软件学报 ISSN 1000-9825, CODEN RUXUEW
2025,36(11):5213−5226 [doi: 10.13328/j.cnki.jos.007373] [CSTR: 32375.14.jos.007373]

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

基于释义知识浮动注入的汉语成语误用诊断*

何 亮, 曹永昌, 黄琰琛, 吴 震, 戴新宇, 陈家骏

(计算机软件新技术全国重点实验室 (南京大学), 江苏 南京 210023)

通信作者: 吴震, E-mail: wuz@nju.edu.cn

摘 要: 汉语成语作为汉语写作的重要组成部分, 具有凝练的表现力和深厚的文化内涵. 它们通常是经过长期使用而固定下来的词组或短句, 来源广泛, 含义相对固定. 然而, 由于汉字的形意属性和汉语词汇、语义的古今变迁, 成语的字面意思与实际含义往往存在偏差, 呈现出特有的非组合性特点, 这种特点使得成语在使用过程中极易产生误用现象, 研究显示, 某些成语的误用率甚至高达 98.6%. 与其他语言不同, 汉语成语的误用通常不会导致词法或语法错误, 因此传统的拼写或语法错误检测方法无法有效识别成语误用. 一种直观的方法是将成语的释义融入模型中, 但是简单的拼接释义会导致句子过长难以处理和知识噪声等问题. 为了解决这一问题, 提出一种基于释义知识浮动注入的模型. 该模型通过引入可学习的权重因子来控制知识注入, 并探讨有效的释义知识注入策略. 为了验证模型的有效性, 构建一套针对汉语成语误用诊断的数据集. 实验结果显示, 该模型在所有测试集上均取得了最优效果, 特别是在长文本多成语的复杂场景中, 性能比基线模型提高了 12.4%–13.9%, 同时训练速度提升了 30%–40%, 测试速度提升了 90%. 这证明了所提出的释义知识浮动注入模型不仅有效融合了成语释义特征, 还显著降低了成语释义拼接对模型处理能力和效率的负面影响, 从而提升了成语误用诊断的性能, 并增强了模型处理多成语和长释义等复杂场景的能力.

关键词: 汉语成语; 误用诊断; 释义知识; 浮动注入; 成语误用数据集

中图法分类号: TP18

中文引用格式: 何亮, 曹永昌, 黄琰琛, 吴震, 戴新宇, 陈家骏. 基于释义知识浮动注入的汉语成语误用诊断. 软件学报, 2025, 36(11): 5213–5226. http://www.jos.org.cn/1000-9825/7373.htm

英文引用格式: He L, Cao YC, Huang YC, Wu Z, Dai XY, Chen JJ. Chinese Idiom Misuse Diagnosis Based on Levitating Injection of Interpretation Knowledge. Ruan Jian Xue Bao/Journal of Software, 2025, 36(11): 5213–5226 (in Chinese). http://www.jos.org.cn/1000-9825/7373.htm

Chinese Idiom Misuse Diagnosis Based on Levitating Injection of Interpretation Knowledge

HE Liang, CAO Yong-Chang, HUANG Yan-Chen, WU Zhen, DAI Xin-Yu, CHEN Jia-Jun

(State Key Laboratory for Novel Software Technology (Nanjing University), Nanjing 210023, China)

Abstract: Chinese idioms, as an essential part of Chinese writing, possess concise expressiveness and profound cultural significance. They are typically phrases or short sentences that have become fixed through long-term use, with diverse origins and relatively stable meanings. However, due to the pictographic nature of Chinese characters and the historical evolution of Chinese vocabulary and semantics, there is often a discrepancy between the literal and actual meanings of idioms, which exhibits a unique non-compositional characteristic. This feature makes idioms prone to misuse of idioms in practice, with research showing that certain idioms are misused at a rate as high as 98.6%. Unlike in other languages, the misuse of Chinese idioms does not typically result in lexical or grammatical errors, which makes traditional spelling and grammar error detection methods ineffective at identifying idiom misuse. An intuitive approach is to incorporate the interpretations of idioms into the model, but simply combining these interpretations can lead to problems such as excessively long

* 基金项目: 国家自然科学基金 (62206126, 62376120)
收稿时间: 2024-08-21; 修改时间: 2024-11-08; 采用时间: 2024-12-07; jos 在线出版时间: 2025-04-30
CNKI 网络首发时间: 2025-05-06

sentences that are hard to process and noise in knowledge. To address this, this study proposes a novel model that uses levitating knowledge injection to incorporate idiom interpretations. This model introduces learnable weight factors to control the injection process and explores effective strategies for knowledge infusion. To validate the model's effectiveness, a dataset specifically for diagnosing the misuse of Chinese idioms is created. Experimental results show that the model achieves optimal performance across all test sets, particularly in complex scenarios involving long texts and multiple idioms, where its performance improves by 12.4%–13.9% compared to the baseline model. At the same time, training speed increases by 30%–40%, and testing speed is improved by 90%. These results demonstrate that the proposed model not only effectively integrates the interpretative features of idioms but also significantly reduces the negative impact of interpretation concatenation on the model's processing capacity and efficiency, thus enhancing the performance of Chinese idiom misuse diagnosis and strengthening the model's ability to handle complex scenarios with multiple idioms and lengthy interpretations.

Key words: Chinese idiom; misuse diagnosis; interpretation knowledge; levitating injection; idiom misuse dataset

成语作为汉语中凝练且富有表现力的语言形式, 不仅蕴含着丰富的文化内涵, 而且在语言表达中扮演着关键角色. 成语通常是长期使用、定型化的词组或短句, 具有整体性, 其来源广泛, 含义相对固定. 然而, 汉字的形意属性以及汉语词汇、语义的古今变迁导致很多词素的意义已发生了翻天覆地的变化, 于是在理解词义的过程中, 人们不免会以"今"度"古"[1], 仅简单地以字面意思组合来理解成语的含义往往会有偏差, 这造成了汉语成语含义具有特有的非组合性特点, 使得成语在使用过程中容易出现误用现象. 表 1 中展示了日常写作中常见的成语误用实例和类型. 其中, 加粗文字是被误用的成语, 括号内的文字是对应成语的正确含义. "浮光掠影"本意形容印象不深刻, 但因为看到"光""影", 人们就误以为可以用来形容"河面", 这种误解来源于汉语言文字中"因形推义"的习惯; "毫发不爽"比喻一点点细小的差错也没有, 但因为其组合方式与"一毛不拔" (形容为人非常吝啬自私) 相似, 使人们容易对这两个词进行"类推同化"从而导致误用; "首当其冲"比喻最先受到攻击或遇到灾难, 而从每个字的现代意义上去理解容易误解为首先要做的事; "美轮美奂"多用于形容建筑物雄伟壮观富丽堂皇, 也用来形容雕刻或建筑艺术的精美效果, 是典型的有明确适用对象的成语, 而人们经常犯"对象错置"的使用错误; "无所不为"虽然表达了什么事都干的意思, 但则带有强烈的贬义, 在上下文中明显是误用了. 统计发现, 不仅是汉语作为外语 (Chinese as foreign language, CFL) 学习者, 即使是母语为汉语的作者, 对部分成语的误用率甚至高达惊人的 98.6%[2], 这说明成语误用现象已非常严重. 成语误用作为一种消极的语言现象, 不仅影响了语言的准确性和美感, 还对跨文化交流和语言学习者的语言习得造成了障碍.

