计算机系统应用 ISSN 1003-3254, CODEN CSAOBN
Computer Systems & Applications,2022,31(9):352−359 [doi: 10.15888/j.cnki.csa.008717]

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

显式融合词法和句法特征的抽取式机器阅读理解模型①

闫维宏[1,2], 李少博[2], 单丽莉[2], 孙承杰[2], 刘秉权[2]

[1](人民网 传播内容认知国家重点实验室, 北京 100733)
[2](哈尔滨工业大学 计算学部, 哈尔滨 150006)
通信作者: 刘秉权, E-mail: liubq@hit.edu.cn

摘　要: 预训练语言模型虽然能够为每个词提供优良的上下文表示特征, 但却无法显式地给出词法和句法特征, 而这些特征往往是理解整体语义的基础. 鉴于此, 本文通过显式地引入词法和句法特征, 探究其对于预训练模型阅读理解能力的影响. 首先, 本文选用了词性标注和命名实体识别来提供词法特征, 使用依存分析来提供句法特征, 将二者与预训练模型输出的上下文表示相融合. 随后, 我们设计了基于注意力机制的自适应特征融合方法来融合不同类型特征. 在抽取式机器阅读理解数据集 CMRC2018 上的实验表明, 本文方法以极低的算力成本, 利用显式引入的词法和句法等语言特征帮助模型在 *F*1 和 EM 指标上分别取得 0.37% 和 1.56% 的提升.

关键词: 机器阅读理解; 词法特征; 句法特征; 深度学习; 预训练模型; 特征融合; 注意力机制

引用格式: 闫维宏,李少博,单丽莉,孙承杰,刘秉权.显式融合词法和句法特征的抽取式机器阅读理解模型.计算机系统应用,2022,31(9):352–359. http://www.c-s-a.org.cn/1003-3254/8717.html

Extractive Machine Reading Comprehension Model with Explicitly Fused Lexical and Syntactic Features

YAN Wei-Hong[1,2], LI Shao-Bo[2], SHAN Li-Li[2], SUN Cheng-Jie[2], LIU Bing-Quan[2]

[1](State Key Laboratory of Communication Content Cognition, People's Daily Online, Beijing 100733, China)
[2](Faculty of Computing, Harbin Institute of Technology, Harbin 150006, China)

Abstract: Language models obtained by pre-training unstructured text alone can provide excellent contextual representation features for each word, but cannot explicitly provide lexical and syntactic features, which are often the basis for understanding overall semantics. In this study, we investigate the impact of lexical and syntactic features on the reading comprehension ability of pre-trained models by introducing them explicitly. First, we utilize part of speech tagging and named entity recognition to provide lexical features and dependency parsing to provide syntactic features. These features are integrated with the contextual representation from the pre-trained model output. Then, we design an adaptive feature fusion method based on the attention mechanism to fuse different types of features. Experiments on the extractive machine reading comprehension dataset CMRC2018 show that our approach helps the model achieve 0.37% and 1.56% improvement in *F*1 and EM scores, respectively, by using explicitly introduced lexical and syntactic features at a very low computational cost.

Key words: machine reading comprehension; lexical features; syntactic features; deep learning; pre-trained models; feature fusion; attention mechanism

① 基金项目: 国家自然科学基金 (62176074)
收稿时间: 2021-12-23; 修改时间: 2022-01-24; 采用时间: 2022-02-18; csa 在线出版时间: 2022-06-16

对于机器而言, 自动地阅读并理解文本是一项颇具挑战的任务, 它需要机器能够依照现实世界中的事实和常识来剖析自然语言所表述的内容[1]. 机器阅读理解 (machine reading comprehension, MRC) 旨在以问答的形式来理解文章, 其输入是自然语言形式的问题, 以及包含了能够支撑该问题回答的证据文章, 输出则是问题对应的答案. 抽取式机器阅读理解规定问题的正确答案会以文本片段的形式出现在输入文章中, 要求在文章中“抽取”出正确的答案片段.

