计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2023,32(5):157−163 [doi: 10.15888/j.cnki.csa.009099]

E-mail: csa@iscas.ac.cn
http://www.c-s-a.org.cn
Tel: +86-10-62661041

基于 RoBERTa-ND 的中文实词辨析①

孙晨瑜, 王振琦, 张宝宇, 张卫山, 侯召祥, 陈　涛

(中国石油大学 (华东) 计算机科学与技术学院, 青岛 266580)
通信作者: 孙晨瑜, E-mail: scyuige@163.com

摘　要: 在机器阅读理解任务中, 由于中文实词的组合性和隐喻性, 且缺乏有关中文实词辨析的数据集, 因此传统方法对中文实词的理解程度和辨析能力仍然有限. 为此, 构建了一个大规模 (600k) 的中文实词辨析数据集 (Chinese notional word discrimination cloze data set, CND). 在数据集中, 一句话中的一个实词被替换成了空白占位符, 需要从提供的两个候选实词中选择正确答案. 设计了一个基线模型 RoBERTa-ND (RoBERTa-based notional word discrimination model) 来对候选词进行选择. 模型首先利用预训练语言模型提取语境中的语义信息. 其次, 融合候选实词语义并通过分类任务计算候选词得分. 最后, 通过增强模型对位置及方向信息的感知, 进一步加强了模型的中文实词的辨析能力. 实验表明, 该模型在 CND 上准确率达到 90.21%, 战胜了 DUMA (87.59%), GNN-QA (84.23%) 等主流的完形填空模型. 该工作填补了中文隐喻语义理解研究的空白, 可以在提高中文对话机器人认知能力等方向开发更多实用价值. 数据集 CND 及 RoBERTa-ND 代码均已开源: https://github.com/2572926348/CND-Large-scale-Chinese-National-word-discrimination-dataset.

关键词: 隐喻语义理解; 中文实词辨析; 机器阅读理解

引用格式: 孙晨瑜,王振琦,张宝宇,张卫山,侯召祥,陈涛.基于 RoBERTa-ND 的中文实词辨析.计算机系统应用,2023,32(5):157–163. http://www.c-s-a.org.cn/1003-3254/9099.html

Chinese Notional Word Discrimination Based on RoBERTa-ND

SUN Chen-Yu, WANG Zhen-Qi, ZHANG Bao-Yu, ZHANG Wei-Shan, HOU Zhao-Xiang, CHEN Tao

(College of Computer Science and Technology, China University of Petroleum, Qingdao 266580, China)

Abstract: Chinese notional words are combinatorial and metaphorical in nature, and there is a lack of data sets on Chinese notional word discrimination. As a result, the understanding and discriminative capability of traditional methods for Chinese notional words are still limited in machine reading comprehension tasks. For this reason, a large-scale (600k) Chinese notional word discrimination cloze data set (CND) is constructed. In the dataset, a notional word in a sentence is replaced with a blank placeholder, and the correct answer needs to be selected from the two candidate notional words provided. A baseline model, RoBERTa-based notional word discrimination model (RoBERTa-ND), is designed to select candidate words. The model first extracts semantic information in the context using a pre-trained language model. Second, the semantics of candidate notional words are fused, and the scores of candidate words are computed by a classification task. Finally, the model's ability to discriminate Chinese notional words is further enhanced by enhancing the model's perception of locations and orientation information. Experiments show that the model achieves the accuracy of 90.21% on CND, beating mainstream cloze test models such as DUMA (87.59%) and GNN-QA (84.23%). This work fills the gap in the research on Chinese metaphorical semantic understanding and can develop more practical value in improving the cognitive ability of Chinese Quiz Bot. The codes of CND and RoBERTa-ND are open-source: https://github.com/

① 基金项目: 国家自然科学基金 (62072469); 中国科学院自动化研究所复杂系统管理与控制国家重点实验室 2021 年开放课题 (20210114)
收稿时间: 2022-11-03; 修改时间: 2022-12-10; 采用时间: 2022-12-23; csa 在线出版时间: 2023-03-17
CNKI 网络首发时间: 2023-03-19

2572926348/CND-Large-scale-Chinese-National-word-discrimination-dataset.