表 1　常见汉语成语误用实例及类型

成语误用实例	误用类型
浮光掠影 (形容印象不深刻)的河面、郁郁葱葱的林园	因形推义
这家伙办事**毫发不爽** (比喻一点点细小的差错也没有)、小气极了	类推同化
发展低碳经济**首当其冲** (比喻最先受到攻击)的是要坚持节约资源	望文生义
它的玻璃屏幕和纤薄机身**美轮美奂** (形容建筑物雄伟壮观)	对象错置
这些年轻的科学家决心以**无所不为** (指什么坏事都干或干尽了坏事)的勇气	褒贬颠倒

与其他语言不同, 汉语成语误用通常不会造成词法或语法错误, 导致传统的拼写或语法错误检测方法无法有效诊断成语误用. 预训练语言模型在很多自然语言处理任务中表现出非凡的性能, 但因其预训练所使用的互联网语料库中本身存在大量成语误用实例, 大多数基于预训练模型的中文文本校对模型对成语误用并不敏感. 表 2 展示了一些现有的优秀模型在汉语成语误用诊断任务上的性能 (该评估基于本文构建的成语误用数据集, 详见第 3.1 节). 其中 Soft-masked BERT[3] 和 ReaLiSe[4] 是中文拼写纠错任务的 SOTA 模型, ResBERT[5] 和 Cao 等人[6]是中文语法错误诊断任务上的 SOTA 模型. 初步实验结果表明, 中文拼写纠错或语法错误诊断任务的模型在诊断成语误用任务上的 $F1$ 分数都仅为 20% 左右, 远低于这些模型在检测字符或单词错误时的效果. 因此, 需要研究更有效的针对成语误用诊断的模型.

为了让模型能够正确理解成语含义, 直接且有效的方法是将成语的完整释义知识引入模型. Zheng 等人[7]验证了将成语视为独立的语义单元优于对成语字符进行单独建模, 因此出现了一系列使用成语的元信息或大规模语料库来实现成语嵌入训练的工作, 并成功应用到成语阅读理解 (或者叫完形填空) 任务中[8−11]. 另一类方法则基于

动态释义编码, 例如 K-BERT[12] 利用软位置和可见矩阵将相关的元组拼接到 BERT 的输入序列中来注入领域知识, 为诊断模型提供更可靠的词汇特征并缓解知识噪声的问题. 以上方法在成语阅读理解任务中取得了不错的效果, 但在成语误用诊断任务中依然面临着更多挑战. 首先, 在成语较多以及释义较长的情况下, 受到预训练语言模型的最大长度约束, 直接拼接自然语言形式的成语释义知识会因过长的输入而被截断, 难以保留更加完整的词汇和释义信息, 从而降低了诊断的能力和效率. 其次, 成语是否误用不仅依赖于成语本身的含义, 还与上下文密切相关. 例如, 在成语使用与文化背景紧密相关的情况下, 可能需要更多地依赖释义特征, 而在成语的使用更多依赖于句子结构和上下文的情况下, 则可能需要更多地依赖上下文表示. 缺少上下文表示和释义特征的融合调节会导致信息的不完整或噪声的引入.

表 2　现有中文拼写纠错和语法诊断模型在成语误用诊断上的性能表现 (%)

方法	原始任务			成语误用诊断		
	精确率	召回率	*F*1值	精确率	召回率	*F*1值
Soft-masked BERT	73.70	73.20	73.50	10.70	3.90	**5.70**
ReaLiSe	77.30	81.30	79.30	20.50	6.80	**10.20**
ResBERT	66.80	91.67	77.28	12.30	18.90	**14.90**
Cao等人[6]	85.65	97.57	91.22	15.80	33.70	**21.50**

为了解决上述问题, 本文提出一种基于释义知识浮动注入的模型, 通过融合控制知识注入的可学习权重因子, 并探索释义知识注入的有效策略, 在保证成语释义特征融合有效性的同时, 显著降低成语释义拼接操作对诊断模型处理能力和效率的限制, 从而提高成语误用诊断的性能, 并有效增强模型对多成语和长释义等复杂场景的处理能力和效率. 同时, 本文基于汉语成语完形填空数据集 (ChID[7]) 和中国高考 (Chinese college entrance examination, CEE) 成语选择题数据构建了一个针对汉语成语误用诊断的大规模数据集. 实验结果表明, 本文所提出的模型在汉语成语误用诊断任务上优于基线诊断模型, 并在多成语和长释义等复杂场景下取得了显著的提升.

综上所述, 本文的主要贡献如下: (1) 本文明确提出了汉语成语误用诊断这一研究任务, 并针对该任务构建了一个大规模的数据集. 该数据集基于汉语成语完形填空数据集 (ChID) 和中国高考 (CEE) 成语选择题数据构建, 为成语误用诊断的研究提供了丰富的实验材料和评估基准. (2) 本文提出了一种创新的模型, 该模型通过浮动注入成语的释义知识, 能够在保持释义特征融合有效性的同时, 显著降低知识噪声和因模型最大输入长度限制而导致的信息丢失问题, 增强了模型对多成语和长释义等复杂场景的处理能力. (3) 通过一系列实验, 本文充分验证了所提出模型在成语误用诊断任务上的有效性, 证明了该方法在实际应用中的潜力和价值.

本文第 1 节重点介绍成语词嵌入学习和词汇知识增强两类汉语成语误用诊断的相关工作. 第 2 节介绍本文构建的基于释义知识浮动注入的模型. 第 3 节通过对比实验验证所提模型的有效性. 最后总结全文.

1　相关工作

据文献检索, 目前尚未有文献对中文成语误用诊断的任务展开专门研究. 尽管如此, 已经有许多工作探索了成语词嵌入学习和词汇知识增强方法, 这些方法都有助于开展中文成语误用诊断任务的研究.

1.1　成语词嵌入学习

大量的研究已证实, 在大型语料库上预训练的模型能够学到通用的语言表示, 这对后续的自然语言处理 (NLP) 任务大有裨益. 自从谷歌提出了 Transformer 模型以来[13], 基于该架构的不同变体及其训练任务的预训练语言模型层出不穷. Transformer 架构主要包括编码器和解码器两部分. 编码器由多头注意力模型 (multi-head attention, MHA) 和前馈网络 (feed-forward network, FFN) 组成, 自注意力机制允许模型直接捕捉输入序列中任意两个词之间的关系, 从而有效地解决序列间的长距离依赖问题. 解码器除了包含类似的自注意力机制外, 还引入了跨注意力模块 (cross-attention), 用于与编码器产生的表示进行交互.

预训练语言模型 (pre-trained language model, PLM) 之间的差异主要体现在选择的编码器架构、预训练任务

以及最终的应用目标上[14]. 例如, OpenAI 的 GPT[15]系列采用了 Transformer 解码器结构, 并以语言模型任务 (预测给定文本序列后的下一个词) 作为无监督学习任务. 而 Google 的 BERT[16]和 Facebook 的 RoBERTa[17]则利用双向 Transformer 编码器, 并通过掩码语言模型 (masked language model, MLM) 进行预训练. Facebook 的 BART 模型[18]结合了 BERT 的双向编码特性和 GPT 的非自回归解码特性, 使用去噪自编码器 (denoising autoencoder, DAE) 对被噪声破坏的文本进行重建. 除了这些基础框架外, 还有多种针对特定领域的改进, 例如增强特定领域知识的模型[19]、多语言模型[20]、多模态模型[21]以及为特定任务定制的模型, 如 Uni-Fold[22] 等.

