectional）的系统。BERT 模型的训练过程如图 2 所示，*Ei* 为单词的编码表示，*Trm* 为 Transformer 结构，T_i 为训练好的目标单词的词向量。使用 Masked LM 和 Next Sentence Prediction 两种方法分别捕捉词语和句子级别的表示。其中“双向”表示模型在处理某一个词时，能同时利用前面的词和后面的词两部分信息，这种“双向”不是在给定所有前面词的条件下预测最可能的当前词，而是随机遮掩一些词，并利用所有没被遮掩的词进行预测。

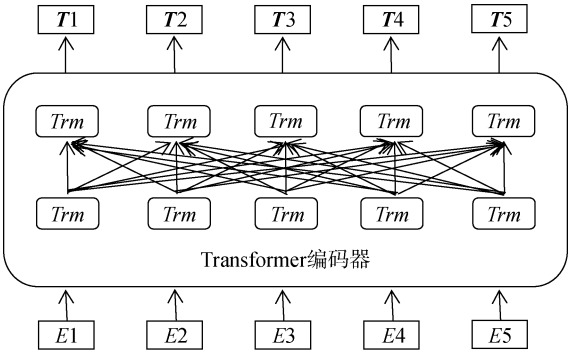

图 2　BERT 模型训练过程

Fig.2　Training Process of BERT Model

BERT 模型能进一步增加词向量模型泛化能力，充分描述字符级、词级、句子级甚至句间关系特征。具有如下三个新特点：

（1）真正的双向 Encoding：Masked LM，类似完形填空，尽管仍旧看到所有位置信息，但需要预测的词已被特殊符号代替，可以双向 Encoding。

（2）Transformer 做 Encoder 实现上下文（Context）相关：使用 Transformer 而不是 Bi-LSTM 做 Encoder，可以有更深的层数、具有更好并行性。并且线性的 Transformer 比 LSTM 更易免受 Mask 标记影响，只需要通过 Self-Attention 减小 Mask 标记权重即可，而 LSTM 类似黑盒模型，很难确定其内部对于 Mask 标记的处理方式。

（3）提升至句子级别：学习句子/句对关系表示，句子级负采样。给定的一个句子，下一句子正例（正确词），随机采样一句负例（随机采样词），句子级上做二分类（即判断句子是当前句子的下一句还是噪声），类似 Word2Vec 的单词级负采样。

3.3　Transformer 模型

Vaswani 等[21]提出一个仅基于 Attention 结构的 Transformer 模型处理序列编码问题。传统的神经网络模型大都是利用 RNN 或者 CNN 作为 Encoder-Decoder 的模型基础[15]，而 Transformer 模型摒弃了固有的定式，并没有使用任何 CNN 或者 RNN 结构。Transformer 模型可以高度并行地工作，所以在提升性能的同时训练速度也较快，结构如图 3 所示。