Key words: metaphorical semantic understanding; Chinese notional word discrimination; machine reading comprehension

1 引言

机器阅读理解技术能够自动处理文本数据并提取语义知识, 在智慧社区、医疗和司法等各个领域的应用越来越广泛. 作为机器阅读理解的一项重要任务, 完形填空任务旨在通过句法结构和语义语境实现空白词的填补. 对实词辨析的研究可以有效提高完形填空任务的准确性, 并改善其他任务 (如问答) 的效果. 目前已经有很多相关的数据集用于完形填空任务的研究, 包括 CNN/Daily Mail[1], children's book test (CBT)[2], ChID[3] 等.

汉语实词是汉语句子信息的主要载体, 数量庞大, 语义丰富, 在阅读理解中至关重要. 在使用汉语实词时, 要注意语义的轻重、使用范围、适用对象、文体色彩、情感色彩、搭配关系、语法功能、不同词语的意义强调等因素, 部分例子见表 1.

表 1 汉语实词辨析示例

因素	候选词A	候选词B
语义轻重	询问: 指日常提问	讯问: 采取强制措施的提问, 常用于司法质询
适用范围	局面: 指一个时期内事情的态势	场面: 指肉眼可见情景
适用对象	年龄: 用于年轻人	年纪: 多用于老年人, 如上了年纪
情感色彩	称赞: 褒义词, 多用于正面人物的赞美	奉承: 贬义词, 反面人物的有意赞美
搭配关系	爆发: 常常与具体的事物搭配使用	暴发: 与突然发生的事件搭配使用, 带有贬义
语法功能	阻碍: 常做动词, 指使不能顺利通过或发展	障碍: 用作名词, 指阻挡前进的东西
意义不同	化装: 改变装束、容貌, 假扮之意	化妆: 针对面部的修饰

然而, 关于完形填空任务的相关研究中忽略了中文实词对文本语义分析的影响, 以及缺乏相关研究的数据集. 因此, 本文提出了一个大规模的中文实词辨析数据集 (CND), 该数据集包含了 70 万个句子, 涵盖了各种不同的领域, 并通过完形填空任务进行评估. 首先, 收集并筛选了一些经典的实词辨析问题, 并根据每个问题中的选项构建了候选词语列表. 将候选词语作为关键词, 在小说、新闻语料库中对包含这些关键词的句子进行提取. 一个句子仅匹配一个关键词, 用一个空白的占位符替换该关键词. 该关键词所在候选词列表作为候选项, 用于填充空白处. 通过这种方式得到了大量的实词辨析问题, 且包含相应的候选项, 并在数据清洗后获得了高质量的数据集. 最后, 对这个数据集设计了微调任务, 提出了一个基线模型, 评估了多种预训练模型, 并对基础模型进行了改进. 实验结果表明, 模型的性能已经超越了人工, 但该任务还有一定的提升空间.

整体创新或贡献如下.

(1) 通过设计数据集和模型, 提出了一种新的中文实词研究方法.

(2) 据对该领域研究情况的调研所知, 这是第 1 个从中文实词辨析的角度来提高文本语义理解的研究, 并在中文实词辨析的任务中进行了实验.

(3) 提供了一个大规模的中文实词辨析完形填空数据集 (CND).

(4) 建立了一个中文实词辨析的基线模型 (RoBERTa-ND), 使用不同的预训练模型对所提出的方法进行了广泛的评估, 以显示其有效性.

2 研究现状

最近, 关于中文机器阅读理解能力的研究逐渐兴起. 目前, 已经发表的中文机器阅读理解的完型填空式数据集包括 ChID[3], CMRC-2017 和 People Daily & Children's Fairy Tale[4]. 这些数据集的基本信息如表 2 所示, 表中答案来源列表示答案是否来自原文. 候选项栏表示是否为查询提供了候选答案. 答案类型列表示候选词属于哪个语篇. 领域列表示语料库的来源.

表 2 数据集 (CND) 与其他已发表的数据集的比较

数据集	答案来源	候选项	答案类型	领域
People Daily & Children's Fairy Tale	是	否	名词、命名实体	新闻、儿童故事
CMRC-2017	是	否	名词、命名实体	儿童故事
ChID	否	是	成语	新闻、小说、散文
CND	否	是	中文实词	中学生的考试题目

人民日报 (PD) 从新闻文章中收集数据. CFT 和 CMRC-2017 从儿童的阅读材料中收集数据. PD 和 CFT 以及 CMRC-2017 的实验方法非常相似. 这些数据集通常用一个空白的占位符替换文档中的名词或命名实体, 并将包含该词的原始句子作为查询. 这意味着, 这些数据集直接从原始句子中提取正确答案. 两个数据集都不提供候选词.