为了评估成语表示的质量, Zheng 等人[7]创建了一个成语完形填空形式的数据集 ChID, 用于检验模型对于成语语义的理解能力. 实验结果表明, 将成语视为一个整体语义单元进行建模比基于单个字符的方法更为有效. 当前的成语嵌入训练策略大致可以分为两大类: 一类是基于成语词汇信息的方法, 例如 Long 等人[8]提出了一种利用成语近义关系图来增强成语嵌入的方法, 并通过图注意力网络整合了成语的近义关系. Tan 等人[9]利用上下文感知池化操作设计了一个双嵌入模型, 用于同时融合成语的词汇属性和全局上下文信息. 这一类模型的训练语料库都受到 ChID 的限制, 导致成语集合的不完整. 另一类则是基于大规模语料进行预训练的方法, 如 Tan 等人[10]基于网络电子书文本资源构建了一个包含数百万条成语实例的语料库, 并使用掩码语言模型 (MLM)[16]来预训练专门的成语表示模型 chengyuBERT. 虽然这种方法大大提高了成语覆盖度, 但互联网上的电子书语料库不可避免地包含误用噪声样本, 导致 chengyuBERT 对成语的误用不够敏感.

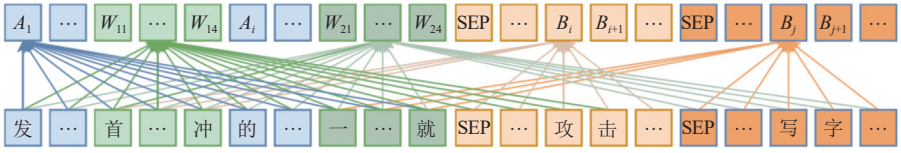
1.2 词汇知识增强

由于通用基于字符的中文 PLM[23−25]没有融合词汇信息, 先前的研究将词汇知识与这些 PLM 的优点集成起来. 利用中文词汇信息进行增强的策略主要可分为两种方式, 其一是直接在字符嵌入中融合词汇信息, Ma 等人[26]通过在 NER 任务上使用基于 LSTM 的浅融合层来融合输入序列的 BERT 表示、离散特征以及词汇信息, 直接在字符表示中融合词汇信息编码. 其二是利用动态结构在编码器中注入词汇信息, 如 FLAT 模型[27]通过相对位置编码策略将用于融合词汇信息的传统格结构[28]转化为有向无环图式的平面结构, 并成功在 Transformer 架构中引入词汇信息. LEBERT [29]直接在 Transformer 层间集成词典信息, 将词典信息适配器应用在相邻的编码器层间.

K-BERT[12]将知识库三元组作为词汇知识拼接到输入序列, 然而, 引入成语释义的方式存在改变原句语义的风险, 即知识噪声 (knowledge noise, KN) 问题, 因此非成语字符不应将自注意力权重分配给引入的释义字符. K-BERT 通过调整位置编码和注意力可视矩阵避免知识噪声问题. 借鉴该做法, 针对成语误用纠错任务时, 假设原输入句表示为 X, 释义句表示为 C, 则 ([CLS], x_1, …, x_n, [SEP], c_1, …, c_t) 即为拼接释义后的组合句序列, 其中 x_i 表示原输入句第 i 个字符, c_i 表示释义句第 i 个字符. 使用自注意力可见矩阵来遮蔽自注意力查询的范围, 自注意力可见矩阵的定义如公式 (1) 所示.

$$\begin{cases} M_{ij}=\begin{cases}-\infty, & (i,j)\in \hat{x}\times C\\ 0, & \text{otherwise}\end{cases}\\ S=Softmax\left(\dfrac{QK^{\top}+M}{\sqrt{d_k}}\right)\end{cases} \tag{1}$$

其中, × 表示笛卡尔积, $\hat{x}$ 表示非成语字符, S 表示 Transformer 编码器中的自注意力分数. 基于公式 (1), 原输入句中的字符与拼接的释义字符可以忽略彼此的编码信息, 仅通过输入中蕴含的成语字符进行信息交互. 图 1 展示了组合句的自注意力可视化示例, 其中第 1 个 [SEP] 之前的字符为原始输入句子, W_*为成语部分, 每一个 [SEP] 开头的字符串为对应各成语的释义. 图中的连接表示允许从编码器上一层到编码器下一层的自注意力分配.



发展经济首当其冲的是科技, 不能一挥而就. [SEP]首当其冲: 比喻最先受到攻击. [SEP]一挥而就: 形容写字、写文章快.

图 1 使用固定释义拼接的注意力可视化示例

2 具体方法

2.1 问题形式化

借鉴通用语法错误诊断任务, 本文使用非重叠的序列标注作为汉语成语误用诊断任务的基本形式. 汉语成语误用诊断问题可以形式化如下: 给定一个包含成语的中文输入序列 $X=(x_1,x_2,\ldots,x_i,\ldots,x_n)$, 其中 x_i 表示句子中成语的字符. 模型需要识别 X 中蕴含的成语是否符合其上下文语义, 并标记出误用成语的位置. 最终模型输出一组误用成语的位置 $P=\{(b,e)_i\}_{i=1}^{\|P\|}$, 其中 $(b, e)_i$ 表示一个误用成语的位置区间, b 和 e 分别代表误用成语的起始位置和结束位置, 误用成语区间在 X 中是不重叠的.

2.2 释义知识浮动注入模型总体框架

尽管以 K-BERT 为代表的基于注意力可见矩阵的释义注入方法可以为诊断模型提供更加可靠的词汇特征并缓解知识噪声的问题, 但预训练语言模型通常具有最大长度约束, 如 BERT-based 模型的最大输入长度为 512 个字符, 而成语释义的长度远大于 K-BERT 原始框架输入中规范化三元组知识的长度, 这使得固定释义拼接的方式严重限制了诊断模型的处理能力和效率. 过长的原始输入或较多的成语释义都将导致输入被截断, 而类似 XLNet[30] 此类能够处理更长序列的预训练语言模型很难适配特定的注意力可视化矩阵技术.

为了能够将原始输入与成语释义句分离, 本文提出了一种释义知识浮动注入模型, “浮动”是指释义知识不是静态地与输入句子拼接, 而是动态地根据模型的需要进行注入. 具体来说, 模型允许将编码的成语词汇特征在不同的输入句子编码层之间动态注入, 而不是固定在某一特定层, 基本框架如图 2 所示. 通过解耦, 编码器的输入句被转化为一个原输入句 $([\mathrm{CLS}],x_1,\ldots,x_i,\ldots,x_n)$ 和一到多个成语释义句 $([\mathrm{CLS}],c_1,\ldots,c_*)$. 当 x_i 是成语字符时, 选取对应成语释义 [CLS] 标签的隐层表示作为相应成语的词汇特征 w_i, 当 x_i 不是成语字符时, 对应的成语词汇特征 w_i 被替换为 [PAD] 填充. 同时, 引入一个可学习的权重因子 α 来控制成语词汇特征与句子表示的融合程度, 将基于成语释义编码得到的成语词汇特征与原输入句融合, 这个权重因子在训练过程中自动调整, 以达到最优的注入效果. 释义知识浮动注入的运算过程如公式 (2) 所示, 其中 LN 表示 LayerNorm 算子, $Transformers_\ell$ 表示编码器的第 ℓ 层. ℓ 和 j 不一定具有对应关系. 释义知识浮动注入使得模型能够更加灵活地处理不同长度和复杂度的成语释义, 从而提高了模型的适应性和效果.