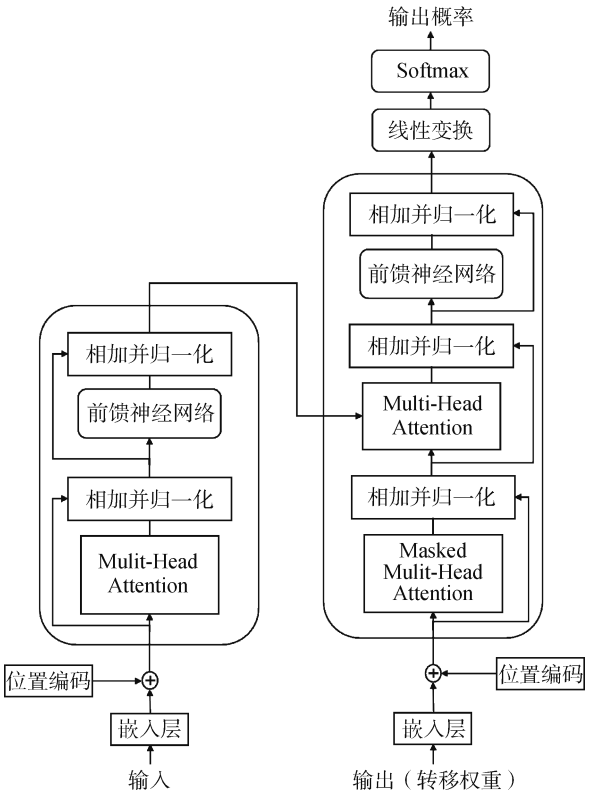

图 3　Transformer 模型结构

Fig.3　Structure of Transformer Model

模型分为编码器和解码器两个部分。编码器由6个相同的层堆叠在一起,每一层又有两个支层。第一个支层是一个多头注意力机制(Muti-Head Attention),第二个支层是一个简单的全连接前馈网络。在两个支层外面都添加了一个残差(Residual)连接,然后进行归一化操作。模型所有支层以及嵌入层的输出维度都是d。解码器也是堆叠了6个相同的层,每层除了编码器中那两个支层,还加入第三个支层,同样也用了残差连接以及归一化。

编码器和解码器的输入就是利用学习好的向量将Tokens(一般应该是词或者字符)转化为d维向量。对解码器来说,利用线性变换以及Softmax函数将解码的输出转化为一个预测下一个Token的概率。由于模型没有任何循环或者卷积,为使用序列的顺序信息,需要将Tokens的相对以及绝对位置信息注入到模型中。在输入向量的基础上加了一个"位置编码"。由于位置编码和输入向量具有同样的维度d,所以两者可以直接相加。

Transformer以三种不同的方式使用多头注意力机制。

(1)在编码器-解码器的Attention层,序列来自于之前的解码器层,而键和值都来自于编码器的输出。类似于很多Seq2Seq模型所使用的Attention机制。

(2)在编码器含有Self-Attention层。在一个Self-Attention层中,所有的键和值以及序列都来自于同一个地方,即编码器前一层的输出。

(3)类似地,解码器中的Self-Attention层也是一样。不同的是在Scaled点乘Attention操作中加了一个Mask的操作,这个操作是保证Softmax操作之后不会将非法的Values连到Attention层中。

4 实 验

4.1 数据集

本文使用的数据集为CCL2018测评中文隐喻检测任务所使用的评测数据①。共计4 394个中文句子,其组成如表1所示。

由于本文所提出的算法主要针对名词隐喻进行识别,所以将数据集划分为两类:所有名词隐喻句子的集合作为一类;动词隐喻与非隐喻句子的并集作为另一类。名词隐喻句子、动词隐喻句子和非隐喻句子的具体示例如表2所示。

表1 数据集组成

Table 1 Data Set Composition

类别	数量	比例
动词隐喻	2 040	46.43%
名词隐喻	2 035	46.31%
非隐喻	319	7.26%
总计	4 394	100%

表2 数据样例(部分)

Table 2 Sample Data Set

类别	示例
动词隐喻	知了在树上唱歌
名词隐喻	他像孔雀一样高傲
非隐喻	对任何不屈服于美国的国家实行制裁

"知了在树上唱歌"这一句子,"知了"本身并不具备"唱歌"的能力和属性,因此"唱歌"是"知了"的动词隐喻,其本体是知了的鸣叫声。而"他像孔雀一样高傲"则是明显的名词隐喻句子,将本体"他",比喻为喻体"孔雀",用来更加形象地表达本体的特征和属性。

4.2 对比实验

与常用的深度学习文本分类模型进行对比,对比模型包括:

(1)卷积神经网络(Convolutional Neural Networks, CNN)[22]:在文本领域,模型输入为使用Word2Vec预训练的词向量[23],通常进行一维卷积,即卷积核宽度为词向量的长度。TextCNN使用多种尺寸的卷积核,具有较好的局部特征抽取能力。本文采用TextCNN模型对用户问题和标准问题分别进行编码,然后将编码后的向量进行拼接,经全连接层和Sigmoid层完成二分类输出。

(2)长短期记忆网络(Long Short-Term Memory, LSTM)[16]:以预训练的词向量为输入,相比于原始的循环神经网络,能够提取出远距离依赖的语义信息。在本文任务中,LSTM模型被用来参与句子的

①http://ir.dlut.edu.cn/news/detail/496.