ChID 收集的数据来自新闻、小说和散文. ChID 的答案类型是成语, 这是中文的一种独特的语言现象. ChID 为查询提供了 7 个选项. 这些选项包括 1 个最佳选项, 3 个干扰的近义词和 3 个随机获得的选项, 随机选项与正确选项的词义相关性较低. 在实验过程中, ChID 并不局限于已经设定好的方案. ChID 的候选项还将包含 1 个黄金选项和 6 个同义词选项, 或者 1 个黄金选项和 6 个普通干扰选项. ChID 分为两个数据集: 相对简单的域内数据和包含一些低频词的域外数据. 域内数据来自新闻和小说. 域外数据来自散文, 用来检查模型的泛化能力. 相比之下, 域内数据的难度要高于域外数据.

此外, 本文调研并参考了一些英语机器阅读理解的完形填空数据集. 这些数据集包括 CNN/Daily Mail, CBT, Who-did-What[5]. CNN/Daily Mail 是最早的英语机器阅读理解数据集. CNN/Daily Mail 的数据集与 PD 非常相似. CNN/Daily Mail 也从新闻文章中收集数据, 并从段落中提取数据, 不提供查询的候选项. 儿童阅读测试 (CBT) 数据集为每个查询提供了一个候选选项列表, 并删除了一些类型的词, 包括名词、动词和介词. Who-did-What 从新闻中收集语料, 并提供类似 CBT 的查询候选项. 每个问题由两篇独立的文章组成: 一篇文章是要阅读的上下文, 另一篇则是需要进行填空的问题.

根据现有的关于文学领域实词的研究成果[6,7], 当组织数据集时, 更关注那些与上下文关系更密切的实词. 这可以使任务更具挑战性, 并可以验证现有的研究成果. 数据集 CND 的风格与 ChID 数据集有些相似. 语料库中的数据来自中学试题和高质量的开源中文语料库. 对于数据集 CND 中的每个样本, 都有两个候选词. 这两个候选词在词形和字符上有很高的相似性, 但在具体语境中却有完全不同的含义.

3 研究方法

本文提出了一个新的数据集 CND, 用于中文完形填空形式的阅读理解, 从实词辨析的角度提高机器阅读理解的能力. 第 3.1 节主要介绍了数据集的构建过程: 首先介绍了如何选择候选词, 如何对候选词进行筛选和标准化. 然后根据候选词从语料库中提取文本, 并对文本进行 MASK. 最后, 对提取的文本数据进行数据清洗, 形成最终的数据集. 第 3.2 节描述了 RoBERTa-ND 模型的构建和优化过程: 该模型使用预训练模型从数据中提取特征. 从预训练模型中取出空白占位符处的状态, 使用 Embedding 层将候选词进行编码, 利用 Einsum 积运算将两者进行特征融合, 最后通过全连接层得到预测结果. 鉴于 Transformer 网络对位置信息的不敏感性, 该模型通过增加一个 LSTM 层[8] 进行了优化. 实验过程如图 1.

3.1 构建中文实词辨析数据集

表 3 是一个实词辨析任务的例子, 在每个问题中, 句子中的一个词将被一个空白占位符所取代. 每个空白处提供两个候选词, 包括正确答案. 中文的实词辨析任务是根据空白处的上下文选择正确答案. 通常情况下, 正确答案不会出现在上下文中, 这与大多数现有的完形填空数据集不同. 接下来, 将详细介绍 CND 构建的两个步骤: (1) 选择候选词. (2) 用高质量文本段落中的空白替换候选词.