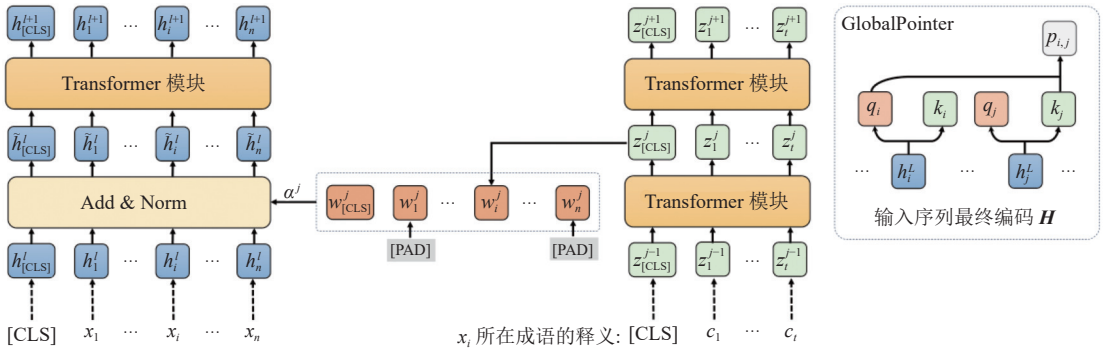


图 2 释义知识浮动注入模型基本框架

$$\begin{cases}\widetilde{h}_i^{\ell}=LN\left(h_i^{\ell}+\alpha^j w_i^j\right)\\ H^{\ell+1}=Transformers_{\ell}\left(\widetilde{H}^{\ell}\right)\end{cases}\tag{2}$$

预测误用成语区间时, 考虑到以实体粒度进行预测更加符合成语误用检测的场景, 本文采用 GlobalPointer[31] 预测头代替基线模型中常用的 CRF 层. 假设输入序列的最终编码为 $\boldsymbol{H}=(h_1^L,h_2^L,\ldots,h_n^L)$, 判定区间 (i,j) 是否为一个误用成语的过程如公式 (3) 所示.

$$\begin{cases} \boldsymbol{q}_i = \boldsymbol{W}_q \boldsymbol{h}_i^L + \boldsymbol{b}_q \\ \boldsymbol{k}_j = \boldsymbol{W}_k \boldsymbol{h}_j^L + \boldsymbol{b}_k \\ p_{i,j} = \boldsymbol{q}_i^{\top} \boldsymbol{k}_j \\ \text{s.t.} \quad 1 \leqslant i \leqslant j \leqslant n \end{cases} \tag{3}$$

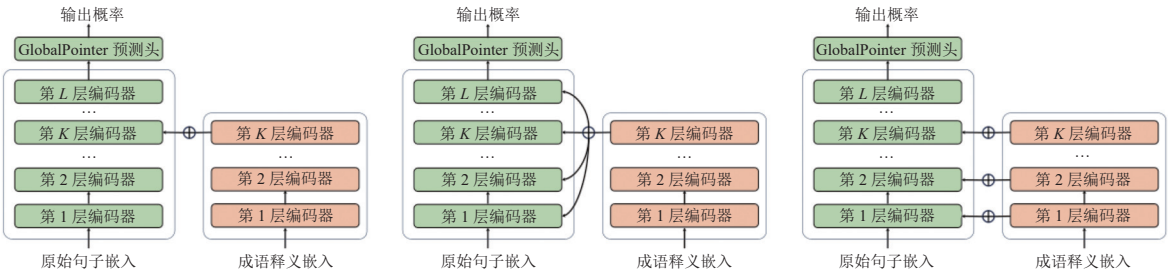
其中, $\boldsymbol{q}_i$ 和 $\boldsymbol{k}_i$ 分别是位置 i 处的查询向量和键向量, $p_{i,j}$ 表示从位置 i 处到位置 j 处为一个误用成语序列的概率. GlobalPointer 预测头的损失函数计算公式如公式 (4) 所示:

$$\mathcal{L} = \log\left(1 + \sum_{(i,j)\in P} \mathrm{e}^{-p_{i,j}}\right) + \log\left(1 + \sum_{(i,j)\in Q} \mathrm{e}^{p_{i,j}}\right) \tag{4}$$

其中, P 是所有误用的成语实体区间的标签集合, 而 Q 表示所有非实体和非误用实体区间的集合.

2.3 释义知识注入策略

受领域特定的任务中对词汇知识利用的启发[27,29], 本文将预训练的成语释义嵌入与通用的基于预训练语言模型相结合. 诊断模型对输入序列进行编码, 预训练的词汇特征被注入到编码器的中间层. 考虑到释义编码器的低层同样在动态编码过程中捕获了成语的表层特征, 这些特征同样有利于建模词汇特征和诊断误用成语, 本文提出了 3 种策略对成语释义知识如何更有效地注入到编码器的中间层进行探索: 顶层适配注入 (top-layer adapter, TLA)、跨层适配注入 (cross-layer adapter, CLA) 和全适配注入 (all-layer adapter, ALA). 顶层适配注入 (TLA) 策略将释义编码器的顶层编码注入到对应的字符编码器层. 顶层编码通常包含更抽象和高层次的语义信息, 可以减少低层噪声的影响, 有助于模型捕捉成语的深层语义特征, 同时计算成本较低, 从而提高模型的计算效率和推理速度. 跨层适配注入 (CLA) 策略利用释义编码器的顶层编码来增强字符编码器的所有层. 通过在多个层次上融合成语释义信息, 模型能够更全面地理解和使用成语的语义特征, 提高对成语误用的识别能力. 全适配注入 (ALA) 策略则将释义编码器的每一层都注入到对应的字符编码器层, 以实现一个完整的层次化信息交互. 通过在每个层次上获取详细的成语释义信息, 模型能够充分利用成语的所有可用信息, 提高模型对成语误用诊断的敏感性和准确性. 这 3 种策略提供了不同层次的信息融合方法, 旨在通过不同方式将成语的释义知识融入到模型中, 以提高模型对成语误用诊断的准确性和效率. 通过对这些策略的探索, 期望能够发现更有效地利用成语释义嵌入提升汉语成语误用诊断性能的方法. 图 3 直观地展示它们之间的差异.



(a) 顶层适配注入 (top-layer adapter, TLA) (b) 跨层适配注入 (cross-layer adapter, CLA) (c) 全适配注入 (all-layer adapter, ALA)

图 3 3 种释义知识浮动注入策略的对比

3 实验分析

3.1 实验数据

本文构建了专门针对中文成语误用诊断任务的大规模数据集 (数据集访问地址: https://github.com/NJUNLP/CIMD), 该数据集主要包含 3 个部分: 1) 训练集及测试集基于 ChID[7]构建, 但进行了一些必要的修改以适应成语误用诊断任务. ChID 数据集的基本形式是完型填空, 即针对句子中每一个成语槽位, ChID 数据集提供 7 个候选成语, 其中包含 1 个正确成语和 6 个错误候选. 本文在错误候选中随机挑选一个成语填入原始句子中作为误用成语,

以模拟实际使用中的误用情况. 并且根据从 ChID 中挑选错误候选方式的不同, 测试集分为随机测试集 (random set, 缩写为 Ran) 和相似测试集 (similar set, 缩写为 Sim). 随机测试集 Ran 中的错误候选是从成语库中随机选取, 相似测试集 Sim 中的错误候选是与正确成语语义相近的成语, 即错误候选与正确成语之间的相似度大于特定的阈值, 这种区分有助于评估模型在处理不同难度的成语误用时的表现. 2) 高考测试集采集了来自中国高考试题 (CEE) 中的成语误用数据. 高考题目设计会追求一定的难度和区分度, 以考查学生的深入理解能力, 利用该数据集可以更好地测试模型在实际场景中的表现. 3) 挑战集包含了输入长度不少于 400 或者包含成语不少于 5 个的长文本实例, 以测试模型在长文本、多成语场景下的性能表现. 表 3 展示了本文所用成语误用诊断数据集的详细统计信息.