编码过程，除编码结构以外的其余结构与CNN模型一致。

（3）长短期记忆网络+注意力机制（LSTM+Attention）[18]：在LSTM模型基础上增加Self-Attention结构，获得LSTM各隐层的权重，经过加权平均得到最终的向量输出，取代原始LSTM模型中仅输出最后一个隐藏层向量的方式。

（4）神经网络（Neural Networks，NN）[19]：仿照原文的网络结构以及参数设置，采用三层全连接神经网络结构，隐藏层神经元维度设置为100，同时使用Dropout方法随机失活部分神经元来防止过拟合现象的产生。并且在原始句子序列基础上，尝试加入词性（POS）子序列和主谓宾（SVO）子序列，对此进行实验验证。

（5）深度双向长短期记忆网络（DBi-LSTM）[24]：由多个LSTM层构建的神经网络结构，能够利用多个LSTM层提取文本特征，缺点是训练速度较慢。本文仿照Li等[24]提出的网络结构，使用连续的双层LSTM，既可以利用多层LSTM的优势，又兼顾了训练效率。

（6）卷积神经网络+支持向量机（CNN+SVM）[25]：原始句子序列通过卷积神经网络进行隐喻特征抽取，将得到的隐喻特征向量作为SVM（Support Vector Machine）的输入。仿照原文的参数设置，并且与NN方法类似，在原始句子序列基础上，尝试加入词性（POS）子序列和主谓宾（SVO）子序列，对此进行了实验验证。

（7）胶囊网络（CapsNet）[26]：Hinton等[26]在2011年首次介绍了胶囊网络模型CapsNet。Capsule是一组神经元，其输入输出向量表示特定实体类型的实例化参数（即特定物体、概念实体等出现的概率与某些属性）。一个胶囊网络基本上是一个试图执行反向图形解析的神经网络，由许多胶囊组成，一个胶囊是一个函数，试图给特定位置的目标预测其存在性以及实例化参数。与CNN类似，胶囊网络模型最初也是被用于计算机视觉领域，并且在图像识别上取得了比CNN更好的效果。仿照原文的网络结构与参数设置，与本文模型进行对比。

4.3　评价指标

本文模型以及对比实验通过4个指标进行分析：准确率（Acc）、精确率（P）、召回率（R）以及F_1值。计算过程分别如公式（5）–公式（8）所示。

$$Acc = \frac{T_p + T_n}{T_p + T_n + F_p + F_n} \tag{5}$$

$$P = \frac{T_p}{T_p + F_p} \tag{6}$$

$$R = \frac{T_p}{T_p + F_n} \tag{7}$$

$$F_1 = \frac{2PR}{P + R} \tag{8}$$

对于给定的预测标签，通过使用混淆矩阵（Confusion Matrix），可以快速计算其精确率和召回率。二分类的混淆矩阵总共包含4个不同的结果：真正例（T_p）、假正例（F_p）、真反例（T_n）以及假反例（F_n）。列表示真实值，行表示预测值。行和列的交叉点就是这4种结果，如表3所示。

表3　混淆矩阵的字符含义

Table 3　Character Meaning in Confusion Matrix

实际＼预测	True	False
True	T_p	F_n
False	F_p	T_n

5　实验结果

经过多次试验，本文模型最终使用Adam优化器[27]，学习率设置为0.01，Epoch数设为20，Batch数设置为16，Dropout值设置为0.3。同时使用谷歌预训练的Word2Vec词向量模型[23]，词向量维度设置为300，去掉停用词后，选择词频最高的10 000个单词，作为特征提取。经过试验验证，如此设置参数可以在本文的实验数据上获得最佳结果。

将数据集按9∶1的比例划分为训练集和测试集。使用十折交叉验证方式，在训练集和验证集上进行参数调整和验证，在测试集上对本文模型进行测试和评价。

使用开源的Python深度学习工具包Keras搭建模型①。同时使用由谷歌提供的预训练的BERT模

①https://github.com/keras-team/keras.

型进行特征提取[①]。

本文所提模型以及对比实验在4个指标上的结果如表4所示。

表4　名词隐喻识别的实验结果

Table 4　Results of Noun Metaphor Identification

模型	*Acc*	*P*	*R*	F_1
CNN	0.870 9	0.879 6	0.834 6	0.856 5
LSTM	0.843 6	0.850 0	0.803 1	0.825 9
NN	0.746 7	0.742 8	0.743 1	0.747 8
LSTM+ATT	0.850 9	0.870 6	0.795 2	0.831 2
DBi-LSTM	0.744 8	0.743 0	0.743 8	0.744 5
CNN+SVM	0.784 0	0.781 2	0.780 2	0.784 6
Capsule	0.878 1	0.875 5	0.858 2	0.866 7
Transformer	0.856 3	0.895 9	0.779 5	0.833 6
BERT	0.883 6	0.874 0	0.874 0	0.874 0
BERT+Transformer	**0.900 0**	**0.896 4**	**0.885 8**	**0.891 0**