表 3 实词辨析任务的例子

上下文 & 占位符	#synonym#处的候选项	
	正确选项	干扰选项
紧急关头, 政委王志明#synonym#成谈判人员, 与躲在车内的犯罪嫌疑人进行谈判.	化装 (人装扮成特定的角色, 有伪装、改变装束和外表的意思)	化妆 (使用化妆品修饰头部和脸部, 使其外观美丽)

3.1.1 候选词的选择

通过计算词嵌入的余弦相似度, 从《现代汉语词典》中收集了 3 000 对候选词. 这些候选词对在发音和字符上有很高的相似性. 由于大多数候选词由两个汉字组成, 因此只在词汇表中保留两个汉字的词. 然而, 这部分候选词对可能是同义词而不是近义词. 也就是说, 这对候选词在不同语境中所表达的意思是相同的. 因此, 对嵌入相似度和词的同义性之间的相关性进行手动评估, 当两个候选词的同义性太高时, 这对候选词将被删除. 为了便于进一步选择候选词对, 使用了 Xu[9] 提供的高质量开源中文语料库及 Sun 等人[10] 提供的新闻语料集. 为了统计每个词在语料库中的出现次数, 语料库中的一些词将被标准化. 这些词只在某个

字符上有细微的变化, 并且有相同的解释和意义 (例如, “祛寒”也叫“驱寒”), 因此, 这些词被标准化了 (“驱寒”被替换为“祛寒”), 以便正确计算某个词的出现频率.

然后, 计算出每个词在语料库中的出现频率, 并删除了那些两个词的出现频率都低于 30 次的候选词对. 最后, 筛选出 677 对候选词, 它们在语料库中的出现频率统计见表 4. 最小和最大的频率分别为 30 和 438. 此外, 还从《高考中文实词辨析集》中挑选了 100 对高质量的候选词. 这些词对都是由高水平的文学工作者筛选出来的. 它们在词形和字形上有很高的相似性, 但在具体语境中却有完全不同的含义. 最后, 该词汇表共有 777 对优质候选词.

