基于注意力机制的概念化句嵌入研究

王亚珅[1] 黄河燕[1] 冯 冲[1] 周 强[2]

摘 要 大多数句嵌模型仅利用文本字面信息来完成句子向量化表示, 导致这些模型对普遍存在的一词多义现象缺乏甄别能力. 为了增强句子的语义表达能力, 本文使用短文本概念化算法为语料库中的每个句子赋予相关概念, 然后学习概念化句嵌入 (Conceptual sentence embedding, CSE). 因此, 由于引入了概念信息, 这种语义表示比目前广泛使用的句嵌入模型更具表达能力. 此外, 我们通过引入注意力机制进一步扩展概念化句嵌入模型, 使模型能够有区别地选择上下文语境中的相关词语以实现更高效的预测. 本文通过文本分类和信息检索等语言理解任务来验证所提出的概念化句嵌入模型的性能, 实验结果证明本文所提出的模型性能优于其他句嵌入模型.

关键词 句嵌入, 短文本概念化, 注意力机制, 词嵌入, 语义表达

引用格式 王亚珅, 黄河燕, 冯冲, 周强. 基于注意力机制的概念化句嵌入研究. 自动化学报, 2020, **46**(7): 1390−1400

DOI 10.16383/j.aas.2018.c170295

Conceptual Sentence Embeddings Based on Attention Mechanism

WANG Ya-Shen[1] HUANG He-Yan[1] FENG Chong[1] ZHOU Qiang[2]

Abstract Most sentence embedding models typically represent each sentence only using word surface, which makes these models indiscriminative for ubiquitous homonymy and polysemy. In order to enhance representation capability of sentence, we employ short-text conceptualization algorithm to assign associated concepts for each sentence in the text corpus, and then learn conceptual sentence embedding (CSE). Hence, this semantic representation is more expressive than some widely-used text representation models such as latent topic model, especially for short-text. Moreover, we further extend CSE models by utilizing an attention mechanism that select relevant words within the context to make more efficient prediction. In the experiments, we evaluate the CSE models on three tasks, text classification and information retrieval. The experimental results show that the proposed models outperform typical sentence embed-ding models.

Key words Sentence embedding, short-text conceptualization, attention mechanism, word embedding, semantic representation

Citation Wang Ya-Shen, Huang He-Yan, Feng Chong, Zhou Qiang. Conceptual sentence embeddings based on attention mechanism. *Acta Automatica Sinica*, 2020, **46**(7): 1390−1400

很多自然语言处理任务都依赖于文本的定长向量表示. 其中, 句子定长向量表示 (又称 “句嵌入”) 是很重要的. 可能最常见的文本定长向量表示方法是词袋模型或者 N-Gram 词袋模型[1]. 但是这类模型面临严峻的数据稀疏性和高维度挑战, 并且无法对词语语义进行建模, 也损失了词语的距离和顺序信息. 近来, 很多研究尝试使用深度神经网络 (Deep neural network, DNN) 来学习句子向量表示, 这类基于深度学习的方法达到了目前句嵌入研究的最好结果[2−4]. 尽管这类基于深度学习的研究取得不错的研究进展, 但是目前的句嵌入模型面临如下挑战: 1) 大多数句嵌入模型只使用文本字面信息来表示句子, 导致这些模型对于普遍存在的 “一词多义” 现象缺乏甄别能力; 2) 有研究尝试将句法结构或者主题建模引入句嵌入, 但是对于短文本, 由于缺乏足够的信号用于统计和推导, 所以无论是句法分析还是主题建模都无法在短文本上取得良好效果; 3) 大多数句嵌入模型平等地处理句子中每个词语, 这种 “一视同仁” 的建模理念不符合人类阅读习惯和人类注意力机制; 4) 设定上下文语境窗口大小比较困难. 为了解决上述问题, 我们必须另辟蹊径、从有限的输入句子中捕获更多语义信号, 例如: 概念 (Concept). 此外, 我们需要为不同的上下文词语赋予不同的注意力, 来增强真正对每次预测有帮助的词语的重要程度.

本文提出概念化句嵌入 (Conceptual sentence

收稿日期 2017-06-02 录用日期 2018-03-24
Manuscript received June 2, 2017; accepted March 24, 2018
国家自然科学基金重点项目 (61751201) 资助
Supported by National Natural Science Foundation of China (61751201)
本文责任编委 赵铁军
Recommended by Associate Editor ZHAO Tie-Jun
1. 北京理工大学计算机学院北京市海量语言信息处理与云计算应用工程技术研究中心 北京 100081 2. 百度在线网络技术 (北京) 有限公司 北京 100085
1. Beijing Engineering Research Center of High Volume Language Information Processing and Cloud Computing Applications, School of Computer, Beijing Institute of Technology, Beijing 100081 2. Baidu Inc., Beijing 100085

embedding, CSE) 模型. 这是一个用于学习句子向量化表示的无监督框架: 在创新性地引入概念信息基础之上, 学习得到的概念层面的句向量被用于预测上下文片段中的目标词或者语境词. 本文是受近来基于深度学习的词嵌入研究的启发[3, 5]. 我们首先使用短文本概念化算法[6] 获取句子的概念分布, 进而生成相应的概念向量; 随后, 句向量、语境词词向量以及该句的概念向量, 被平均来预测给定上下文片段中的目标词. 所有句向量和词向量都是通过随机梯度方法和反向传播技术来训练得到的.

注意力机制能够实现不同数据形态之间的自动对齐, 能够有倾向性地重点关注某些对解决问题起最关键作用的数据元素, 在许多自然语言处理任务中获得较大认可[7−9]. 不难发现, 在我们的概念化句嵌入模型中, 在给定上下文片段中, 目标词的预测仅与窗口中的某些词有关, 而并非与窗口中所有词语有关. 这与人类阅读习惯是一致的, 人类注意力机制会自动增强某些词语而相对忽略另一些词语, 相关研究表明词语类型 (Word type) 和惊异度 (Surprisal) 与人类阅读行为有直接关系[10−11]. 根据人类阅读行为, 本文使用注意力机制扩展概念化句嵌入模型, 使模型能够在预测目标词的时候, 根据词语类型和惊异度来有区别地对待上下文语境词. 所以, 本文提出基于词语类型的注意力机制和基于惊异度的注意力机制等两种注意力机制.

总的来说, 概念化句嵌入的核心思想是: 在引入概念信息和注意力机制后, 概念层面的句嵌入模型允许每个词语在不同概念下拥有不同的意义、拥有不同的嵌入形式. 例如, 对于词语 “apple”, 在概念 *Food* 下可能指一个水果, 而在概念 *Information company* 下可能指一家 IT 公司. 所以, 概念信息会有效提升句向量的语义甄别能力和表达能力. 此外, 本文所提出模型的一个重要优势在于, 可以在无标注数据上完成自动训练, 且相较于已有模型标注成本和训练成本大幅降低; 另一大优势是该模型可以利用词语顺序信息, 这是目前很多句嵌入模型所欠缺的, 其效果类似于 N-Gram 词袋模型但是大大降低向量维度和存储开销. 在文本分类任务和信息检索任务上的实验结果, 充分证明了这种概念层面句子向量化表示模型的性能.

本文内容安排如下: 第 1 节总结国内外相关研究工作; 第 2 节为本文所涉及相关概念和研究任务进行形式化定义; 第 3 节详细介绍基于注意力机制的概念化句嵌入模型的相关研究细节; 第 4 节分析讨论实验及实验结果; 最后, 第 5 节总结全文.

1 相关工作

本文所研究的基于注意力机制的概念化句嵌入模型与句嵌入模型、短文本概念化算法以及注意力机制等研究密切相关.