可以发现，BERT+Transformer模型在所有的4个指标上均有最好的表现，相比于目前主流的深度学习模型都有不同程度上的提高。说明本文所提模型能够更有效地处理名词隐喻中本体和喻体之间的相似性和差异性，并且能够建立起本体和喻体之间的映射关系。而*P*和*R*之间的差别不大，能够证明BERT+Transformer模型对数据的预测结构性较好，与数据的相关性较好。

BERT+Transformer的准确率最高，达到0.900 0，而仅使用BERT模型的准确率为0.883 6，说明Transformer网络对隐喻识别的有效性。Transformer和BERT+Transformer模型在精确率上表现较为出色，分别达到0.895 9和0.896 4，与对比实验相比具有优势。两者相差不大说明BERT在精确率上的提升没有明显的效果，特别当仅使用BERT时，精确率为0.874 0，进一步证明了这点。本文所提模型在召回率指标上领先于所有的对比模型，达到0.885 8。BERT模型次之，达到0.874 0。然而Transformer模型的表现较差，仅为0.779 5，充分说明了BERT对于识别召回率提升的意义，经过预训练的BERT能够有效提取到名词隐喻中本体和喻体的相关性。BERT+Transformer的F_1值达到0.891 0，Capsule和BERT分别达到0.866 7和0.874 0，均领先于其他对比模型中效果最好的CNN（0.856 5）。Transformer模型的表现较差，仅为0.833 6，与LSTM（0.825 9）和LSTM+ATT（0.831 2）基本持平。实验结果说明，预训练BERT模型对于识别的综合能力提升很大。尤其是BERT能够提取文本深层信息的特点，对于在名词隐喻中找到本体和喻体之间的关联性具有极大的帮助。

从预测结果来看，对于其他模型难以预测的一些句子，BERT+Transformer模型也能够给出正确预测。如“安贫乐道的达观修养，成了中国文化人格结构中一个宽大的地窖，尽管有浓重的霉味，却是安全而宁静。”，此句没有出现例如“像”、“似乎”等明显的名词隐喻特征。但是经过训练后的BERT+Transformer模型，准确地找到了“修养”和“地窖”之间隐含的关联，正确地做出了判断。而其他主流深度学习模型，对于类似隐喻特征不明显的句子，很难得到正确的结论。

但是BERT+Transformer模型也存在一定局限，特别是句子中包含成语或冷僻词汇的情况。例如：“美丽的野鸡拖着绚丽的翎翅在住家庭院倘佯。”中，“翎翅”这一词汇比较冷僻，因此没有被准确识别。再如“当你亲眼看到了猎豹追赶斑马、狮子撕裂羚羊、红鹤与冠鸡相争，鹈鹕与火烈鸟并飞的时候，你就会感叹这里确是动物界相依相克、强者为王的地方。”以及“他的学识纵贯古今，融会中外，真正能读懂陈寅恪的人却不多。”这些存在成语的句子，预测准确率不高。另外，如果句子中存在干扰词汇，对于模型的判断也会造成影响。例如“邵老师把慈父般的关爱倾洒在孩子们身上，让远在他乡的少数民族学生感受到‘家’的温暖。”中，“慈父般的”是一个误导词，实际上这一句子是其字面意的表达。对于这种人类也可能判断出错的句子，模型仍然存在局限，难以超越人类。

针对这些局限，笔者认为有两种解决方法：

（1）扩大训练的数据，增加训练的样本。学习的样本越多，对于模型来说冷僻词就越少，模型的表现也就会越来越好。本文仅仅使用4 000余条数据进行训练和测试就达到目前的效果。

①https://github.com/google-research/bert.