表 3　汉语成语误用诊断数据集统计信息

统计项	训练集	Ran测试集	Sim测试集	高考测试集	挑战集
平均长度	158.1	158.5	158.5	35.7	226.6
包含成语总数	629902	24240	24240	394	62
误用成语数量	286406	10606	10606	311	28
误用比例 (%)	45	44	44	79	45
成语平均数量	1.21	1.21	1.21	1.0	6.2

3.2　评价指标及基准模型

在成语误用诊断任务中, 模型需要准确识别出误用的成语片段, 并判断其使用的正确性. 因此, 本文借鉴机器翻译质量评估任务的做法 (https://www2.statmt.org/wmt24/qe-subtask2.html), 选择了跨度级别 (span-level) 的评测方法来评价成语误用诊断任务的性能, 该方法主要用于评估模型对于文本中特定连续片段 (span) 的识别能力, 模型需要识别出文本中的特定区域, 并对其进行分类或标注. 为了全面衡量模型的性能, 本文采用了跨度级别的精确率 (Precision)、召回率 (Recall) 以及 $F1$ 值作为评价指标. 评价指标的计算是以成语为单位的. 当输入句子中包含多个成语时会对每个成语分别进行检测.

本文选取了多个基线模型来评估它们在成语误用诊断任务上的性能, 为了更好地适配这一任务, 本文对这些模型进行了必要的调整: BERT-based 模型利用朴素的 BERT 和 GlobalPointer 预测器作为诊断器, 基线模型不引入任何成语词汇特征信息. FLAT[27]模型采用平面格 (FlatLattice) 结构在 Transformer 中同时融入词汇特征. 本文采用 chengyuBERT 中的预测层权重作为预训练的成语嵌入向量, 将原始 FLAT 论文中的 CRF 预测器替换为 GlobalPointer 预测器, 并去除在其他基线模型中未被使用的二元文法 (bigram) 特征, 以保证各个模型之间的可比性. LEBERT[29] 模型通过词典适配器层直接将外部的词典知识集成到 BERT 层中, 这种适配器被应用于 BERT 的相邻 Transformer 层之间, 使得词典特征和 BERT 表示能够在 BERT 的多层编码器内充分交互. 该模型同样使用 chengyuBERT 的预测层权重作为预训练成语嵌入, 并将 GlobalPointer 作为误用成语区间预测器. K-BERT [12] 模型基于第 1.2 节中固定释义拼接的诊断框架, 引入了 soft-position 以及注意力可视化矩阵技巧以注入自然语言形式的领域知识, 同时建立了一个释义查询表来替代原始模型所利用的三元组知识库, 并利用 GlobalPointer 作为误用成语区间预测器.

3.3　实验细节

本文中的所有实验均基于组件 Transformer (https://github.com/huggingface/transformers) 和 GlobalPointer (https://github.com/gaohongkui/GlobalPointer_pytorch). 本文采用预训练的 BERT-wwm 模型[24]对所有模型进行初始化. 在训练过程中, 本文使用 AdamW 优化器[32]进行 5 个周期 (epochs) 的训练, 并采用 CosineAnnealingScheduler 将学习率从 5E–5 逐渐衰减至 1E–5. 训练批次大小设置为 64, 原始输入句子的最大长度设定为 300. 公式 (3) 中的 $\boldsymbol{q}_i$ 和 $\boldsymbol{k}_i$ 向量的维度设定为 64. 除了上述参数外, 其余实验设置遵循原始论文中所描述的配置.

3.4　实验结果

3.4.1　整体性能对比

表 4 展示了各个模型在汉语成语误用诊断数据集上的实验性能, 实验结果表明:

• 成语词汇特征增强是有效的. 包含成语词汇特征增强的模型, 无论是基于预训练成语嵌入 (FLAT 及 LEBERT) 或编码成语释义特征 (K-BERT 及本文提出的模型), 其诊断结果都显著优于朴素的 BERT 模型, 验证了成语词汇特征增强策略对非组合性的成语建模的有效性.

• 编码成语释义特征总体优于预训练的成语上下文嵌入. 比较基于预训练成语嵌入的诊断模型和基于成语释义特征提取的诊断模型, 后者拥有明显的性能优势. 这是由于以互联网数据为主的用于预训练的原始语料库在未经清洗的情况下广泛存在着成语误用, 预训练的成语嵌入不可避免地包含噪声并且可能造成对成语误用的不敏感. 这表明从成语释义中学习到的成语词汇特征更加适用于汉语成语误用诊断任务.

• 各个模型在 Sim 测试集上的表现普遍低于在 Ran 测试集上的性能. 首先 Sim 测试集中的误用成语与正确成语之间存在更大的相似度, 加大了诊断任务的难度. 其次, 虽然在数据构建的过程中, 尽管 ChID 已经将与正确答案相似度特别高的候选在 Sim 集去除, 以排除两个词是完全同义词的情况, 然而过滤后的候选词仍有可能在上下文中是一个次优的正确结果, 并不构成成语误用实例, 现有的数据构建方法并不能有效地剔除这部分数据.

表 4　模型性能对比 (%)

方法	Ran测试集			Sim测试集		
	精确率	召回率	*F*1值	精确率	召回率	*F*1值
BERT-based	78.0	78.2	78.1	62.3	59.3	60.8
FLAT	82.6	85.0	83.8	73.8	66.4	69.9
LEBERT	84.6	87.0	85.8	77.7	66.4	71.6
K-BERT	85.6	92.4	88.9	82.9	75.3	78.9
TLA	86.2	91.8	88.9	81.8	72.9	77.1
ALA	88.6	92.3	**90.4**	85.2	73.8	**79.1**
CLA	87.7	92.3	89.9	85.0	73.9	**79.1**

3.4.2　高考测试集性能对比

如表 5 所示, 虽然释义知识浮动注入在高考测试集上取得了最优的效果, 但依然有较大的改进空间.

表 5　高考测试集性能对比 (%)

方法	精确率	召回率	*F*1值
BERT-based	80.4	25.1	38.2
LEBERT	81.6	30.5	44.4
K-BERT	88.4	40.8	55.8
TLA	87.5	40.0	54.9
ALA	**90.0**	40.5	55.9
CLA	89.2	**41.5**	**56.5**

所有模型在高考测试集上的性能都明显弱于在 Ran 和 Sim 测试集上的表现, 表明高考测试集具备更大的挑战性. 首先, 高考测试集的平均句长仅有 35.7, 远低于 Ran 和 Sim 测试集的 158.5, 因此模型需要在更短的上下文中判断成语是否匹配. 其次, 高考测试集中的数据是由专家精心设计, 误用成语通常与上下文语境存在较高的相关性, 需要更多知识才能诊断, 如判断情感一致性、表达对象一致性等.

3.4.3　长文本多成语性能对比

表 6 显示在挑战集上, 释义知识浮动注入模型的性能下降明显小于释义固定拼接策略, 这验证了其在更长输入文本、多成语条件下良好的适应性. 释义知识浮动注入策略可以在成语较多以及释义或输入较长时避免截断, 从而保留更加完整的词汇信息. 同时还可以通过可学习因子显式地调节上下文表示和释义特征的融合权重, 进一步降低多成语带来的知识噪声风险.