1.1 句嵌入模型

句嵌入研究旨在将句子转化成可计算的定长向量, 在很多自然语言处理任务中发挥着重要作用, 例如文本分类[12]、文本相似度计算[13]、自动问答[14]、情感分析[15] 等. 在句子表示领域, One-Hot 方法曾被广泛应用, 例如词袋模型 (Bag-of-word, BOW) 等[1]. 近来, 基于深度学习的方法在句子表示中取得最优效果[2−4, 13, 16], 相关工作可以分为三类: 1) 通过组合词向量生成句向量; 2) 受词嵌入模型[5] 启发的句嵌入模型; 3) 基于通用深度神经网络架构生成句向量. 在第一类工作中, 基于词向量加权求和 (或平均) 的方式获得句向量, 是最直观、最简单的研究思路, 但是丢失了词序及词语之间的关联关系信息, 例如文献 [4]、文献 [17] 等. 第二类工作的代表工作包括文献 [3−4]、文献 [13] 等, 文献 [3] 提出段落向量 (Paragraph vector, PV) 模型将不定长文档表示成向量, 该向量被训练用于预测文档中的词语, 文献 [13] 提出一个神经网络架构来获取高质量词嵌入, 并直接用于优化句子表示. 但是, 他们的方法仅依赖文本字面信息, 而忽略了诸如文本概念、文本主题等更高层次语义信息. 第三类方法则基于卷积神经网络[18]、循环神经网络[2, 19] 等, 将词向量序列有效地编码成短语向量或者句向量, 此类研究大多在监督学习框架下展开, 往往针对特定应用场景或任务来训练和优化模型, 导致其语料依赖性高、领域移植能力欠缺、模型灵活性较差.

此外, 很多学者为探讨深度学习和语言结构的关系, 提出很多基于句法分析的句嵌入模型来利用长距离依存关系[20−21]. 但是句子作为短文本, 通常难见完整规范的书面语言句法结构, 而且缺乏足够的信号用于统计推理. 综上, 由于输入文本中缺乏足够信息号, 无论句法分析还是主题建模都很难奏效, 因此需要从有限的输入中捕获更多语义信号, 例如概念信息. 综上, 本文通过引入概念信息和注意力机制扩展了 PV 模型.

1.2 短文本概念化算法

短文本概念化 (Short-text conceptualization) 研究是近来新兴的热门研究方向, 旨在挖掘与给定短文本 (中的词语) 最相关的 “概念”[22−25].

基于贝叶斯推理机制, 文献 [23] 使用条件概率对来自概率化词汇知识库 Probase[26] 中的概念进行排序, 选择能够使上述条件概率取得最大值的概念集合来表示给定短文本. 为了从充满噪声且稀疏性明显的短文本中挖掘更多信号, 文献 [6] 尝试引入词语的动词修饰信息、形容词修饰信息和词语的属性

信息等, 在一定程度上为理解词语提供了有益线索; 随后, 基于随机游走的方法, 在每轮迭代中都给候选概念重新打分, 最终在算法收敛的时候获得相关概念及概念分布.

短文本概念化通过引入概念这一更高层面的语义信息来扩展原始短文本, 进而帮助有效地理解短文本意义和内涵[27]. 因此, 本文使用短文本概念化算法来挖掘概念级别的句子内涵, 将文献的工作[6]引入到句嵌入模型中.

1.3 注意力机制

注意力机制 (Attention mechanism) 源于认知心理学中的人脑注意力机制[28], 能够有选择性地重点关注源数据的某些部分, 近来已被应用于提升多种自然语言处理任务性能[7−8, 16−17]. 本文中, 注意力机制将被用于为上下文片段中不同的语境词语赋予不同的注意力值. 本文中注意力机制的设计源于人类阅读行为研究[29−30]. 相关研究已经证明词性、组合范畴文法、词语长度、词频等都是影响人类阅读行为的因素[31−32], 而近来很多研究已在尝试通过模拟人类阅读行为来提高自然语言处理任务效率[17, 30−32].

2 问题描述

本文首先对所研究的任务进行形式化定义:

定义 1. 概念: 参照以往研究[24], 我们将 "概念 (Concept)" 定义为一组 (类) 实体或事物的集合, 属于相近 (或相同) 类别的词语有相似 (或相同) 的概念表达. 例如, 词语 "Jeep" 和词语 "Honda" 都属于概念 *Car*. 本文所使用的词汇语义知识库是 Probase[26], Probase 同样被作为本文的词表. 相关研究已经证明[6], 在理解短文本 (例如理解搜索引擎中用户所提出的查询或者理解问答系统中用户所提出的问题等) 方面, 我们需要的是关于语言和语用的知识, 或者说是词语在一种语言之中是如何彼此交互的[22], 因此在这种情况下, 词汇知识库的应用价值和必要性要大于百科知识库 (例如 Wikipedia、DBpedia、Freebase 和 Yago 等). 因此, 本文选用目前规模最大、质量最优的词汇知识库 Probase, 该知识库已被成功应用于针对多种类型短文本 (如新闻类文本[23]、社交媒体文本 (如 Twitter 数据等)[6, 23, 25]、搜索引擎查询项[6] 等) 的概念化和理解任务, 并体现出良好的性能.

定义 2. 短文本概念化: 给定句子 $S = \{w_1, w_2, \cdots, w_l\}$, 其中, w_i 表示词语. 通过短文本概念化算法[6], 我们可以: 1) 从知识库中获得概念分布 (Concept distribution) $\theta_C = \{c_i, p_i | i = 1, \cdots, k_C\}$ 来表示句子 S. 其中, p_i 表示概念 c_i 的概率. 2) 同时获得句子 S 的关键词集合 $W = \{\langle w_j, RS(w_j)\rangle | j = 1, \cdots, k_W\}$. 其中, $RS(w_j)$ 表示词语 w_j 的打分, 表征词语 w_j 对句子 S 整体语义建模的重要程度. 文献 [6] 设计实现了一个融合词语搭配知识和基于随机游走框架的短文本概念化模型, 取得目前短文本概念化研究的最佳实验效果. 因此, 本文使用该算法来产生概念分布 θ_C 和关键词集合 W.

定义 3. 概念化句嵌入: 给定句子 $S = \{w_1, w_2, \cdots, w_l\}$, 及其概念分布 θ_C, 概念化句嵌入旨在生成 d 维向量 $\boldsymbol{s}$ 来对句子 S 的语义进行表示.

3 概念化句嵌入

本文提出 5 种概念化句嵌入模型. 其中, 第一个概念化句嵌入模型是基于词嵌入研究中的 CBOW 模型, 记为 CSE-CBOW; 第二个概念句嵌入模型是基于词嵌入研究中的 Skip-Gram 模型, 记为 CSE-SkipGram. 基于上述基本概念化句嵌入模型, 我们通过引入不同类型的注意力机制来得到相应变体: 1) 引入基于词语类型的注意力机制, 提出 aCSE-TYPE 模型; 2) 引入基于惊异度的注意力机制, 提出 aCSE-SUR 模型; 3) 引入基于词语类型和惊异度的注意力机制, 提出 aCSE-ALL 模型.

3.1 CBOW 模型和 Skip-Gram 模型

本文所提出的概念化句嵌入模型是受词嵌入模型启发. CBOW 模型和 Skip-Gram 模型是比较经典的学习词向量的框架[5, 33], 本文首先简单回顾这两个词嵌入模型.

CBOW 模型的框架如图 1 (a) 所示. 每个词语被映射成唯一向量, 被表示成词矩阵 $W \in \mathbf{R}^{d \times |V|}$ 的一列. 其中, V 表示词表 (词表规模记为 $|V|$), d 表示词向量维度. 这些词语向量的平均通常被称为语境向量, 作为特征来预测当前上下文片段的目标词 w_t.