（2）针对中文和隐喻的特点，重新训练一个语言模型。本文所使用的BERT，是在由谷歌开源的中文预训练模型基础上微调得到的，只有中文当中“字”级别的嵌入。如果能够结合中文和隐喻的特点，例如成语古语等，有针对性地训练一个语言模型，对于名词隐喻识别任务效果的提升会有相当大的帮助。不过缺点在于，训练语言模型的开销是巨大的。

6　结　语

隐喻识别是自然语言中语义理解的重要环节，对于机器智能的提升至关重要，如何有效识别语言表达中的隐喻是当前研究亟待攻克的难题。针对目前隐喻识别研究中语义信息利用不足以及名词隐喻识别中的关系表征问题，本文提出基于BERT与Transformer模型的名词隐喻识别方法，使用BERT模型替代词向量，在语义表示中同时包含词与词之间的位置关系等信息，利用Transformer模型进行特征提取并通过神经网络分类器进行识别。BERT+Transformer模型可以注意到多个关键点信息，从而提高分类效果。

该方法能够端到端充分考虑语义表示的名词隐喻识别机制，利用语义资源和上下文神经语言环境提供的语义信息，从不同维度获取语义空间，解决名词隐喻的一词多义问题，从局部和全局的角度识别出名词隐喻。实验结果表明，本文所提方法可以有效地完成名词隐喻识别任务，并且在4项模型评价指标上取得了最先进的结果，超过了现有基于人工特征的分类模型及主流深度学习模型。

参考文献：

[1] Lakoff G, Johnson M. Metaphors We Live by[M]. University of Chicago Press, 2008.

[2] Richards I A. The Philosophy of Rhetoric[M]. New York: Oxford University Press, 1965.

[3] Ausubel D P. The Acquisition and Retention of Knowledge: A Cognitive View[M]. Springer Science & Business Media, 2012.

[4] 田嘉, 苏畅, 陈怡疆. 隐喻计算研究进展[J]. 软件学报, 2015, 26 (1): 40-51. (Tian Jia, Su Chang, Chen Yijiang. Computational Metaphor Processing[J]. Journal of Software, 2015, 26(1):40-51.)

[5] Brunner G, Liu Y, Pascual D, et al. On the Validity of Self-Attention as Explanation in Transformer Models[OL]. arXiv Preprint, arXiv: 1908.04211.

[6] Khandelwal U, Clark K, Jurafsky D, et al. Sample Efficient Text Summarization Using a Single Pre-Trained Transformer[OL]. arXiv Preprint, arXiv: 1905.08836.

[7] Liu J, Cohen S B, Lapata M. Discourse Representation Structure Parsing with Recurrent Neural Networks and the Transformer Model[C]// Proceedings of the 2019 IWCS Shared Task on Semantic Parsing. 2019.

[8] Yang W, Xie Y, Lin A, et al. End-to-End Open-Domain Question Answering with Bertserini[OL]. arXiv Preprint, arXiv:1902.01718.

[9] Alberti C, Lee K, Collins M. A Bert Baseline for the Natural Questions[OL]. arXiv Preprint, arXiv: 1901.08634.

[10] Xu P, Ma X, Nallapati R, et al. Passage Ranking with Weak Supervision[OL]. arXiv Preprint, arXiv: 1905.05910.

[11] Fass D. Met: A Method for Discriminating Metonymy and Metaphor by Computer[J]. Computational Linguistics, 1991, 17 (1): 49-90.

[12] Wilks Y, Dalton A, Allen J, et al. Automatic Metaphor Detection Using Large-Scale Lexical Resources and Conventional Metaphor Extraction[C]// Proceedings of the 1st Workshop on Metaphor in NLP, Atlanta, Georgia, USA. 2013: 36-44.

[13] Jang H, Moon S, Jo Y, et al. Metaphor Detection in Discourse [C]// Proceedings of the 16th Annual Meeting of the Special Interest Group on Discourse and Dialogue, New York, USA. 2015: 384-392.

[14] Rai S, Chakraverty S, Tayal D K, et al. A Study on Impact of Context on Metaphor Detection[J]. The Computer Journal, 2018, 61(11): 1667-1682.

[15] Do Dinh E L, Gurevych I. Token-level Metaphor Detection Using Neural Networks[C]// Proceedings of the 4th Workshop on Metaphor in NLP, San Diego, California, USA. 2016: 28-33.

[16] Sun S, Xie Z. BiLSTM-Based Models for Metaphor Detection [C]// Proceedings of the 2017 National CCF Conference on Natural Language Processing and Chinese Computing (NLPCC2017), Dalian, China. 2017: 431-442.