表 6　基于挑战集测试的长文本多成语复杂场景性能 (%)

方法	Ran测试集$F1$值	Sim测试集$F1$值
K-BERT	68.2 (−20.7)	60.5 (−18.4)
TLA	80.9 (−8.0)	70.0 (−7.1)
ALA	**83.6** (**−6.8**)	**73.1** (**−6.0**)
CLA	82.7 (−7.2)	72.9 (−6.2)

注: 括号内的数值表示长文本多成语场景和综合数据对比性能下降情况

3.4.4　复杂度分析

释义的引入无疑将增加模型的计算成本. Transformer 架构中 self-attention 模块的计算复杂度与序列长度的平方项成正比. 尽管与释义拼接方式额外引入的释义句的长度相似, 但释义拼接策略是将输入转化为一个长句子, 相比之下, 本文提出的释义知识浮动注入策略将输入转化为两个或多个较短的句子. 后者几乎不增加最大输入长度, 进而缓解了计算成本的增加. 假设释义句在组合句中的长度占比为ρ, 可以推导出在 self-attention 模块中释义知识浮动注入相对于固定的释义拼接方法的理论加速比 $\tau = 1/(2\rho^2 - 2\rho + 1)$, 该方程在 $\rho = 0.5$ 时达到极大值, 此时 τ 达到 2.

表 7 比较了在相同的释义序列长度下, 基于释义知识浮动注入以及基于固定释义拼接方式的模型计算效率. 可以看出, 浮动注入有着明显的计算效率提升, 其中 TLA 引入了最少的注入模块, 因此其具有最明显的加速效果. 与 TLA 和 ALA 相比, CLA 增大了前向传播和反向传播过程中模型的最大堆叠层数, 从而降低了加速比, 但相较于基线模型依然具有显著的性能优势.

表 7　模型运行效率对比

方法	训练加速比	测试加速比
K-BERT	1.0	1.0
TLA	1.4	1.9
ALA	1.3	1.9
CLA	1.1	1.7

3.4.5　参数分析

释义编码器层数选择. 不同的释义编码器会影响诊断模型的性能和运行效率. 尽管 Liu 等人[29]验证了预训练词汇特征在较低层的融合有利于外部知识的融合, 但由于成语释义浮动注入模型所利用的词汇特征是基于动态编码的, 因此仅使用低层编码表示会遗漏词汇特征中的高层语义信息. 为了确定最佳释义编码器层数 K, 本文在验证集上进行了超参数搜索. 图 4 展示了不同的 K 对诊断模型性能的影响. 随着释义编码器层的加深, TLA 和 ALA 呈现出性能增长的趋势, 当 $K = 12$ 时达到最佳性能, 表明深层释义编码器可以捕获成语的语义特征, 有利于与上下文语义对齐. 而 CLA 在中间层 ($K = 6$) 表现最好, 表明在跨层融合时, 高层词汇特征和低层字符特征融合产生的偏差过大可能会损伤模型整体的性能.

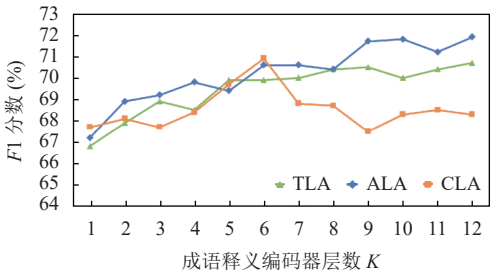


图 4　不同释义编码器层数的模型性能实验分析

编码器微调开关. 尽管不进行微调的释义编码器能够进一步加速推理过程, 但由于原始字符编码器与释义编码器在信息提取方面的需求存在差异, 因此不对其进行微调会导致诊断性能下降. 为此, 本文进行了对比实验, 旨在验证释义编码器的微调对诊断模型性能的具体影响. 实验结果表明, 在未对释义编码器进行微调的情况下, 模型在 Ran 测试集上的 $F1$ 值从 90.4% 下降至 85.9%, 减少了 4.5%; 在 Sim 测试集上则从 79.1% 下降至 72.9%, 减少了 6.2%. 这些结果强调了释义编码器微调的重要性, 证实了其对于提升模型诊断性能是不可或缺的.

3.4.6 大语言模型在成语误用诊断上的表现

大语言模型 (large language model, LLM) 已经在各种领域和任务中展现出较强的处理能力, 为此, 本文在不同的大语言模型上进行了测试, 以探究大语言模型在中文成语误用诊断任务上的表现. 本次实验选取了国内外目前成熟的大语言模型, 包括 GLM-4 (https://github.com/THUDM/GLM-4), GPT-4 (https://openai.com/index/gpt-4/), GPT-3.5-Turbo (https://platform.openai.com/docs/models/gpt-3-5-turbo), Qwen-Turbo (https://github.com/QwenLM/Qwen), Qwen-max (https://github.com/QwenLM/Qwen), ERNIE-bot-4 (https://qianfan.cloud.baidu.com/), baichuan2-13b (https://huggingface.co/baichuan-inc/Baichuan2-13B-Chat) 和 Spark v3.0 (https://xinghuo.xfyun.cn/sparkapi).

实验在高考测试集 (CEE) 上进行. 本文挑选了 3 条高考测试集之外的数据作为大语言模型的样本, 以限制其输出格式. 所有模型都使用同一套提示词, 提示词设置如下.

你是一名精通中国成语的语言学者, 现在需要你完成成语纠错任务, 任务描述如下: 你的输入是一个字符串文本, 你首先需要理解字符串文本的语义, 然后你要识别出字符串文本中所包含的成语, 然后你要根据成语的释义判断在这个字符串中成语是否被正确使用. 最后以字典形式返回判断结果. 你只需要给出最终判断结果, 不需要给出解释. 请参照下面的例子进行输入输出:
输入: 这条小河看上去浮光掠影, 旁边的树林郁郁葱葱.
输出: “浮光掠影”: false, “郁郁葱葱”: true
输入: 这家伙办事毫发不爽, 小气极了.
输出: “毫发不爽”: false
输入: 发展低碳经济首当其冲的是要坚持节约资源.
输出: “首当其冲”: false
接下来轮到你:
输入: input_text
输出:

大语言模型将返回一个字典, 包含其定位到的成语以及该成语是否被正确使用的信息, 如果定位到的成语在句子中被正确使用了, 则返回 true, 反之, 则返回 false. 在高考测试集中的每条测试数据, 若存在成语使用错误的情况, 则其标签 (label) 设置为 1, 若没有成语使用错误, 则其标签设置为 0. 如果某条数据经大语言模型测试后, 在返回的字典中包含 false, 则认为大语言模型判断该条数据存在成语误用情况. 如果某条数据经过大语言模型测试后, 返回字典中不包含 false, 则认为大语言模型判断该条数据中的成语使用均正确. 最终实验结果如表 8 所示. 实验结果表明, ERNIE-bot-4 和 Qwen 家族的 3 个大模型在 $F1$ 值上本文方法接近, 但这些模型主要表现在高召回率, 对成语是否误用的诊断精确率很低, 很难实际运用.

同时, 本文选择表现较为突出的 GLM-4、Qwen-max 和 ERNIE-Bot-4 大模型, 在提示词中加入了成语对应的释义, 探索成语释义是否能对大模型对成语误用情况判断产生正向的影响. 提示词设置如下.

你是一名精通中国成语的语言学者, 现在需要你完成成语纠错任务, 任务描述如下: 你的输入是一个字符串文本, 你首先需要理解字符串文本的语义, 然后你要识别出字符串文本中所包含的成语, 然后你要根据提供的成语的释义判断在这个字符串中成语是否被正确使用. 最后以字典形式返回判断结果. 你只需要给出最终判断结果, 不需要给出解释.