在以 GPT[2] 和 BERT[3] 为代表的大规模预训练语言模型出现之前, 该类任务通常的解决方法是通过循环神经网络对输入问题和文章进行编码并交互, 其模型结构主要包括 4 个部分, 分别是嵌入层、编码层、交互层和输出层. 问题和文章以词序列的形式分别输入到模型中, 嵌入层首先将问题和文章的输入序列转换为词向量序列, 编码层对词向量序列进行编码后得到自然语言序列对应的上下文编码, 交互层负责将问题和文章的上下文编码进行交互, 加强问题和文章之间的相互感知, 最后由输出层计算出答案片段在文章中的具体位置. 在这一经典结构的基础上, 许多工作对其进行了修改. Attentive Reader[4] 将细粒度注意力机制应用到模型中来加强交互层的理解能力. Match-LSTM[5] 将 Match-LSTM 以及 Answer Pointer 模型相结合, 将 Pointer Net 中指针的思想首次应用于阅读理解任务. BiDAF[6] 通过双流注意力机制来提高问题与文章的交互能力. QA-NET[7] 利用自注意力机制和 CNN[8] 来进行文本的编码, 相比于 RNN[9], 其并行运算能力提高了训练速度, 也取得了当时在 SQuAD[1] 数据集上的最优预测精度.

尽管上述各类方法使得机器阅读理解模型性能逐渐提高, 但是这些模型仅仅使用固定的检索表 (look-up table) 映射得到词编码的方式具有一些无法避免的缺陷, 例如无法解决一词多义等问题[10]. 而 BERT[3] 等预训练模型引入动态编码的方式, 利用大规模语料来获取更深层且更加匹配上下文的语义表征, 极大地提高了各类模型的性能, 在机器阅读理解数据集 SQuAD 1.0[1] 和 SQuAD 2.0[11] 上的表现甚至超越了人类. BERT 做到如此出色的性能提高引起很多相关领域研究者的兴趣, Jawahar 等人[12] 通过探测任务挖掘 BERT 中的语言学信息, 实验表明 BERT 的低层网络学习到了短语级别的信息表征, 中层网络学习到了丰富的语言学特征, 而高层则学习到了丰富的语义信息特征. 而针对阅读理解任务, Si 等人[13] 的工作表明对 BERT 的微调主要学习到文本中的关键词如何引导模型进行正确的预测, 而非学习语义理解和推理. Albilali 等人[14] 则通过对抗样例表明基于预训练的语言模型仅仅依靠表面的线索, 如词汇重叠或实体类型匹配, 就能获得有竞争力的性能; 同时, 预测的错误可以由 BERT 的低层网络所识别. Aken 等人[15] 的工作则从 BERT 不同编码层的粒度揭示了 BERT 回答问题的过程, 作者将问答模型由低层至高层的输出分别表示为语义聚类, 聚类后语义与问题中相关实体的链接, 对于支持问题事实的抽取以及答案片段抽取 4 个阶段, 并将该过程与人类阅读理解的过程进行了类比.

目前的工作大都关注于为什么 BERT 的内部表征能够如此有效地完成机器阅读理解任务, 对显式地在 BERT 引入额外的特征的研究则较少. 类似工作是 SemBERT[16], 该模型通过将 BERT 输出的上下文特征与语义角色特征相拼接, 显式地利用这两种特征对来抽取答案片段, 在 SQuAD 2.0 数据集上, SemBERT 取得了优于原始 BERT 模型的表现.

受此启发, 语义角色之外其他的词法或句法特征同样值得我们关注. 人类在理解文本的过程中是先验地知道某些词法或者文法特征的, 例如, CMRC2018 中的问题“前秦对前燕发动的灭国战争是谁主导的?”中, 我们可以通过问题中的疑问代词“是谁”, 推断出问题的答案是“人名”, 从而更加关注文章中命名实体特征为“人名”的文本片段“慕容垂”. 而对于类似这样的词法、句法特征, 人类同样具有对其理解的能力, 但是在当前的主流模型中并未体现. 为了填补这部分工作的缺失, 我们提出融合词法和句法特征的抽取式机器阅读理解模型. 我们的主要工作如下:

(1) 在 BERT 输出的上下文表示的基础上, 显式地引入多词法和句法特征, 来探究这些特征是否能够在 BERT 预训练语言模型所提供上下文特征的基础上, 进一步增强机器阅读理解的性能. 其中词法特征包括命名实体特征和词性特征, 句法特征则包括依存分析特征.