[17] 汪梦翔,饶琪,顾澄,等 . 汉语名词的隐喻知识表示及获取研究[J]. 中文信息学报, 2017,31(6): 1-9.(Wang Mengxiang, Rao Qi, Gu Cheng, et al. Metaphorical Knowledge Expression and Acquisition for Chinese Nouns[J]. Journal of Chinese Information Processing, 2017, 31(6): 1-9.)

[18] Bizzoni Y, Ghanimifard M. Bigrams and BiLSTMs Two Neural Networks for Sequential Metaphor Detection[C]// Proceedings of the 2018 Workshop on Figurative Language Processing, Louisiana, USA. 2018: 91-101.

[19] Gao G, Choi E, Choi Y, et al. Neural Metaphor Detection in Context[OL]. arXiv Preprint, arXiv:1808.09653.

[20] Devlin J, Chang M W, Lee K, et al. Bert: Pre-training of Deep Bidirectional Transformers for Language Understanding[OL]. arXiv Preprint, arXiv:1810.04805.

[21] Vaswani A, Shazeer N, Parmar N, et al. Attention is All You Need [C]// Proceedings of the 31st International Conference on Neural Information Processing Systems. 2017: 5998-6008.
[22] Moriya S, Shibata C. Transfer Learning Method for Very Deep CNN for Text Classification and Methods for Its Evaluation[C]// Proceedings of the IEEE 42nd Annual Computer Software & Applications Conference. 2018: 153-158.
[23] Mikolov T, Chen K, Corrado G, et al. Efficient Estimation of Word Representations in Vector Space[OL]. arXiv Preprint, arXiv:1301.3781.
[24] Li C, Zhan G, Li Z. News Text Classification Based on Improved Bi-LSTM-CNN[C]// Proceedings of the 9th International Conference on Information Technology in Medicine and Education (ITME). IEEE, 2018: 890-893.
[25] Maldonado S, López J. Dealing with High-Dimensional Class-Imbalanced Datasets: Embedded Feature Selection for SVM Classification[J]. Applied Soft Computing, 2018, 67: 94-105.
[26] Hinton G E, Krizhevsky A, Wang S D. Transforming Auto-encoders[C]// Proceedings of the 21st International Conference on Artificial Neural Networks. Springer, 2011: 44-51.
[27] Kingma D P, Ba J. Adam: A Method for Stochastic Optimization [OL]. arXiv Preprint, arXiv:1412.6980.

作者贡献声明：

张冬瑜，林鸿飞：提出研究思路，设计研究方案；
崔紫鹃，张冬瑜：进行实验；
崔紫娟，李映夏：采集、清洗和分析数据；
张冬瑜，张伟：论文起草；
林鸿飞：论文最终版本修订。

利益冲突声明：

所有作者声明不存在利益冲突关系。

支撑数据

支撑数据由作者自存储，E-mail: zhangdongyu@dlut. edu. cn。
[1] 林鸿飞, 张冬瑜. Noun. xml. 名词隐喻数据.

收稿日期：2019-07-30
收修改稿日期：2019-11-08

Identifying Noun Metaphors with Transformer and BERT

Zhang Dongyu[1] Cui Zijuan[2] Li Yingxia[1] Zhang Wei[1] Lin Hongfei[3]
[1](School of Software, Dalian University of Technology, Dalian 116620, China)
[2](International Office, Dalian University of Technology, Dalian 116024, China)
[3](School of Computer Science and Technology, Dalian University of Technology, Dalian 116023, China)

Abstract: [**Objective**] This paper proposes a new method to address the issues facing semantic information and relationship representation, aiming to improve the recognition of noun metaphors. [**Methods**] First, we used the BERT model to replace the word vector, and added position relationship among words for the semantic representation. Then, we utilized the Transformer model to extract features. Finally, we identified the noun metaphors with the help of used neural network classifier. [**Results**] The proposed model got the highest scores in accuracy (0. 900 0), precision (0. 896 4), recall (0. 885 8), and F1(0. 891 0). It covered multiple key points to improve the classification results of noun metaphors. [**Limitations**] The proposed method could not process the Chinese ancient idioms, as well as rare or dummy vocabularies. [**Conclusions**] The proposed model could more effectively identify Noun Metaphors than the existing models based on artificial features and deep learnings.
Keywords: Metaphor Recognition Noun Metaphor Semantic Comprehension Transformer Model BERT