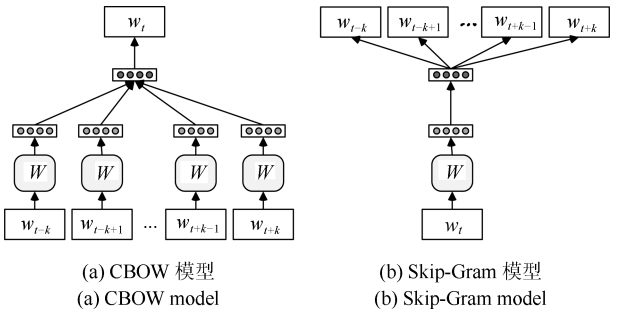


图 1 CBOW 模型和 Skip-Gram 模型

Fig. 1 CBOW model and Skip-Gram model

通常, 给定句子 $S = \{w_1, w_2, \cdots, w_l\}$, CBOW

模型在句子 S 上最大化如下目标函数:

$$\ell(S)=\frac{1}{l-2k}\sum_{t=k+1}^{l-k}\lg P(w_t|Context(w_t)) \quad (1)$$

其中, k 控制目标词 w_t 的上下文窗口规模, $Context(w_t)$ 表示词语 w_t 的上下文词语集合, 定义为 $Context(w_t)=\{w_{t+c}|-k\le c\le k, c\neq 0\}$. 式 (1) 中概率 $P(w_t|Context(w_t))$ 的求解方式, 如下所述.

为了便于描述, 我们定义词语 w_t 的语境向量 $\boldsymbol{context}_t$, 由词矩阵 W 中的词向量 $\{\boldsymbol{w}_{t-k},\cdots,\boldsymbol{w}_{t+k}\}$ 通过平均构成:

$$\boldsymbol{context}_t=\frac{1}{2k}\sum_{-k\le c\le k, c\neq 0}\boldsymbol{w}_{t+c} \quad (2)$$

其中, $\boldsymbol{w}_{t+c}$ 表示词语 w_{t+c} 在词矩阵 W 中对应的词向量, 且 $\boldsymbol{context}_t\in\mathbf{R}^{m\times 1}$. 进而计算

$$\boldsymbol{y}_{w_t}=U\cdot\boldsymbol{context}_t+\boldsymbol{b} \quad (3)$$

其中, $U\in\mathbf{R}^{|V|\times d}$ 是权值矩阵, 是 $\boldsymbol{b}\in\mathbf{R}^{|V|\times 1}$ 偏置向量, 均是需要训练得到的参数. 经过上述步骤计算得到的 $\boldsymbol{y}_{w_t}=(y_{w_1},y_{w_2},\cdots,y_{w_t},\cdots,y_{w_{|V|}})^{\mathrm{T}}$ 是一个长度为 $|V|$ 的向量, 但其分量 y_{w_t} 不能表示概率. 如果想要使 $\boldsymbol{y}_{w_t}$ 的分量 y_{w_t} 表示上下文为 $Context(w_t)$ 时目标词语恰好为词表 V 中第个 t 词语 w_t 的概率, 则还需要执行一个 Softmax 归一化操作. 经过归一化后, $P(w_t|Context(w_t))$ 表示为如下形式:

$$P(w_t|Context(w_t))=\frac{\mathrm{e}^{y_{w_t}}}{\sum\limits_{i=1}^{|V|}\mathrm{e}^{y_{w_i}}} \quad (4)$$

其中, 分母中的 i 表示词语 w_i 在词表 V 中的索引. Skip-Gram 模型的框架如图 1 (b) 所示. 该模型旨在给定目标词 w_t 来预测上下文片段中的语境词, 而不是像 CBOW 模型那样基于语境词来预测目标词. 通常, Skip-Gram 模型的目标函数是最大化如下平均对数概率:

$$\ell(S)=\frac{1}{l-2k}\sum_{t=k+1}^{l-k}\sum_{-k\le c\le k, c\neq 0}\lg P(w_{t+c}|w_t) \quad (5)$$

通常使用 Softmax 函数计算上式中的条件概率 $P(w_c|w_t)$, 如下:

$$P(w_c|w_t)\frac{\mathrm{e}^{\boldsymbol{w}_c{}^{\mathrm{T}}\boldsymbol{w}_t}}{\sum\limits_{w_i\in V}\mathrm{e}^{\boldsymbol{w}_i{}^{\mathrm{T}}\boldsymbol{w}_t}} \quad (6)$$

其中, $\boldsymbol{w}_t$ 和 $\boldsymbol{w}_c$ 分别是目标词 w_t 和语境词 w_c 的词向量. 通常, 在训练 CBOW 模型和 Skip-Gram 模型时: 1) 层次化 Softmax 和负采样技术被用于提高学习效率[5]; 2) 使用随机梯度方法来训练词向量, 梯度由反向传播产生[33]. 训练收敛之后, 语义相似的词语被映射到语义向量空间的相近位置, 例如, 词语 “powerful” 和词语 “strong” 在语义向量空间中的位置比较接近.

直观地, 本文所提出的概念化句嵌入模型是受词嵌入研究的启发. 在词嵌入研究中: 1) 词向量被用于预测目标词或者上下文语境词; 2) 虽然词向量被随机初始化, 但是作为训练过程的一个间接结果, 它们最终会赋予准确的语义含义. 因此, 我们将这个理念移植到概念化句嵌入模型中: 在给定上下文语境中, 被赋予概念信息的句向量, 被用于预测目标词或者上下文语境词. 此外, 注意力机制会为上下文片段中不同语境词赋予不同注意力.

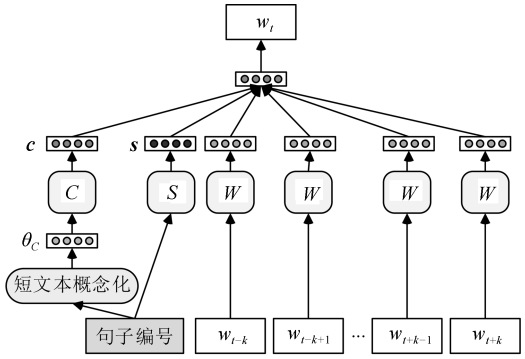
3.2 基于 CBOW 模型的概念化句嵌入

首先介绍本文提出的第一个概念化句嵌入模型 (CSE-CBOW), 该句嵌入模型是基于词嵌入研究中的 CBOW 模型. CSE-CBOW 模型的框架如图 2 (a) 所示, 每个句子由唯一的句子编号 (句 ID) 标识, 通过句矩阵 S 被映射成唯一向量 $\boldsymbol{s}$. 概念分布 θ_C 是由一个基于词汇知识库的短文本概念化算法所生成[6]. 类似于词嵌入方法, 每个词语 w_i 被映射成为唯一向量 $\boldsymbol{w}_i$, 表示为词矩阵 W 的一列. 上下文窗口中的词语 $Context(w_t)=\{w_{t+c}|-k\le c\le k, c\neq 0\}$, 句 ID 以及这个句子对应的概念分布 θ_C, 作为输入. 图 2 (a) 中, C 是一个固定的线性变换器, 将概念分布 θ_C 转换成为概念向量 $\boldsymbol{c}$. 这也使我们的模型与 Mikolov 的模型[3] 产生很大区别: 他们的工作仅使用字面信息, 而没有利用概念信息. 本文实验结果充分证明了引入概念信息的有效性.

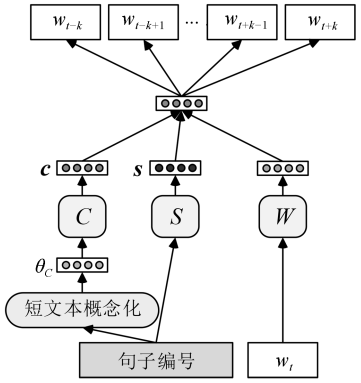
随后, 句向量 $\boldsymbol{s}$、上下文词向量 $\{\boldsymbol{w}_{t-k},\cdots,\boldsymbol{w}_{t+k}\}$ 和概念向量 $\boldsymbol{c}$ 被平均来预测上下文窗口的目标词 w_t. 需要注意的是, 句向量在该句所有上下文片段上共享, 而不会跨句共享. 所以, 可以认为句 ID 扮演着一个存储器的角色, 用于记录在当前上下文片段所缺失的句子全局信息. 其中, 上下文片段是从当前句子上通过滑动窗口采样得到的, 窗口长度都是固定的 (长度为 $2k$). 词矩阵 W 则在不同句子间共享. 总结来说, CSE-CBOW 的算法流程如下所述.