(2) 设计基于注意力机制的自适应特征选择方法对各类特征进行融合, 并探究不同文本特征对 BERT 模型的影响.

(3) 在公开数据集 CMRC2018 上, 与基准模型进行

对比, 本文所提出的显式融合词法和句法特征的抽取式机器阅读理解模型在 $F1$ 和 EM 指标上分别取得了 0.37% 和 1.56% 的提升.

1 显式融合词法和句法特征的抽取式机器阅读理解模型

在本节中, 我们首先对本文方法进行概述, 随后对基于 BERT 的抽取式阅读理解模型进行详细介绍, 并阐述我们使用到的词法句法特征, 最后描述各类特征融合的动态融合方法, 并得到最终的输出.

1.1 概述

抽取式机器阅读理解可以形式化地定义为: 给定一个包含m个字符的问题$q=(q_1,q_2,\cdots,q_m)$, 一个包含n个字符的文章$p=(p_1,p_2,\cdots,p_n)$以及一个包含l个字符的答案$a=(a_1,a_2,\cdots,a_l)$, 其为p中的一个子序列. 我们的目标是学习一个机器阅读理解模型f, 来根据输入文章p和问题q得到输出答案a, 如式 (1) 所示:

$$f(p,q)\rightarrow a \tag{1}$$

本文工作的结构如图 1 所示, 我们利用已有的模型对输入数据进行预处理, 得到文本特征. 我们的问答模型首先使用 BERT 对问题和文章进行编码得到编码后的输出H^{C}. 接下来, 将H^{C}与表征为向量的文本特征通过自适应的注意力机制进行融合, 再使用多层 Transformer encoder 进行编码, 得到融合特征的编码表示H^{F}. 将二者通过自适应的注意力特征融合层, 得到我们最终的输出H, 并利用一个答案位置分类器得到最终的答案开始位置得分和结束位置得分.

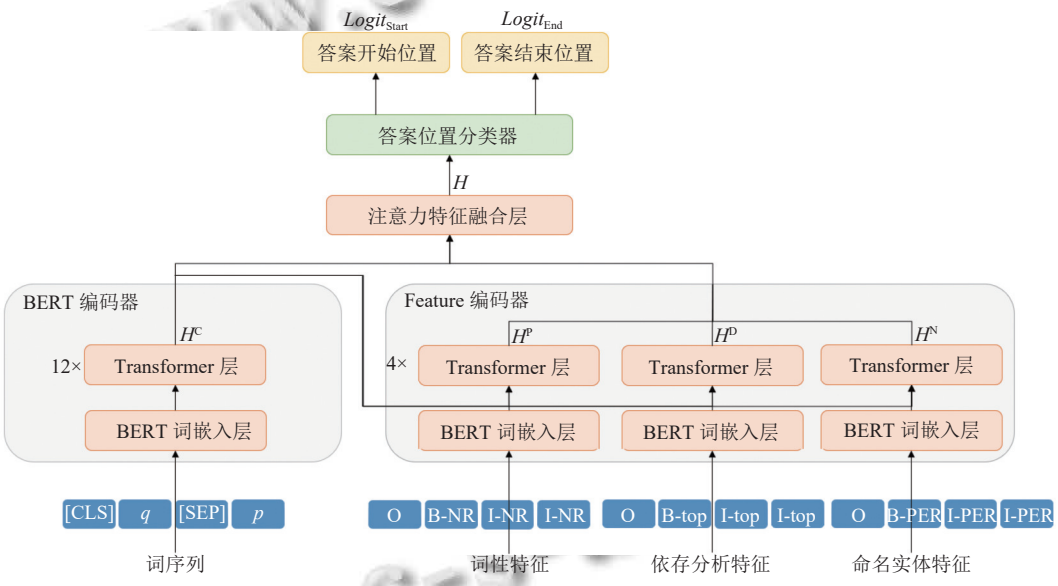

图 1 融合词法和句法特征的抽取式机器阅读理解模型结构

1.2 基于 BERT 的抽取式阅读理解模型

本文使用 BERT 作为基准模型来解决阅读理解问题. 本文任务输入的问题和文章是一串字符, 神经网络无法直接处理这样的数据. 在预训练模型出现前通常的做法是使用静态词向量将不同的词映射为对应的高维向量, 例如基于局部上下文窗口编码单词的 Word2Vec[17] 和引入全局统计信息的 GloVe[18]. 而 BERT则使用基于注意力机制的 Transformer encoder, 利用大规模语料通过其强大编码能力将文本编码为具有上下文信息的文本向量.