请参照下面的例子进行输入输出:
输入: 这条小河看上去浮光掠影, 旁边的树林郁郁葱葱. 释义: 浮光掠影, 像水面的光和掠过的影子一样, 一晃就消逝, 形容印象不深刻; 郁郁葱葱, 意思是形容草木苍翠茂盛, 也形容气势美好蓬勃, 生机勃勃的样子.
输出: “浮光掠影”: false, “郁郁葱葱”: true
输入: 这家伙办事毫发不爽, 小气极了. 释义: 毫发不爽, 意思是形容一点不差.
输出: “毫发不爽”: false
输入: 发展低碳经济首当其冲的是要坚持节约资源. 释义: 首当其冲, 比喻最先受到攻击或遇到灾难.
输出: “首当其冲”: false
接下来轮到你:
输入: {input_text} 释义: {explanation}
输出:

实验结果如表 9 所示, 成语释义的引入并没有提升大模型诊断成语误用的性能, ERNIE-Bot-4 及 Qwen-max 模型还出现了明显的性能下降. 考虑到目前大模型预训练语料库十分丰富, 提供成语释义信息并不能很大程度上影响大模型的行为, 相反的, 提供的成语释义可能在一定程度上对大模型产生了误导的作用, 使得其性能出现了下降. 因此, 在成语误用诊断任务中, 通过提示词提供成语释义并不能有效提升大模型的诊断性能.

表 8　大语言模型在高考测试集上的性能对比 (%)

模型	精确率	召回率	*F*1值
本文方法 (CLA)	89.2	41.5	**56.5**
ERNIE-Bot-4	38.6	96.8	55.2
Qwen-max	39.3	89.1	54.5
Qwen-Turbo	39.9	84.0	54.1
GLM-4	32.6	91.6	48.1
Spark v3.0	31.6	59.6	41.3
GPT-4	25.8	80.7	39.1
GPT-3.5-Turbo	21.0	68.7	32.2
Baichuan2-13B	34.2	16.7	22.4

表 9　成语释义对大模型成语误用诊断的影响 (%)

模型	未提供释义			提供释义		
	精确率	召回率	*F*1值	精确率	召回率	*F*1值
ERNIE-Bot-4	38.6	96.8	55.2	33.2	98.7	49.7
Qwen-max	39.3	89.1	54.5	37.6	94.9	53.8
GLM-4	32.6	91.6	48.1	33.1	98.7	49.5

表 10 展示了在高考数据集上表现出色的 4 个大模型在 5 个典型例句上的诊断能力. 其中, √表示诊断正确, ×表示诊断错误. 由于大模型在成语误用诊断任务上的低精确率表现, 所有模型在典型例句上的诊断效果都不理想, 仅有 Qwen-max 诊断正确两条, ERNIE-Bot-4 全部诊断错误. 对比之下, 本文提出的模型可以准确诊断出所有的成语误用现象. 表明大语言模型虽然在许多 NLP 任务上表现出色, 但在成语误用诊断这一特定任务上, 它们的表现并不理想. 这可能是因为大语言模型在预训练过程中主要依赖于大规模的文本数据, 而这些数据可能缺乏对成语使用准确性的细致标注. 成语的使用往往与文化背景、语境和情感色彩紧密相关, 这对于模型来说是一个挑战, 尤其是对于预训练的大语言模型, 它们在处理特定领域的知识时存在局限性. 需要将成语相关的知识库与模型结合, 增强模型对成语深层含义的理解, 并对大语言模型进行领域适应训练, 使其更好地理解和处理成语的使用.

表 10 大语言模型在典型例句上的表现

例句	GPT-4	GLM-4	Qwen-max	ERNIE-Bot-4	本文方法
浮光掠影的河面、郁郁葱葱的林园	×	√	√	×	√
这家伙办事**毫发不爽**, 小气极了	√	×	×	×	√
发展低碳经济**首当其冲**的是要坚持节约资源	×	×	×	×	√
他的玻璃屏幕和纤薄机身**美轮美奂**	×	×	√	×	√
这些年轻的科学家决心以**无所不为**的勇气	×	×	×	×	√

4 总 结

本文提出了汉语成语误用诊断研究任务, 旨在识别中文语料中存在的成语误用现象, 并针对该任务构建了一个大规模的数据集, 为成语误用诊断的研究提供了丰富的实验材料和评估基准. 本文提出了一种模型, 该模型通过浮动注入成语的释义知识, 能够在保持释义特征融合有效性的同时, 显著降低知识噪声和因模型最大输入长度限制而导致的信息丢失问题, 增强了模型对多成语和长释义等复杂场景的处理能力. 通过一系列实验充分验证了所提出模型在成语误用诊断任务上的有效性, 证明了该方法在实际应用中的潜力和价值. 在未来工作中, 将继续围绕成语所特有的使用场景, 深入探索包括情感一致性、表达对象一致性等知识在成语误用诊断任务中的融合, 以提升模型在高考测试集等高难度真实场景数据上的性能.

References:

[1] Peng H. Discussion on the type, reason and mentality implied in the wrong usage of idiom. Journal of Zhongzhou University, 2014, 31(4): 85–89 (in Chinese with English abstract). [doi: 10.13783/j.cnki.cn41-1275/g4.2014.04.018]

[2] Yang H, Juan YW. Pragmatic interpretive method of the use of idioms and the influence in idioms' evolution. Seeking Truth, 2009, 36(6): 121–127 (in Chinese with English abstract). [doi: 10.3969/j.issn.1000-7504.2009.06.023]

[3] Zhang SH, Huang HR, Liu JC, Li H. Spelling error correction with Soft-masked BERT. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 882–890. [doi: 10.18653/v1/2020.acl-main.82]

[4] Xu HD, Li ZL, Zhou QY, Li C, Wang ZZ, Cao YB, Huang HY, Mao XL. Read, listen, and see: Leveraging multimodal information helps Chinese spell checking. In: Proc. of the 2021 Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2021. 716–728. [doi: 10.18653/v1/2021.findings-acl.64]

[5] Wang SL, Wang BX, Gong JF, Wang ZY, HuX, Duan XY, Shen ZZ, Yue G, Fu RJ, Wu DY, Che WX, Wang SJ, Hu GP, Liu T. Combining ResNet and Transformer for Chinese grammatical error diagnosis. In: Proc. of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou: Association for Computational Linguistics, 2020. 36–43. [doi: 10.18653/v1/2020.nlptea-1.5]

[6] Cao YC, He L, Ridley R, Dai XY. Integrating BERT and score-based feature gates for Chinese grammatical error diagnosis. In: Proc. of the 6th Workshop on Natural Language Processing Techniques for Educational Applications. Suzhou: Association for Computational Linguistics, 2020. 49–56. [doi: 10.18653/v1/2020.nlptea-1.7]

[7] Zheng CJ, Huang ML, Sun AX. ChID: A large-scale Chinese idiom dataset for cloze test. In: Proc. of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 778–787. [doi: 10.18653/v1/P19-1075]

[8] Long SY, Wang R, Tao K, Zeng JL, Dai XY. Synonym knowledge enhanced reader for Chinese idiom reading comprehension. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 3684–3695. [doi: 10.18653/v1/2020.coling-main.329]

[9] Tan MH, Jiang J. A BERT-based dual embedding model for Chinese idiom prediction. In: Proc. of the 28th Int'l Conf. on Computational Linguistics. Barcelona: Int'l Committee on Computational Linguistics, 2020. 1312–1322. [doi: 10.18653/v1/2020.coling-main.113]

[10] Tan MH, Jiang J, Dai BT. A BERT-based two-stage model for Chinese Chengyu recommendation. Trans. on Asian and Low-resource Language Information Processing, 2021, 20(6): 92. [doi: 10.1145/3453185]

[11] Ju SG, Huang FY, Sun JP. Idiom cloze algorithm integrating with pre-trained language model. Ruan Jian Xue Bao/Journal of Software, 2022, 33(10): 3793–3805 (in Chinese with English abstract). http://www.jos.org.cn/1000-9825/6307.htm [doi: 10.13328/j.cnki.jos.