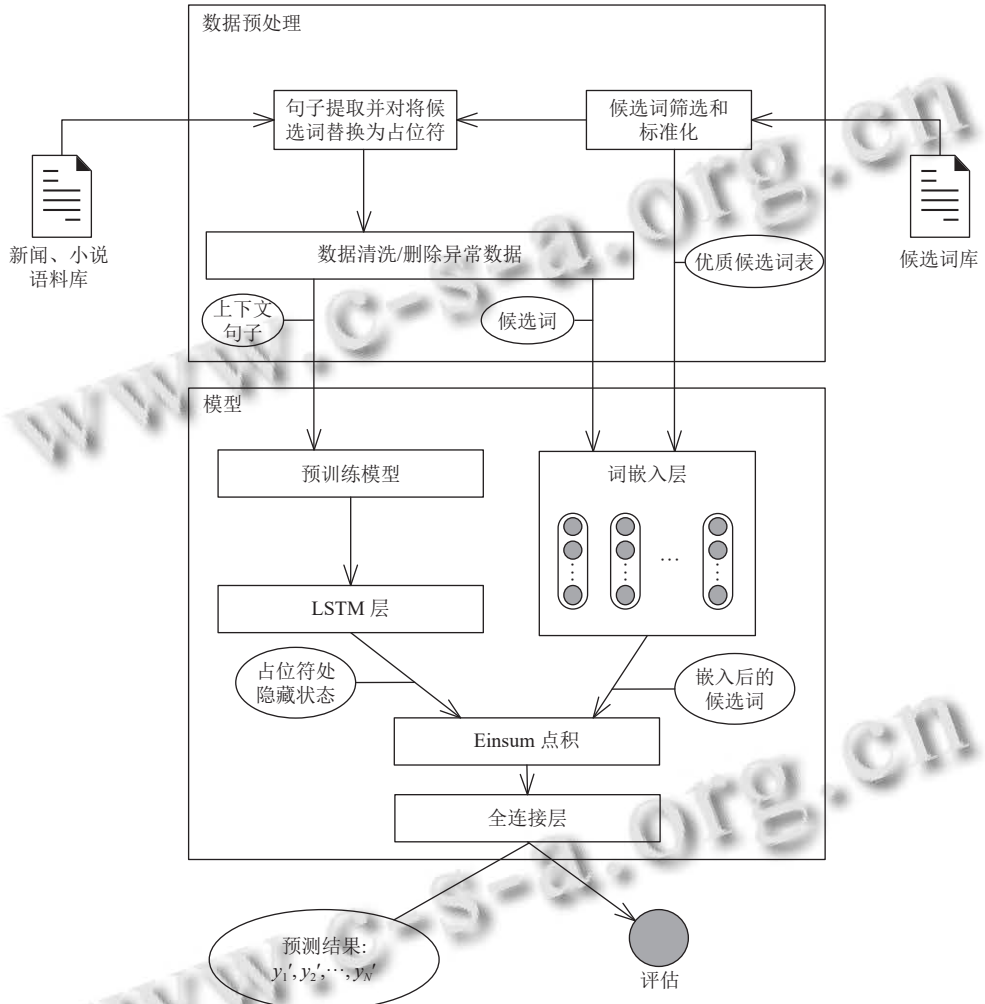

图 1 研究过程

表 4 语料库中的词频统计

出现频率	频率区间	出现次数	占比 (%)
非常低	[30, 50]	127	18.8
低	[50, 100]	133	19.7
适中	[100, 200]	142	20.9
高	[200, 300]	113	16.7
非常高	[300, 438]	162	23.9
总计	[30, 534]	677	100.0

3.1.2 问题的提取

为了使上下文的主题和领域更加多样化, 使用了现代中国小说的开源数据集和搜狗实验室提供的高质量新闻数据集. 由于一些文本可能非常长, 将句子作为基本单位来分割这些数据集. 当候选词表中的一个词出现在一个句子中时, 用一个空白符号替换该词并记录该句子. 最后, 总共获得了 7 000k 条数据. 由于一些句子的语义不完整, 短于 100 个字符的句子将被删除, 以确保最终数据的语义完整性. 同时, 删除带有特殊符号或超过 700 个字符的句子, 并将每两个句子的长度差控制在一定范围内.

值得注意的是, 如果某些词的词频远远高于其他词, 那么模型可能会倾向于选择那些出现频率较高的词. 因此, 为词汇表中的每个词在新闻数据集中提取 750 个句子, 在小说数据集中提取 750 个句子, 并将每个词的句子控制在 1 500 左右. 使经常出现的词和不经常出现的词更加平衡. 最后, 在数据集中还剩下 600k 高质量的数据. 此外, 在早期的实验中, 实验结果表明如果正确答案集中在第 1 个候选词上, 那么模型会倾向于选择第 1 个候选词作为正确答案, 使得模型的泛化效果变差. 因此, 对候选词顺序进行了随机打乱, 以进一步增加选词任务的难度.

3.2 模型构建

RoBERTa-ND 模型包含一个预训练模型层, 一个长短期记忆 (LSTM) 层, 一个嵌入层, 一个 Dropout 层和一个全连接层.

文学领域中汉语实词的研究[11] 充分证明了语境对实词意义的重要影响. 为了提高模型的预测精度, 实验决定在模型中记录这些实词的语境信息. 因此, 将数据清洗后的上下文数据传递给预训练模型层. 该层对上下文信息进行编码, 并记录这些名义词的上下文信息.

在实验过程中, 选择的预训练模型采用了基于自注意力的 Transformer 编码器. 自注意机制在编码过程中削弱了位置和方向信息. 借助于直观的经验, 在实词辨析的任务中, 位置和方向信息对候选词的选择有一定影响[12]. LSTM 是一个有记忆能力的循环神经网络. 它可以保存位置和方向等信息供后续网络层使用, 也可以防止梯度的消失. 因此, 考虑在预训练模型层的基础上使用双向 LSTM 层来进一步感知序列的位置依赖关系.

除了处理上下文文本, 对候选词进行了处理. 首先, 模型通过词典获得两个候选词的索引. 其次, 嵌入层将候选词的索引转换为词向量. 该层的操作保留了候选词的关键信息, 并通过合理地组织词向量的形式提高了存储空间的利用率. 接下来, 使用 Einsum 点积运算[13], 将从预训练网络中提取的占位符处的状态与编码后的候选词进行融合, 该模型参考点积运算的分数来决定最终的单词选择.

此外, 在模型中加入了一个 Dropout 层. 该层随机选择神经网络中的某些节点进行抑制, 防止模型训练过程中的过度拟合, 并提高模型的泛化能力. 最后在模型中加入了一个线性层, 以降低张量的维度, 决定最终的单词选择, 用于最终的预测和评估.