针对本文所涉及的问答任务这类的上下句任务, BERT 通常会将分字后的问题的词序列和文章的词序列连接起来, 输入序列如式 (2) 所示:

$$X=[\mathrm{CLS}],q_1,\cdots,q_m,[\mathrm{SEP}],p_1,\cdots,p_n,[\mathrm{SEP}] \tag{2}$$

其中, q_i表示问题的词序列中第i个字符, p_i表示文章的词序列中第i个字符; [CLS]和[SEP]为 BERT 中定义的特殊标记, [CLS]表示序列开始, [SEP]则是分隔标记, 用来分隔问题和文章以及标识输入序列的结束. 随后, 我们利用 BERT 的 Embedding 层将输入序列分别映射为词向量 (token embedding)、类型向量 (segment embedding) 和位置向量 (position embedding), 将三者相加即为 BERT 的最终输入特征. 接着通过多层 Transformer encoder 进行编码, 进而获得问题与文章交互的向量表示H^{C}:

$$H^{\mathrm{C}}=\left\{h_1^{\mathrm{C}},h_2^{\mathrm{C}},\cdots,h_l^{\mathrm{C}}\right\}=BERT(Emb_{\mathrm{BERT}}(X)),h_i^{\mathrm{C}}\in\mathbb{R}^d \quad (3)$$

其中, l表示输入序列长度, d表示 BERT 输出的每个词对应的向量的维度, 在本文中$d=768$, h_i^{C}表示经由 BERT 编码后的第i个词对应的上下文表征.

接下来, 通常的做法是使用一个全连接层作为分类器得到答案开始位置和结束位置的得分向量, 如式 (4) 和式 (5):

$$LogitBERT_{\mathrm{Start}}=Linear\left(H^{C}\right) \quad (4)$$

$$LogitBERT_{\mathrm{End}}=Linear\left(H^{C}\right) \quad (5)$$

在后续的特征融合模块中, 我们将利用 BERT 输出的上下文向量H^{C}与词法和句法特征进行融合.

1.3 词法与句法特征

当前的数据集仅仅包括文本形式的问题和文章, 并未包含所需的额外词法和句法特征, 为了获取额外特征, 我们利用现有模型进行标注, 并将这些特征进行组合, 其中词法特征包括词性特征和命名实体特征, 句法特征包括依存分析特征. 为了使得我们的文本特征阅读模型尽可能地与 BERT 阅读模型在输入层的分布相同, 我们以单字粒度进行分词, 使得各个特征构建的向量与 BERT 预训练模型最大长度相同, 以便直接进行拼接. 文本的特征示例如图 2 所示.

依存分析	O	B-top	I-top	I-top	B-cop	B-det	I-det	B-dep	I-dep	B-pass	B-root	I-root	B-dobj	I-dobj	B-dep	B-pun
命名实体	O	B-PER	I-PER	I-PER	O	O	O	O	O	O	O	O	O	O	O	O
词性	O	B-NR	I-NR	I-NR	B-VC	B-DT	I-DT	B-NN	I-NN	B-SB	B-VV	I-VV	B-NN	B-NN	B-SP	B-PU
字	[CLS]	范	延	颂	是	什	么	时	候	被	任	为	主	教	的	?
词	[CLS]	范廷颂			是	什么		时候		被	任为		主教		的	?