概念化阶段: 对于给定的句子, 我们首先进行预处理并将该句分割成词语集合; 然后使用基于 Probase 知识库的短文本概念化算法[6], 对该句完成概念化, 得到相应概念及每个概念的概率, 构成前文

提到的概念分布 θ_C. 需要注意的是, 概念分布 θ_C 在整个概念化句嵌入框架中扮演着重要角色, 对整个句子的语义表达有重要作用.



(a) CSE-CBOW 模型
(a) CSE-CBOW model



(b) CSE-SkipGram 模型
(b) CSE-SkipGram model

图 2 CSE-CBOW 模型和 CSE-SkipGram 模型
Fig. 2 CSE-CBOW model and CSE-SkipGram model

训练阶段: 在训练阶段, 我们的目的是在可观察到的句子上, 训练得到词矩阵 W、句矩阵 S 以及 Softmax 权重 $\{U, \boldsymbol{b}\}$. 同时, 本文使用层次化 Softmax 技术来提高模型学习效率. W 和 S 由随机梯度方法训练得到: 在随机梯度方法的每一步, 我们从整个句子中采样的一个定长的上下文片段, 计算通过反向传播得到的误差梯度, 然后使用该梯度来更新参数.

3.3 基于 Skip-Gram 模型的概念化句嵌入

对于上述预测过程, 还存在另外一种建模方式: 在输入中忽略上下文片段中的语境词, 而让模型在输出层预测从上下文片段中随机采样得到的词语. 这种建模方式类似于词嵌入研究中的 Skip-Gram 模型[5]. 如图 2 (b) 所示, 只有句向量 $\boldsymbol{s}$、概念向量 $\boldsymbol{c}$ 和中心词 w_t 的词向量 $\boldsymbol{w}_t$ 被用于预测上下文片段中的词语; 而上下文词语不在用于输入, 而变成输出层所要预测的内容. 在随机梯度下降的每一轮迭代中, 我们采样一个上下文片段 $Content(w_t)$, 然后从该片段中随机采样一个词语, 对当前句向量 $\boldsymbol{s}$、概念向量 $\boldsymbol{c}$ 及 $\boldsymbol{w}_t$ 形成一个分类任务. 上述概念化句嵌入模型记作 CSE-SkipGram. CSE-SkipGram 的训练过程与 CSE-CBOW 类似. 为了区分下文介绍的基于注意力机制的概念化句嵌入模型, 我们将 CSE-CBOW 和 CSE-SkipGram 称为 "基础概念化句嵌入模型".

由上述对 CSE-CBOW 模型和 CSE-SkipGram 模型的介绍可知, 概念分布 θ_C 在本文提出的概念化句嵌入模型中扮演着重要角色. 这是因为, 这个分布中的每一维都对应着一个概念, 而每一维的数值都表示这个句子对应这个概念的概率. 换言之, 概念分布是一个对该句而言比较可靠的语义表达. 相反, 句向量和词向量中的每个维度上的信息则没有任何实质意义、无法被有效解释.

3.4 基于注意力机制的概念化句嵌入

大部分已有句嵌入研究往往平等地处理句子中的每个词语[3−4, 13], 但实际上人类在阅读过程中, 会自动跳过某些词语或者快速扫视某些词语, 而把主要注意力集中在个别重要词语上[10−11, 29]. 人类这种注意力机制不仅有助于提高阅读效率, 而且能够节省有限的认知计算资源. 同理, 人类阅读习惯中的这种注意力机制有助于句嵌入建模.

此外, 正如前文所提到的, 设定上下文片段窗口规模 $2k$ 的值是一件困难的事情: 1) 如果 k 的值设置过大, 不仅会加大计算开销, 而且会导致大量无关词语被无效引入进而导致模型性能衰退; 2) 如果 k 的值设置过小, 会导致上下文片段范围过小而不足以容纳语义相关的词语[8]. 为了解决这个问题, 我们通过引入多种注意力机制, 来扩展上述基础概念化句嵌入模型, 使模型能够有区别地处理上下文片段中的语境词语. 为了方便表述, 我们使用注意力机制扩展 CSE-CBOW 模型. 重写式 (2), 如下:

$$\boldsymbol{context}_t = \frac{1}{2k} \sum_{-k \le c \le k, c \ne 0} a(w_{t+c}) \cdot RS(w_{t+c}) \cdot \boldsymbol{w}_{t+c} \tag{7}$$

由式 (7) 可知, 我们使用对词向量进行加权平均的方式取代式 (2) 中对词向量平均的方式. 这就意味着, 每个上下文语境词 w_{t+c} 被赋予不同 "注意力权重", 表征其对预测目标词 w_t 的重要程度. 式 (7) 中的 "注意力权重" 包括两部分: 1) 语境词 w_{t+c} 的概念化打分 $RS(w_{t+c})$, 表征词 w_{t+c} 对句子整体表达的贡献程度[6] (详见第 2 节); 2) 语境词 w_{t+c} 的注意力因子 $a(w_{t+c})$, 本章节将重点讨论两种注意力因子建模方法, 分别是基于词语类型的注意力机制 (第 3.4.1 节) 和基于惊异度的注意力机制 (第 3.4.2 节).

3.4.1 基于词语类型的注意力机制

人类阅读行为受词语类型 (如词性等) 影响很大, 研究表明人类注意力力更倾向于将停留在开放性词类 (实词为主, 例如名词、形容词、动词等), 而非封闭性词类 (功能词和结构词为主, 例如连词、介词、感叹词等) 则被投入很少注意力, 甚至被忽略[10−11]. 例如, 给定句子 "microsoft/NNP unveils/VBZ office/NN for/IN apple/NN ipad/NNP", 如果要对该句进行有效的句子表示, 需要对具有名词词性 (NN 和 NNP)、动词词性 (VBZ) 的词语给予更高注意力, 而对于介词词性 (IN) 词语施以较低注意力. 引入词语类型信息, 能够实现对人类阅读行为和人类注意力机理的有效模拟, 进而提高句嵌入水平. 本文所重点关注的词语类型信息主要是词性 (Part-of-speech, POS).

综上, 在给定上下文片段 $\{w_{t-k}, \cdots, w_{t+k}\}$, 要预测目标词 w_t 时, $\{-k, \cdots, k\}$ 表示语境词 $\{w_{t-k}, \cdots, w_{t+k}\}$ 对于目标词 w_t 的相对位置 (距目标词 w_t 左/右的距离), 为了便于表述, 下文用 "相对位置" 表示词语位置. 对于上下文片段中位于相对位置 i 的语境词 w_i, 基于词语类型的注意力因子 $a_{\mathrm{TYPE}}(w_i)$, 可以表示为所有语境词上的 Softmax 函数[7], 如下所示:

$$a_{\mathrm{TYPE}}(w_i) = \frac{\mathrm{e}^{d(w_i)} + ex_i}{\sum\limits_{-k \le c \le k, c \neq 0} \mathrm{e}^{d(w_c)} + ex_c} \tag{8}$$

其中, $d(w_i)$ 是词语类型矩阵 $D \in \mathbf{R}^{|V| \times 2k \times |POS|}$ 的一个元素, 表征位于相对位置 i 的词语 w_i 的词语类型 (即词性) 的重要程度. 本文所采用的词性标注集合是宾州英文树库 (Penn treebank) 词性标注集合, $|POS|$ 表示该词性标注集合的规模. ex_i 是偏置矩阵 $E \in \mathbf{R}^{2k}$ 的一个元素, 由相对位置 i 决定. 虽然以往研究表明注意力机制在检索大规模参数表的时间开销比较高昂[7], 但是本文的注意力机制的计算开销比较小. 在计算给定上下文片段中的所有语境词的注意力值的时候, 我们的只需执行 $4k$ 步查表操作: 1) 从词语类型矩阵 D 中, 为每个语境词检索相应位置和相应词性的词语类型值; 2) 从偏置矩阵 E 中, 为每个语境词检索偏置值. 虽然上述注意力值计算策略不是最优计算方式, 而且已有研究提供了多种其他更加复杂的注意力机制建模方式[7, 34], 但是本文所提出的注意力机制是对模型准确率和计算复杂度有效平衡.