模型结构如图 2 所示. 训练一个基于预训练模型的分类网络来获得最后的隐藏状态, 并使用最后的隐藏状态对候选词进行评分.

$$W_{\text{last}} = Model(S) \tag{1}$$

$$c_i = embedding(W_i) \tag{2}$$

$$\alpha_i = Softmax_i(Linear(W_{\text{last}}^{\text{T}} \cdot c_i)) \tag{3}$$

其中, W_{last} 表示预训练模型最后一层的隐藏状态, S 表示句子, c_i 表示每一个候选词的嵌入表示, W_i 表示候选词. 将预训练模型后 4 层的隐藏状态取出, 并输入到 LSTM 网络中, 获得空白占位符处的上下文信息.

$$W_i = Model(S) \tag{4}$$

$$W_{\text{all}} = [W_9, W_{10}, W_{11}, W_{12}] \tag{5}$$

$$o = LSTM(W_{\text{all}}) \tag{6}$$

$$c_i = embedding(W_i) \tag{7}$$

$$\alpha_i = Softmax_i(Linear(o \cdot c_i)) \tag{8}$$

其中, W_i 表示预训练模型中第 i 层的隐藏状态, W_{all} 表示将后 4 层的隐藏状态拼接, o 表示 LSTM 层的输出. 将 o 与 c_i 进行积操作, 经过全连接层降维, 最终通过 *Softmax* 得到最终得分.

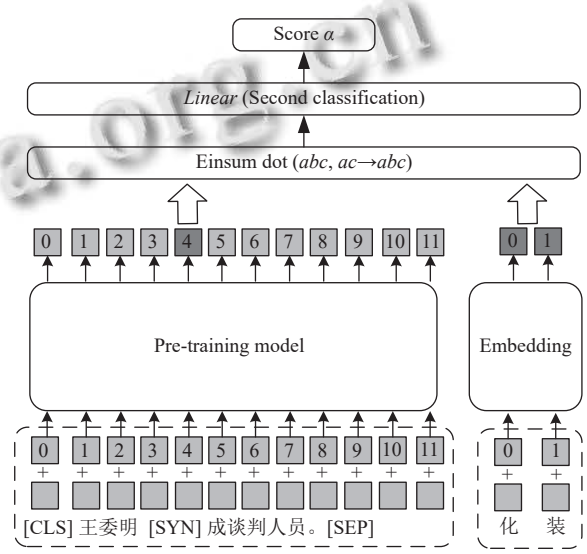

图 2 模型结构示意图

4 实验评估

4.1 评估过程和评价指标

为了评估最先进的预训练模型在中文实词辨析任

务中的表现, 在大规模中文实词辨析完形填空数据集 CND 中评估了 BERT(wwm)[14], RoBERTa[15], ALBERT[16] 的效果. 使用准确度 (ACC) 来评价实验结果.

4.2 超参数

将句子的最大长度设置为 128, 并在嵌入层使用 768 维的单词嵌入. LSTM 的隐藏单元的数量被设置为 768. 在词嵌入上应用 0.5 的丢弃率. 计算训练损失的函数是交叉熵损失函数. 使用 ADAM[17] 作为模型优化器. 所有模型的初始学习率为 2E–5. 将批次大小设置为 16, 当验证集的准确率稳定后停止训练. 训练集的数量为 500k, 验证集和测试集的数量为 50k.

4.3 实验结果

从以下几个方面对实验结果进行分析.

预训练模型的比较: 实验结果见表 5, 模型准确率情况见图 3. 首先, RoBERTa 优于其他所有模型, 其原因可能是: RoBERTa 的预训练模型有更多的训练数据 (160G), 更大的批处理量 (8k) 和更长的训练时间 (500k 轮). 其次, BERT(wwm) 的训练集和测试集之间的准确度差距最小, 为 5.62%, ALBERT 的差距最大, 为 7.78%. 这说明 BERT(wwm) 具有更强的中文理解能力.