图 2 文本特征标注示例

词性 (part of speech, POS) 特征: 我们使用词性标注的 CTB 规范[19], 包括 37 个词性标签. 以单字切分文本后, 使用 BIO 规则对特征进行重构, 即某个词w的词性为P, 按字切分后为$\{z_1,z_2,\cdots,z_n\}$, 我们将其标注为$\{\text{B-P},\text{I-P},\cdots,\text{I-P}\}$. 对于 BERT 中的 3 种特殊标签[CLS]、[SEP]和[UNK], 我们标记为 O. 共计 75 种标签, 我们将其转换为 75 维的 one-hot 向量.

命名实体 (named entity, NE) 特征: 我们使用 MSRA 的命名实体标注规范, 该规范源于中文文本标注规范 (5.0 版), 其中包括专有名词 (NAMEX)、时间表达式 (TIMEX)、数字表达式 (NUMEX)、度量表达式 (MEASUREX) 和地址表达式 (ADDREX) 五大类及其下属的 31 个子类. 我们同样使用 BIO 规则进行标注, 并将其转换为 63 维的 one-hot 向量.

依存分析 (dependency parse, DEP) 特征: 该特征用来表示句法结构中各项之间的依赖关系[20], 共 44 项. 我们同样使用 BIO 规则进行标注, 并将其转换为 89 维的 one-hot 向量.

1.4 特征融合模块

在处理这些特征标签时, 我们需要将其转换为向量的形式. 首先, 我们对词性、命名实体和依存标签分别通过一个嵌入层映射为固定维度的向量, 并分别将这些特征与上下文特征H^{C}通过相加的方式融合, 从而将不同的特征融入上下文表示, 见图 3. 接着我们使用单个浅层的特征编码器对特征向量进行编码, 该编码器同样是 Transformer encoder.

编码后我们便得到了词性特征的向量表示H^{P} (POS), 命名实体识别特征的向量表示H^{N} (NE), 以及依存分析特征的向量表示H^{D} (DEP), 编码过程可以如式 (6)–式 (8) 所示:

$$\begin{aligned}H^{\mathrm{P}}&=\left\{h_1^{\mathrm{P}},h_2^{\mathrm{P}},\cdots,h_l^{\mathrm{P}}\right\}\\&=Transformer_{\mathrm{POS}}(Emb_{\mathrm{POS}}(POS(X))+H^{\mathrm{C}}),h_i^{\mathrm{P}}\in\mathbb{R}^d\end{aligned} \quad (6)$$

$$\begin{aligned}H^{\mathrm{N}}&=\left\{h_1^{\mathrm{N}},h_2^{\mathrm{N}},\cdots,h_l^{\mathrm{N}}\right\}\\&=Transformer_{\mathrm{NE}}(Emb_{\mathrm{NE}}(NE(X))+H^{\mathrm{C}}),h_i^{\mathrm{N}}\in\mathbb{R}^d\end{aligned} \quad (7)$$

$$\begin{aligned}H^{\mathrm{D}}&=\left\{h_1^{\mathrm{D}},h_2^{\mathrm{D}},\cdots,h_l^{\mathrm{D}}\right\}\\&=Transformer_{\mathrm{DEP}}(Emb_{\mathrm{DEP}}(DEP(X))+H^{\mathrm{C}}),h_i^{\mathrm{D}}\in\mathbb{R}^d\end{aligned} \quad (8)$$

其中, Emb和$Transformer$分别表示嵌入层和编码层. 每个词x_i最终对应的融合特征h_i则通过对 3 种不同层次的特征进行加权求和得到, 注意力权重使用双线性注意力机制[21] 得到, 如式 (9)–式 (11):

$$\begin{cases} e_i^{\mathrm{C}} = h_i^{\mathrm{C}} W^{\mathrm{CC}} h_i^{\mathrm{C}} \\ e_i^{\mathrm{N}} = h_i^{\mathrm{C}} W^{\mathrm{CN}} h_i^{\mathrm{N}} \\ e_i^{\mathrm{P}} = h_i^{\mathrm{C}} W^{\mathrm{CP}} h_i^{\mathrm{P}} \\ e_i^{\mathrm{D}} = h_i^{\mathrm{C}} W^{\mathrm{CD}} h_i^{\mathrm{D}} \end{cases} \tag{9}$$