因此, 除了基础概念化句嵌入模型中的参数集合 $\{U, \boldsymbol{b}, S, W\}$ 外, 词语类型矩阵 D 和偏置矩阵 E 是新增需要训练的参数集合. 所有参数通过反向传播算法计算, 并在每次迭代之后通过固定学习率更新. 本文将这种引入基于词语类型的注意力机制的概念化句嵌入模型记作 aCSE-TYPE.

以句子 "microsoft unveils office for apple′s ipad" 为例. aCSE-TYPE 模型预测多义词 "apple" 时的模型架构如图 3 所示, 方框的颜色越深表示注意力权重 ($a(w_{t+c}) \cdot RS(w_{t+c})$) 越高. 我们可以观察到, 介词 "for" 被注意的程度较低, 而名词词性的语境词 (特别是对概念化贡献高的词, 即 $RS(w_i)$ 值高的语境词) 往往被赋予更高的注意力权重, 这些词也最具甄别能力, 例如 "ipad" "office" 和 "microsoft" 等. 没有任何歧义的词语 "ipad" 被赋予最高注意力值, 一方面是因为其靠近被预测的目标词 "apple" 并且是专有名词词性 (NNP), 另一方面是因为其与被预测词在大规模语料中共现频繁, 因此语义范畴非常相近, 提升其 $RS(w_i)$ 值.

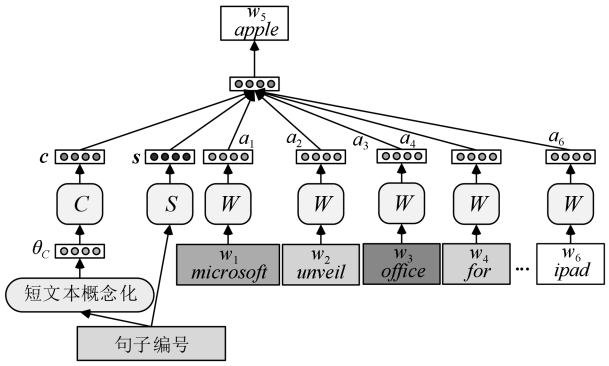


图 3　aCSE-TYPE 模型

Fig. 3　aCSE-TYPE model

3.4.2 基于惊异度的注意力机制

惊异度 (Surprisal) 概念起源于心理学和信息论研究领域[35], 用于预测某个词语在上下文语境中的可预测程度和被处理的困难程度[36], 在众多研究中被作为人类阅读行为的重要预测因素[17, 29]. 心理学和认知学研究认为, 具有较高惊异度值的词语, 包含和传递越丰富的信息, 对其的处理时间和阅读时间会越高, 应该被赋予越高注意力. 因此, 引入惊异度, 能够实现对人类阅读行为和人类注意力机理的有效模拟, 进而提高句嵌入水平. 惊异度通用概念和定义最早由文献 [36] 提出, 通常被定义为给定前序词语序列基础上对该词语的条件概率的负对数形式, 因此一般使用语言模型 (N-Gram 模型) 来计算, 如下所示:

$$s(w_i) = -\lg P(w_i | w_1, w_2, \cdots, w_{i-1}) \tag{9}$$

其中, w_i 表示当前词, $\{w_1, \cdots, w_{i-1}\}$ 表示上下文片段中当前词 w_i 前面的词语. 由式 (9) 可知, 词语 w_i 的惊异度 $s(w_i)$ 表示为, 在给定语境 $\{w_1, \cdots, w_{i-1}\}$

下词语 w_i 的条件概率的负对数函数. 由建模方式可知, 惊异度可以在某种程度上对词语顺序进行建模. 本文所提出的基础概念化句嵌入模型 (CSE-CBOW 和 CSE-SkipGram) 并没有考虑词语之间的顺序, 而词语之间的前后关系 (顺序) 对句子语义表达以及词义消岐都有重要作用. 例如, "Lee Sedol defeats AlphaGo" 和 "AlphaGo defeats Lee Sedol" 这两句话在语义上是不同的. 惊异度能够起到使句嵌入建模过程捕获局部组合信息[17] 的作用, 使得相邻词语能够在语义和句法层面组合起来以增强表情达意能力.

在实际应用中, 惊异度可以通过各类语言模型来进行估计, 例如最简单的统计语言模型以及神经语言模型等. 本文假设, 高惊异度词语涵盖更多信息, 而应该被给予更高注意力程度, 所以我们直接使用惊异度作为注意力权重. 在这种情况下, 基于惊异度的注意力因子可以表示为:

$$a_{\mathrm{SUR}}(w_i) = \frac{\mathrm{e}^{s(w_i)} + ex_i}{\sum\limits_{-k\le c\le k, c\neq 0} \mathrm{e}^{s(w_c)} + ex_c} \tag{10}$$

其中, ex_i 表示相对位置 i 的偏置值, 如上所述. 除了训练 $\{U, \boldsymbol{b}, S, W\}$ 之外, 还需训练偏置矩阵 E, 训练方式同 aCSE-TYPE. 本文将这种引入基于惊异度的注意力机制的概念化句嵌入模型记作 aCSE-SUR.

4 实验

为了验证本文所提出的概念化句嵌入模型的性能, 我们使用文本分类任务和信息检索任务来完成对不同模型的对比. 这些文本理解任务经常被用于评估句嵌入方法[3−4, 9, 25, 27].

4.1 数据集

我们使用了 4 个数据集来进行训练和评估. 其中, 数据集 NewsTitle、TREC 和 Twitter 用于文本分类实验, 数据集 Tweet11 用于信息检索实验. 此外, 我们构建数据集 Wiki 用于训练主题模型进行对比实验.

NewsTitle: 我们从《路透社 (Reuters)》和《纽约时报 (New York Time)》抽取 359 万篇新闻报道, 根据内容分为商业、健康、娱乐、饮食、政治和体育等 6 个类别. 从每个类别随机选取 30 000 篇新闻报道, 仅保留其标题和首句, 构成数据集[23]. 该数据集中句子的平均长度为 9.41 字.

TREC: 该数据集是 TREC 问题分类评测任务的官数据集, 被广泛用于作为文本分类任务. 该数据集包含 5 952 个句子, 被分成人物、缩写、实体、描述、处所和数字等 6 个类别.

Twitter11: 该数据集是 TREC 在 2011 年和 2012 年举办的微博检索评测任务的官方数据集[37]. 使用官方 API, 我们爬取的数据集包括 1 600 万条推特文本 (简称 "推文"). 2011 年查询集 TMB2011 包括 49 个带时间戳查询项, 2012 年查询集 TMB2012 包括 60 个带时间戳查询项.

Twitter: 该数据集是通过手工标注上述数据集 Twitter11 得到. 我们共标注 126 549 条推文, 按照内容分为商业、体育、娱乐和电子产品等 4 个类别. 该数据集中句子的平均长度为 11.03 字. 因为数据集噪音和稀疏性明显, 所以这个社交文本数据集对于所评估模型更加具有挑战性.

此外, 本文还构建了一个维基百科数据集 Wiki. 我们按照如下规则处理维基百科文章: 首先, 去掉少于 100 个词语或者少于 10 个链接的文章; 然后, 去掉所有目录页和消岐页; 最终, 通过爬取内容重定向页面, 得到 287 万篇维基百科文章, 构成数据集.

4.2 对比算法

我们将概念化句嵌入模型与以下算法进行对比:

BOW: 作为一个基础对比算法, 使用词袋模型对句子进行向量化表示[1], 所使用的特征是 TF-IDF 值.

LDA: 使用 LDA 模型[38] 生成主题分布表示句子. 本文使用两种方式训练 LDA 模型: 1) 使用数据集 Wiki 和上述所有评估数据集进行训练; 2) 仅使用评估数据集进行训练. 我们展示二者中性能优者.