表 5 不同预训练模型的性能 (%)

模型	训练集	验证集	测试集
BERT(wwm)	95.31	89.78	89.69
RoBERTa	**96.25**	**90.05**	**90.17**
ALBERT	72.72	64.94	64.52

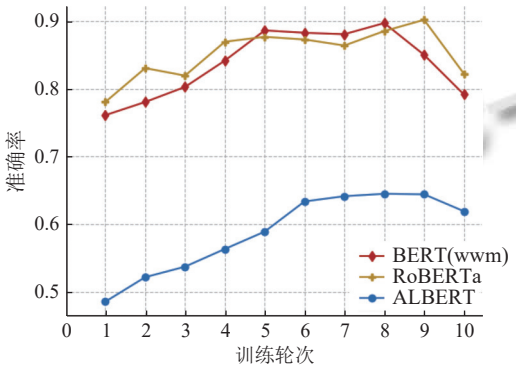

图 3 预训练模型准确率比较

结构消融实验: 实验结果见表 6, 模型准确率情况见图 4. 在模型中加入 LSTM, 从表 6 中可以得出, ALBERT 在测试集上取得了 3.35% 的最大准确率提升, 其他模型在 CND 上的表现在加入 LSTM 层后准确率也得到了提升. 这表明增加 LSTM 层对实词的辨析是有效的.

表 6 使用不同预训练模型+LSTM 的性能比较 (%)

模型	训练集	验证集	测试集
BERT(wwm)+LSTM	95.51	89.96	89.89
RoBERTa+LSTM	**96.42**	**90.24**	**90.21**
ALBERT+LSTM	75.43	67.80	67.88

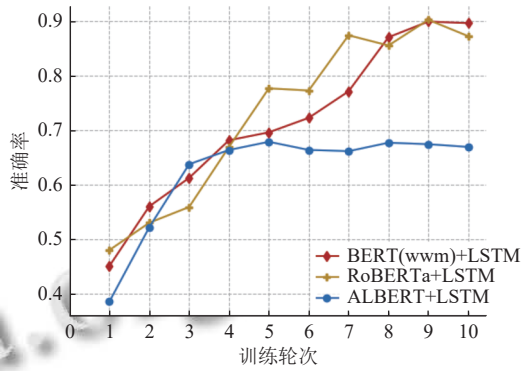

图 4 消融实验准确率比较

对比实验: 将效果最好的 RoBERTa+LSTM 作为 RoBERTa-ND 的编码器与 Contextual LSTMS[2], attention sum reader (AS reader)[18], Chinese attention sum reader (CAS reader)[4], DUMA[19], GNN-QA[20] 分别在 CND 数据集上进行实验, 同时雇佣 15 名研究生完成测试集中的部分题目, 以验证数据集的难度. 实验结果如表 7, 模型准确率情况如图 5. 实验结果表明, 模型的表现远超人工, 且在与主流的完形填空模型的比较中仍取得了富有竞争力的效果.

表 7 对比实验结果 (%)

模型	训练集	验证集	测试集
RoBERTa-ND	96.42	**90.24**	**90.21**
BIDAF	74.21	70.77	70.12
AS reader	84.92	83.83	83.71
CAS reader	**96.62**	88.38	88.06
DUMA	92.22	87.32	87.59
GNN-QA	85.92	84.01	84.23
Human	—	—	64.50

5 结论

在本文中, 提出了一种从加强实词辨析能力的角度来提高机器阅读理解能力的方法. 一方面, 提出了一个大规模的中文实词辨析数据集 CND, 并以完形填空任务作为机器阅读理解的例子. 另一方面, 通过提高实词的表征能力来加强机器对隐喻意义的理解, 并提出了中文实词辨析基线模型 RoBERTa-ND, 从而提高中文实词辨析的准确性. 实验证明, 该方法在 CND 数据集上取得了 90.21% 的测试集准确率, 取得了很好的效果.

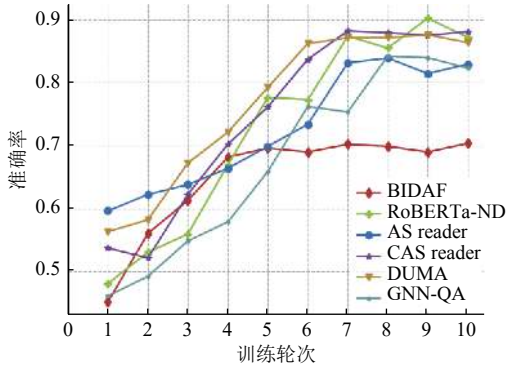

图 5 对比实验准确率比较

由于中文实词的隐喻性, 一字之差就可能导致表述了完全不同的含义. 在今后的工作中, 将继续丰富数据集 CND, 提高任务难度, 进一步提高实词的表征能力. 并探究更有效的语义融合方法, 尝试引入更多外部的特征信息, 提高模型的辨析能力.