006307]
[12] Liu WJ, Zhou P, Zhao Z, Wang ZR, Ju Q, Deng HT, Wang P. K-BERT: Enabling language representation with knowledge graph. In: Proc. of the 34th AAAI Conf. on Artificial Intelligence. New York: AAAI, 2020. 2901–2908. [doi: 10.1609/aaai.v34i03.5681]
[13] Vaswani A, Shazeer N, Parmar N, Uszkoreit J, Jones L, Gomez AN, Kaiser Ł, Polosukhin I. Attention is all you need. In: Proc. of the 31st Int'l Conf. on Neural Information Processing Systems. Long Beach: Curran Associates Inc., 2017. 6000–6010.
[14] Qiu XP, Sun TX, Xu YG, Shao YF, Dai N, Huang XJ. Pre-trained models for natural language processing: A survey. Science China Technological Sciences, 2020, 63(10): 1872–1897. [doi: 10.1007/s11431-020-1647-3]
[15] Yenduri G, Ramalingam M, Chemmalar Selvi G, Supriya Y, Srivastava G, Maddikunta PKR, Deepti Raj G, Jhaveri RH, Prabadevi B, Wang WZ, Vasilakos AV, Gadekallu TR. GPT (generative pre-trained Transformer)—A comprehensive review on enabling technologies, potential applications, emerging challenges, and future directions. IEEE Access, 2024, 12: 54608–54649. [doi: 10.1109/ACCESS.2024.3389497]
[16] Devlin J, Chang MW, Lee K, Toutanova K. BERT: Pre-training of deep bidirectional Transformers for language understanding. In: Proc. of the 2019 Conf. of the North American Chapter of the Association for Computational Linguistics. Minneapolis: Association for Computational Linguistics, 2019. 4171–4186. [doi: 10.18653/v1/N19-1423]
[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:1907.11692, 2019.
[18] Lewis M, Liu YH, Goyal N, Ghazvininejad M, Mohamed A, Levy O, Stoyanov V, Zettlemoyer L. BART: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 7871–7880. [doi: 10.18653/v1/2020.acl-main.703]
[19] Sun Y, Wang SH, Li YK, Feng SK, Chen XY, Zhang H, Tian X, Zhu DX, Tian H, Wu H. ERNIE: Enhanced representation through knowledge integration. arXiv:1904.09223, 2019.
[20] Liu YH, Gu JT, Goyal N, Li X, Edunov S, Ghazvininejad M, Lewis M, Zettlemoyer L. Multilingual denoising pre-training for neural machine translation. Trans. of the Association for Computational Linguistics, 2020, 8: 726–742. [doi: 10.1162/tacl_a_00343]
[21] Radford A, Kim JW, Hallacy C, Ramesh A, Goh G, Agarwal S, Sastry G, Askell A, Mishkin P, Clark J, Krueger G, Sutskever L. Learning transferable visual models from natural language supervision. In: Proc. of the 38th Int'l Conf. on Machine Learning. PMLR, 2021. 8748–8763.
[22] Li ZY, Liu XY, Chen WJ, Shen F, Bi HR, Ke GL, Zhang LF. Uni-Fold: An open-source platform for developing protein folding models beyond alphafold. bioRxiv, 2022. [doi: 10.1101/2022.08.04.502811]
[23] Cui YM, Che WX, Liu T, Qin B, Wang SJ, Hu GP. Revisiting pre-trained models for Chinese natural language processing. In: Proc. of the 2020 Findings of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 657–668. [doi: 10.18653/v1/2020.findings-emnlp.58]
[24] Cui YM, Che WX, Liu T, Qin B, Yang ZQ. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Trans. on Audio, Speech, and Language Processing, 2021, 29: 3504–3514. [doi: 10.1109/TASLP.2021.3124365]
[25] Shao YF, Geng ZC, Liu YT, Dai JQ, Yan H, Yang F, Li Z, Bao HJ, Qiu XP. CPT: A pre-trained unbalanced Transformer for both Chinese language understanding and generation. Science China Information Sciences, 2024, 67(5): 152102. [doi: 10.1007/s11432-021-3536-5]
[26] Ma RT, Peng ML, Zhang Q, Wei ZY, Huang XJ. Simplify the usage of lexicon in Chinese NER. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 5951–5960. [doi: 10.18653/v1/2020.acl-main.528]
[27] Li XN, Yan H, Qiu XP, Huang XJ. FLAT: Chinese NER using flat-lattice Transformer. In: Proc. of the 58th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics, 2020. 6836–6842. [doi: 10.18653/v1/2020.acl-main.611]
[28] Zhang Y, Yang J. Chinese NER using lattice LSTM. In: Proc. of the 56th Annual Meeting of the Association for Computational Linguistics. Melbourne: Association for Computational Linguistics, 2018. 1554–1564. [doi: 10.18653/v1/P18-1144]
[29] Liu W, Fu XY, Zhang Y, Xiao WM. Lexicon enhanced Chinese sequence labeling using BERT adapter. In: Proc. of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th Int'l Joint Conf. on Natural Language Processing. Association for Computational Linguistics, 2021. 5847–5858. [doi: 10.18653/v1/2021.acl-long.454]
[30] Yang ZL, Dai ZH, Yang YM, Carbonell J, Salakhutdinov R, Le QV. XLNet: Generalized autoregressive pretraining for language understanding. In: Proc. of the 33rd Int'l Conf. on Neural Information Processing Systems. Vancouver: Curran Associates Inc., 2019. 517.

[31] Su JL, Murtadha A, Pan SF, Hou J, Sun J, Huang WW, Wen B, Liu YF. Global Pointer: Novel efficient span-based approach for named entity recognition. arXiv:2208.03054, 2022.

[32] Loshchilov I, Hutter F. Fixing weight decay regularization in Adam. arXiv:1711.05101v1, 2019.

附中文参考文献:

[1] 彭慧. 成语误用的类型、诱因及其文化心理分析. 中州大学学报, 2014, 31(4): 85–89. [doi: 10.13783/j.cnki.cn41-1275/g4.2014.04.018]

[2] 杨华, 隽娅玮. 实用主义解读方式及其对成语演变的影响. 求是学刊, 2009, 36(6): 121–127. [doi: 10.3969/j.issn.1000-7504.2009.06.023]

[11] 琚生根, 黄方怡, 孙界平. 融合预训练语言模型的成语完形填空算法. 软件学报, 2022, 33(10): 3793–3805. http://www.jos.org.cn/1000-9825/6307.htm [doi: 10.13328/j.cnki.jos.006307]

何亮(1982－), 男, 博士生, 主要研究领域为自然语言处理, 知识工程.

吴震(1993－), 男, 博士, 助理教授, 博士生导师, 主要研究领域为自然语言处理.

曹永昌(1998－), 男, 硕士, 主要研究领域为自然语言处理.

戴新宇(1979－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为自然语言处理, 知识工程.

黄琰琛(2002－), 男, 学士, 主要研究领域为软件工程, 自然语言处理.

陈家骏(1963－), 男, 博士, 教授, 博士生导师, CCF 专业会员, 主要研究领域为自然语言处理, 软件工程.