PV: Paragraph vector (PV) 模型是最近提出的对不定长文本进行定长向量表示的方法[3], 包括分布式记忆模型 (PV-DM) 和分布式词袋模型 (PV-CBOW) 两类. PV 模型在情感分类等任务中取得目前最优表现, 但是该模型利用文本字面信息.

TWE: 该模型基于 Skip-Gram 模型, 通过借助主题建模的优势, 在一定程度上解决了单纯利用文本字面信息进行句嵌入的歧义性, 实现了主题化词嵌入和主题化句嵌入[4]. 该项工作总共提出 3 种主题化句嵌入模型, 我们展示三者中性能优者. 此外, 对于主题模型的训练方式, TWE 同 LDA.

SCBOW: 受词嵌入研究[3] 启发, 文献 [13] 通过将对 "词向量累加平均得到的句向量" 的对比加入到损失函数中, 直接面向 "词向量累加平均得到句向量" 这个最终目标来优化传统词向量的训练, 提出面向句嵌入的词向量训练模型 Siamese CBOW 模型, 训练得到的词向量被累加取平均得到句向量.

CSE-CBOW: 本文所提出的基于 CBOW 模型的概念化句嵌入模型 (第 3.2 节).

CSE-SkipGram: 本文所提出的基于 Skip-Gram 模型的概念化句嵌入模型 (第 3.3 节).

aCSE-TYPE: 本文所提出的基于注意力机制的概念化句嵌入模型, 该注意力机制基于词语类型 (第 3.4.1 节).

aCSE-SUR: 本文所提出的基于注意力机制的概念化句嵌入模型, 该注意力机制基于惊异度 (第 3.4.2 节).

aCSE-ALL: 本文所提出的基于注意力机制的概念化句嵌入模型, 该注意力机制同时基于词语类型和惊异度, 即 aCSE-ALL 模型是 aCSE-TYPE 和 aCSE-SUR 的融合. 本文使用线性插值 (Linear interpolation) 方法[25] 对模型 aCSE-TYPE 和 aCSE-SUR 进行融合, 这种融合体现在对词语 w_i 的注意力因子的融合:

$$a_{\mathrm{ALL}}(w_i) = \lambda \cdot a_{\mathrm{TYPE}}(w_i) + (1-\lambda) \cdot a_{\mathrm{SUR}}(w_i) \quad (11)$$

其中, $a_{\mathrm{ALL}}(w_i)$ 表示模型 aCSE-ALL 中词语 w_i 的注意力因子, $a_{\mathrm{TYPE}}(w_i)$ 表示模型 aCSE-TYPE 中词语 w_i 的注意力因子 (式 (8)), $a_{\mathrm{SUR}}(w_i)$ 表示模型 aCSE-SUR 中词语 w_i 的注意力因子 (式 (10)); 参数 λ 控制在融合过程中, $a_{\mathrm{TYPE}}(w_i)$ 和 $a_{\mathrm{SUR}}(w_i)$ 的重要程度 (权重), 即决定了词语类型和惊异度对最终结果的影响力程度.

4.3 实验设置

所有句子均使用 *Porter* 工具包进行词干提取, 使用 *InQuery* 停用词表进行去停. 对于 SCBOW、TWE、CSE-CBOW、CSE-SkipGram 及其注意力变体, 我们使用二元霍夫曼树[5, 33] 作为层次化 Softmax 的架构, 越高频的词语的编码越短以便快速检索, 可以提高训练速度[3]. 句向量、词向量、主题向量、概念向量的维度设置均为 500. 同时, 我们探讨不同上下文片段窗口规模取值 (k 取值从 3 到 11) 对实验结果的影响, 结果显示不同 k 值在不同数据集上会产生不同的实验结果, 最终选择 k 值为 5 的实验结果展示在实验表格中, 因为这个窗口规模能够在大部分数据集上产生最优实验结果. 同理, 通过探讨式 (11) 中的参数 λ 的不同取值, λ 取值为 0.6 的时候, 实验取得最好结果. 对于基于主题模型的 LDA 和 TWE, 在文本分类任务中, 我们将主题数量定为分类类别数量或者类别数量的两倍, 取最优结果进行展示; 在信息检索任务中, 我们探讨了不同主题数量 (100～500) 对实验结果的影响, 结果显示主题数量对实验结果影响微弱, 最终选择主题数量为 500 时的结果进行展示. 对于 aCSE-TYPE 模型, 本文使用 *Stanford* 词性标注工具包完成对文本的词性标注, 该工具包所使用词性标注集为宾州英文树库词性标注集. 对于 aCSE-SUR 模型, 本文使用 *SRILM* 工具包计算 N-Gram 语言模型来获得 aCSE-SUR. 与上下文片段窗口规模选择相对应, 本文使用数据集 NewsTitle、Twitter11 和 Wiki 训练 5 阶 N-Gram 语言模型, 并采用 $Kneser-ney$ 平滑方法.

4.4 文本分类实验及结果分析

我们使用各算法所生成的句向量作为特征, 使用线性分类器 *Liblinear*[39] 在数据集 NewsTitle、Twtter 和 TREC 上完成文本分类任务. 使用准确率 (P)、召回率 (R) 和 F-值 (F) 作为评价指标, 实验结果展示在表 1 中. 本文使用显著性检验来验证实验结果, 上标 α 和 β 分别表示针对模型 SCBOW 和 PV-DM 的显著性提升 ($p < 0.05$).

通过表 1 可知, 本文提出的基于注意力机制的概念化句嵌入模型性能明显优于其他对比算法, 特别是数据集 Twitter 上的召回率: aCSE-ALL 相比最优对比算法 SCBOW 提高 5.2%, 相比目前公认的基线算法 PV-DM 提高 9.2%. 这充分说明本文所提出的概念化句嵌入模型, 相比较于基于主题模

表 1　文本分类任务实验结果

Table 1　Evaluation results of text classification task

数据集	NewsTitle			Twitter			TREC		
模型	P	R	F	P	R	F	P	R	F
BOW	0.731	0.719	0.725	0.397	0.415	0.406	0.822	0.820	0.821
LDA	0.720	0.706	0.713	0.340	0.312	0.325	0.815	0.811	0.813
PV-DBOW	0.726	0.721	0.723	0.409	0.410	0.409	0.825	0.817	0.821
PV-DM	0.745	0.738	0.741	0.424	0.423	0.423	0.837	0.824	0.830
TWE	0.810	0.805	0.807	0.454	0.438	0.446	0.894	0.885	0.885
SCBOW	0.812^{β}	0.805^{β}	0.809^{β}	0.455^{β}	0.439	0.449^{β}	0.897^{β}	0.887^{β}	0.892^{β}
CSE-CBOW	0.814	0.811	0.812	0.458	**0.450**	0.454	0.895	0.891	0.893
CSE-SkipGram	**0.827**	**0.819**	**0.823**	**0.477**	0.447	**0.462**	**0.899**	**0.894**	**0.896**
aCSE-SUR	0.828	0.822	0.825	0.469	0.453	0.462	0.906	0.897	0.901
aCSE-TYPE	0.838	0.830	0.834	0.483	0.455	0.468	0.911	0.903	0.907
aCSE-ALL	$\mathbf{0.845}^{\alpha\beta}$	$\mathbf{0.832}^{\alpha\beta}$	$\mathbf{0.838}^{\alpha\beta}$	$\mathbf{0.485}^{\alpha\beta}$	$\mathbf{0.462}^{\alpha\beta}$	$\mathbf{0.473}^{\alpha\beta}$	$\mathbf{0.917}^{\alpha\beta}$	$\mathbf{0.914}^{\alpha\beta}$	$\mathbf{0.915}^{\alpha\beta}$

型的句嵌入模型以及其他类型句嵌入模型, 能够捕获更加精确的语义信息. 因为概念信息能够显著增强句子的语义表达能力, 而且受文本噪音和稀疏性影响较小. 算法 SCBOW 通过将对 "词向量累加平均得到的句向量" 的对比加入到损失函数中, 直接面向 "词向量累加平均得到句向量" 这个最终目标来优化传统词向量的训练. 但是该算法仅利用字面信息对短文本进行建模, 这也是性能劣于本文所提出的算法的原因.