参考文献

1 Cui YM, Liu T, Chen ZP, *et al*. Dataset for the first evaluation on Chinese machine reading comprehension. Proceedings of the 11th International Conference on Language Resources and Evaluation. Miyazaki: European Language Resources Association, 2018.

2 Hill F, Bordes A, Chopra S, *et al*. The Goldilocks principle: Reading children's books with explicit memory representations. Proceedings of the 4th International Conference on Learning Representations. San Juan: ICLR, 2016.

3 Zheng CJ, Huang ML, Sun AX. ChID: A large-scale Chinese idiom dataset for cloze test. Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 778–787.

4 Cui YM, Liu T, Chen ZP, *et al*. Consensus attention-based neural networks for Chinese reading comprehension. Proceedings of the 26th International Conference on Computational Linguistics: Technical Papers. Osaka: The COLING 2016 Organizing Committee, 2016. 1777–1786.

5 Onishi T, Wang H, Bansal M, *et al*. Who did what: A large-scale person-centered cloze dataset. Proceedings of 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 2230–2235.

6 张占山. "抱怨"和"埋怨"辨析与词典释义. 辞书研究, 2006, (3): 46–53. [doi: 10.3969/j.issn.1000-6125.2006.03.008]

7 张占山. 语义角色视角下的谓词同义词辨析 [博士学位论文]. 厦门: 厦门大学, 2006.

8 Graves A. Long short-term memory. Supervised Sequence Labelling with Recurrent Neural Networks. Berlin: Springer, 2012. 37–45.

9 Xu B. NLP Chinese corpus: Large scale Chinese corpus for NLP. https://github.com/safin1120/nlp_chinese_corpus. [2022-10-24].

10 Sun M, Li J, Guo Z, *et al*. THUCTC: An efficient Chinese text classifier. GitHub Repository. http://thuctc.thunlp.org/. (2016-01-25)[2022-10-24].

11 张占山. 『陆续』与『连续』的区别及词典释义. 辞书研究, 2006, (1): 68–77. [doi: 10.3969/j.issn.1000-6125.2006.01.010]

12 丁美荣, 刘鸿业, 徐马一, 等. 面向机器阅读理解的多任务层次微调模型. 计算机系统应用, 2022, 31(3): 212–219. [doi: 10.15888/j.cnki.csa.008417]

13 盛艺暄, 兰曼. 利用外部知识辅助和多步推理的选择题型机器阅读理解模型. 计算机系统应用, 2020, 29(4): 1–9. [doi: 10.15888/j.cnki.csa.007327]

14 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 29: 3504–3514. [doi: 10.1109/TASLP.2021.3124365]

15 Liu YH, Ott M, Goyal N, *et al*. RoBERTa: A robustly optimized BERT pretraining approach. arXiv:1907.11692, 2019.

16 Lan ZZ, Chen MS, Goodman S, *et al*. ALBERT: A lite BERT for self-supervised learning of language representations. Proceedings of the 8th International Conference on Learning Representations. Addis Ababa: OpenReview.net, 2020.

17 Kingma DP, Ba J. Adam: A method for stochastic optimization. Proceedings of the 3rd International Conference on Learning Representations. San Diego: ICLR, 2015.

18 Kadlec R, Schmid M, Bajgar O, *et al*. Text understanding with the attention sum reader network. Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics. Berlin: Association for Computational Linguistics, 2016. 908–918.

19 Zhu PF, Zhang ZS, Zhao H, *et al*. DUMA: Reading comprehension with transposition thinking. IEEE/ACM Transactions on Audio, Speech, and Language Processing, 2021, 30: 269–279.

20 Wang K, Zhang YY, Yang DY, *et al*. GNN is a counter? Revisiting GNN for question answering. Proceedings of the 10th International Conference on Learning Representations. OpenReview.net, 2021.

(校对责编: 孙君艳)