实验结果同样证明, 通过引入注意力机制, 概念化句嵌入模型性能得到全面提升 (例如 aCSE-ALL 与 CSE-CBOW 的对比), 足以证明注意力机制的优势. 其中, aCSE-TYPE 性能优于 aCSE-SUR, 说明相比较于词语惊异度, 词语类型 (词性等) 是构建句嵌入注意力机制更合适的选择. 我们同样探索了这两种注意力机制是否可以实现互补: 融合惊异度和词语类型的注意力模型 aCSE-ALL 的性能优于 aCSE-TYPE 和 aCSE-SUR, 说明惊异度和词语类型是句子语义建模的不同方面, 包含不同的语义信息, 二者结合能够生成更好的注意力机制.

同样是基于词嵌入研究中的 Skip-Gram 模型, 我们的 CSE-SkipGram 性能优于 TWE, 在 F- 值方面, 在数据集 Twitter 和 NewsTitle 上分别较 TWE 提升 3.6 % 和 2.0 %. 二者的不同之处在于, CSE-SkipGram 侧重利用概念信息来增强句子表示, 而 TWE 侧重利用主题信息来增强句子表示. 虽然 TWE 尝试引入主题模型力求增强语义表达能力, 但是由于短文本缺乏足够信号用于推理, 导致无论是句法分析还是主题模型都很难奏效. 此外, TWE 通过简单地将句中所有词的主题化词向量求平均来获得句向量, 忽略了语素之间的语义关联, 极大限制了所产生的句向量的语义表达能力, 同样的问题还存在于例如文献 [17] 的工作.

总的来说, 几乎所有对比算法都在数据集 Twitter 上出现了性能下滑, 特别是 LDA 和 TWE. 这主要因为 Twitter 中的数据噪音大、稀疏和歧义, 给基于主题模型的算法带来很大挑战; 另一个因素在于社交文本的虚词 (例如感叹词、连词、介词等) 占比较高, 有碍对句子核心语义建模, 使基于词语类型的注意力机制 (aCSE-TYPE) 和基于惊异度的注意力机制 (aCSE-SUR) 的性能受限. 例如, 由表 1 可知, aCSE-ALL 在数据集 NewsTitle 上对 CSE-CBOW 的 F-值提升为 3.2 %, 而上述指标数据集 Twitter 上的提升仅为 2.6 %.

4.5 信息检索实验及结果分析

评估相关模型的第二个任务是短文本信息检索任务, 该任务使用推文数据集 Tweet11: 给定一个查询项, 我们评估相关模型是否可以检索到语义最相关的推文. 本文使用 *Indri* 工具包索引推文数据, 并执行一个类似的伪相关反馈 (Relevance-pseudo feedback, PRF)[25] 的检索过程: 1) 给定一个查询项, 我们首先基于初始检索获取相关推文, 作为反馈推文; 2) 我们使用不同算法为原始查询项和反馈推文生成句向量; 3) 计算查询项向量与反馈推文向量之间的余弦相似度, 并按相似度值降序排列推文.

我们使用 TREC 微博检索任务官方评测指标 P@30 和平均准确率 (Mean average precision, MAP) 来检验各算法的推文排序效果. 实验结果如表 2 所示, 其中, TMB2011 表示微博检索任务 2011 年所用查询项集合 (包括 49 个带时间戳查询项), TMB2012 表示 2012 年所用查询项集合 (包括 60 个带时间戳查询项). 在该任务上, 同样使用显著性检验来验证实验结果, 上标 α 和 β 分别表示针对模型 SCBOW 和 PV-DM 的显著性提升 ($p < 0.05$).

表 2 信息检索任务实验结果

Table 2 Evaluation results of information retrieval

查询项集合	TMB2011		TMB2012	
模型	MAP	P@30	MAP	P@30
BOW	0.304	0.412	0.321	0.494
LDA	0.281	0.409	0.311	0.486
PV-DBOW	0.285	0.412	0.324	0.491
PV-DM	0.327	0.431	0.340	0.524
TWE	0.331	0.446	0.347	0.509
SCBOW	0.333	0.448^{β}	0.349^{β}	0.511
CSE-CBOW	0.337	0.451	0.344	0.512
CSE-SkipGram	**0.367**	**0.461**	**0.360**	**0.517**
aCSE-SUR	0.342	0.458	0.349	0.520
aCSE-TYPE	0.373	0.466	0.365	0.525
aCSE-ALL	$\mathbf{0.376}^{\alpha\beta}$	$\mathbf{0.471}^{\alpha\beta}$	$\mathbf{0.369}^{\alpha\beta}$	$\mathbf{0.530}^{\alpha\beta}$

通过表 2 可知, aCSE-ALL 性能显著优于所有对比算法, 在 TMB2011 上将最优对比算法 SCBOW 的 MAP 值和 P@30 值分别提升了 13.4 % 和 5.6 %. 这种显著提升可以归因于本文的模型能够将深层次上下文语义信息嵌入到句向量中. 此外, 同样可以观察到, 基于注意力机制的概念化句嵌入模型要优于基础概念化句嵌入模型. 相比较于基于词语类型的注意力机制, 基于惊异度的注意力机制在社交文本上的对于基础概念化句嵌入的提升有限: 在 TMB2011 上的 MAP 指标, aCSE-SUR 仅将基础概念化句嵌入模型 (CSE-CBOW) 提升了 1.5 %, 性能甚至还不如 CSE-SkipGram, 而基于词语类型的注意力机制 (aCSE-TYPE) 则将基础模型显著提升了 10.7 %. 这主要原因是, 由于书写不规范, 社

交文本中的词语片段在大规模语料中出现次数过少, 导致基于统计的惊异度性能受限. 与文本分类实验类似, 基于主题模型的相关算法 (LDA 和 TWE) 受社交文本数据集的噪音和稀疏性影响严重.

5 结论

通过引入概念信息, 本文所提出的概念化句嵌入模型能够保持和增强句向量的语义表达能力和甄别能力. 在此基础上, 为了模拟人类阅读行为, 我们引入基于词语类型的注意力机制和基于惊异度的注意力机制, 来扩展上述概念化句嵌入模型, 允许模型有选择性地处理上下文窗口中的语境词语, 为对句子语义建模有帮助的语境词语赋予更高注意力值和重视程度, 进一步增强了句向量的表达能力. 在实验中, 我们将所提出的基于注意力机制的概念化句嵌入模型与多种类型公认的基线系统进行比较. 实验结果证明, 概念化句嵌入模型性能优于其他模型, 而且在短文本上具有良好的抗数据噪音和稀疏性能力.

References

1 Harris Z S. Distributional Structure. *Word*, 1954, **10**: 146−162

2 Palangi H, Deng L, Shen Y, Gao J, He X, Chen J, Song X, Ward R. Deep Sentence Embedding Using Long Short-Term Memory Networks: Analysis and Application to Information Retrieval. *IEEE/ACM Transactions on Audio, Speech, and Language Processing*, 2016, **24**: 694−707

3 Le Q, Mikolov T. Distributed Representations of Sentences and Documents. In: Proceedings of the 31st International Conference on Machine Learning. New York, NY, USA: ACM, 2014. 1188−1196

4 Liu Y, Liu Z, Chua T S, Sun M. Topical Word Embeddings. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2015. 2418−2424

5 Mikolov T, Corrado G, Chen K, Dean J. Efficient Estimation of Word Representations in Vector Space. arXiv preprint arXiv: 1301.3781, 2015. http://arxiv.org/abs/1301.3781

6 Wang Z, Zhao K, Wang H, Dean J. Query Understanding through Knowledge-Based Conceptualization. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann, 2015. 3264−3270

7 Bahdanau D, Cho K, Bengio Y. Neural Machine Translation by Jointly Learning to Align and Translate. arXiv preprint arXiv: 1409.0473, 2015. http://arxiv.org/abs/1409.0473

8 Lin Y, Shen S, Liu Z, Luan H, Sun M. Neural Relation Extraction with Selective Attention over Instances. In: Proceedings of the 54th Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. 2124−2133

9 Wang Y, Huang H, Feng C, Zhou Q, Gu J, Gao X. CSE: Conceptual Sentence Embeddings based on Attention Model. In: Proceedings of the 54th Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. 505−515

10 Rayner K. Eye movements in Reading and Information Processing: 20 Years of Research. *Psychological Bulletin*, 1998, **124**: 372−422

11 Nilsson M, Nivre J. Learning where to look: Modeling eye movements in reading. In: Proceedings of the 13th Conference on Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2009. 93−101

12 Lai S, Xu L, Liu K, Zhao J. Recurrent Convolutional Neural Networks for Text Classification. In: Proceedings of the 29th AAAI Conference on Artificial Intelligence. Palo Alto, CA, USA: AAAI, 2015. 2267−2273

13 Kenter T, Borisov A, De Rijke M. Siamese CBOW: Optimizing Word Embeddings for Sentence Representations. In: Proceedings of the 54th Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. 941−951

14 Xiong C, Merity S, Socher R. Dynamic Memory Networks for Visual and Textual Question Answering. In: Proceedings of the 33rd International Conference on Machine Learning. New York, NY. USA: ACM, 2016. 1230−1239

15 Yu J. Learning Sentence Embeddings with Auxiliary Tasks for Cross-Domain Sentiment Classification. In: Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2016. 236−246

16 Yong Z, Meng J E, Ning W, Pratama M. Attention Pooling-based Convolutional Neural Network for Sentence Modelling. *Information Sciences*, 2016, **373**: 388−403

17 Wang S, Zhang J, Zong C. Learning sentence representation with guidance of human attention. In: Proceedings of IJCAI International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann, 2017. 4137−4143

18 Lang Z, Gu X, Zhou Q, Xu T. Combining statistics-based and CNN-based information for sentence classification. In: Proceedings of the 28th IEEE International Conference on Tools with Artificial Intelligence. New York, NY. USA: IEEE, 2017. 1012−1018

19 Wieting J, Gimpel K. Revisiting Recurrent Networks for Paraphrastic Sentence Embeddings. In: Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. 2078−2088

20 Nicosia M, Moschitti A. Learning Contextual Embeddings for Structural Semantic Similarity using Categorical Information. In: Proceedings of the 21st Conference on Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2017. 260−270

21 Peng W, Wang J, Zhao B, Wang L. Identification of Protein Complexes Using Weighted PageRank-Nibble Algorithm and Core-Attachment Structure. *IEEE/ACM Transactions on Computational Biology & Bioinformatics*, 2015, **12**: 179−192

22 Hua W, Wang Z, Wang H, Zheng K, Zhou X. Short text understanding through lexical-semantic analysis. In: Proceedings of the 31st IEEE International Conference on Data Engineering. New York, NY. USA: IEEE, 2015. 495−506

23 Song Y, Wang H. Open Domain Short Text Conceptualization: A Generative + Descriptive Modeling Approach. In: Proceedings of the 24th International Joint Conference on Artificial Intelligence. San Francisco, CA, USA: Morgan Kaufmann, 2015. 3820−3826

24 Wang F, Wang Z, Li Z, Wen J R. Concept-based Short Text Classification and Ranking. In: Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management. New York, NY. USA: ACM, 2014. 1069−1078

25 Wang Y, Huang H, Feng C. Query Expansion Based on a Feedback Concept Model for Microblog Retrieval. In: Proceedings of the 26th International Conference on World Wide Web. New York, NY. USA: ACM, 2017. 559−568

26 Wu W, Li H, Wang H, Zhu K Q. Probase: A probabilistic taxonomy for text understanding. In: Proceedings of the 2012 ACM SIGMOD International Conference on Management of Data. New York, NY. USA: ACM, 2012. 481−492

27 Yu Z, Wang H, Lin X, Wang M. Understanding Short Texts through Semantic Enrichment and Hashing. *IEEE Transactions on Knowledge and Data Engineering*, 2016, **28**: 566−579

28 Itti L, Koch C, Niebur E. A Model of Saliency Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1998, **20**: 1254−1259

29 Hahn M, Keller F. Modeling Human Reading with Neural Attention. *Psychological Bulletin*, 2016, **85**: 618−627

30 Narayanan S, Jurafsky D. A Bayesian Model Predicts Human Parse Preference and Reading Times in Sentence Processing. *Advances in Neural Information Processing Systems*, 2001, **14**: 59−65

31 Demberg V, Keller F. Data from eye-tracking corpora as evidence for theories of syntactic processing complexity. *Cognition*, 2016, **109**: 193−210

32 Barrett M, Sogaard A. Reading behavior predicts syntactic categories. In: Proceedings of the 2015 SIGNLL Conference on Computational Natural Language Learning. Stroudsburg, PA, USA: Association for Computational Linguistics, 2015. 345−349

33 Mikolov T, Sutskever I, Chen K, Corrado G, Dean J. Distributed Representations of Words and Phrases and their Compositionality. In: Proceedings of the 27th Annual Conference on Neural Information Processing Systems. Cambridge, MA, USA: MIT Press, 2013. 1−9

34 Fernandez-Carbajales V, García M A, MartU.S.A.nez J M. Visual attention based on a joint perceptual space of color and brightness for improved video tracking. *Pattern Recognit*, 2016, **60**: 571−584

35 Attneave F. Applications of information theory to psychology: a summary of basic concepts, methods, and results. *American Journal of Psychology*, 1961, **74**(2): 319−324

36 Hale J. A probabilistic earley parser as a psycholinguistic model. In: Proceedings of the 2nd meeting of the North American Chapter of the Association for Computational Linguistics on Language technologies. Stroudsburg, PA, USA: Association for Computational Linguistics, 2001. 1−8

37 Ounis I, Macdonald C, Lin J. Overview of the trec-2011 microblog track. In: Proceedings of the 2011 Text REtrieval Conference. Gaithersburg, MD, USA: NIST, 2001. 1−9

38 Blei D M, Ng A Y, Jordan M I. Latent Dirichlet Allocation. *Journal of Machine Learning Research*, 2003, **3**: 993−1022

39 Fan R E, Chang K W, Hsieh C J, Wang X R, Lin C J. LIBLINEAR: A Library for Large Linear Classification. *Journal of Machine Learning Research*, 2008, **9**: 1871−1874

王亚珅 北京理工大学计算机学院博士研究生. 2012 年获得北京理工大学计算机学院学士学位. 主要研究方向为自然语言处理与社交网络分析.
E-mail: yswang@bit.edu.cn
(**WANG Ya-Shen** Ph. D. candidate at the School of Computer, Beijing Institute of Technology. He received his bachelor degree from the School of Computer, Beijing Institute of Technology in 2012. His research interest covers natural language processing and social media analysis.)

黄河燕 北京理工大学计算机学院教授. 1989 年获得中国科学院计算技术研究所博士学位. 主要研究方向为自然语言处理和机器翻译. 本文通信作者.
E-mail: hhy63@bit.edu.cn
(**HUANG He-Yan** Professor at the School of Computer, Beijing Institute of Technology. She received her Ph. D. degree from Institute of Computer Technology, China Academy of Sciences in 1989. Her research interest covers natural language processing and machine translation. Corresponding author of this paper.)

冯 冲 北京理工大学计算机学院副研究员. 2005 年获得中国科学技术大学博士学位. 主要研究方向为信息抽取和情感分析.
E-mail: fengchong@bit.edu.cn
(**FENG Chong** Associate professor at the School of Computer, Beijing Institute of Technology. He received his Ph. D. degree from University of Science and Technology of China in 2005. His research interest covers information extraction and sentiment analysis.)

周 强 百度公司研发工程师. 2016 年获得北京理工大学计算机学院硕士学位. 主要研究方向为自然语言处理与社交网络分析.
E-mail: qzhou@bit.edu.cn
(**ZHOU Qiang** Research engineer at Baidu Inc. He received his master degree from Beijing Institute of Technology in 2016. His research interest covers natural language processing and deep learning.)