$$\left\{a_i^{\mathrm{C}}, a_i^{\mathrm{N}}, a_i^{\mathrm{P}}, a_i^{\mathrm{D}}\right\} = Softmax\left(\left\{e_i^{\mathrm{C}}, e_i^{\mathrm{N}}, e_i^{\mathrm{P}}, e_i^{\mathrm{D}}\right\}\right) \tag{10}$$

$$h_i = a_i^{\mathrm{C}} h_i^{\mathrm{C}} + a_i^{\mathrm{P}} h_i^{\mathrm{P}} + a_i^{\mathrm{N}} h_i^{\mathrm{N}} + a_i^{\mathrm{D}} h_i^{\mathrm{D}} \tag{11}$$

其中, BERT 输出上下文相关特征向量h_i^{C}对应的权重为a_i^{C}, 词性特征向量h_i^{P}对应的权重为a_i^{P}, 命名实体特征向量h_i^{N}对应的权重为a_i^{N}, 依存分析特征向量h_i^{D}对应的权重为a_i^{D}.

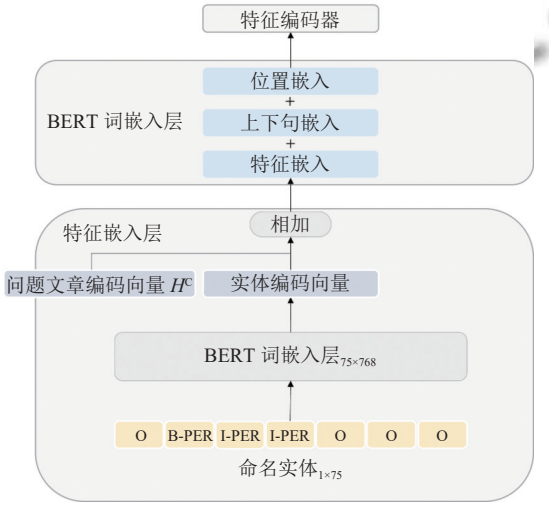

图 3 实体特征输入模块

根据上下文编码对各个特征的注意力权重, 将所有特征进行加权融合, 得到最终的融合了词法与句法特征的问题与文章交互向量表示H, 如式 (12)–式 (15), 我们使用全连接层对H进行二分类, 分别得到每个词作为答案开始位置和结束位置的概率, 并使用 *Softmax* 得到归一化后的起止位置的最终得分.

$$Logit_{\mathrm{Start}} = Linear(H) \tag{12}$$

$$Logit_{\mathrm{End}} = Linear(H) \tag{13}$$

$$Score_{\mathrm{Start}} = Softmax(Logit_{\mathrm{Start}}) \tag{14}$$

$$Score_{\mathrm{End}} = Softmax(Logit_{\mathrm{End}}) \tag{15}$$

2 实验

2.1 数据

本文使用的数据是哈工大讯飞联合实验室机器阅读理解组 (HFL-RC) 于 2018 年发布的中文篇章片段抽取型阅读理解数据集 CMRC2018[22], 由近 20 000 个由人类专家在维基百科段落中注释的真实问题组成. 我们使用其给出的训练数据集来进行模型的训练, 用其验证数据集来对模型进行评估. 图 4 是该数据集的样例, 包括 1 篇文章以及 2 个问题, 其中蓝色文字表示与问题 1 相关的内容, 红色文字表示与问题 2 相关的内容.

【文章】: 计算神经科学 (Computational neuroscience) 为一种跨领域科学, 包含神经科学、认知科学、资讯工程、电脑科学、物理学及数学。这个词首次出现于1985年, 由于加州卡莫市主办的会议中提出。其后出现的类似名词包含神经模型、脑理论及神经网络。后来相关的解释定义皆收录于麻省理工学院出版(1990)之《计算神经科学》("Computational Neuroscience") 一书内。

【问题1】计算神经科学包含哪些科学?
【答案1】包含神经科学、认知科学、资讯工程、电脑科学、物理学及数学

【问题2】计算神经科学一词首次出现是在什么时候?
【答案2】这个词首次出现于1985年

图 4 CMRC2018 数据集示例

针对基于 BERT 模型的机器阅读理解任务, 我们对数据进行了一些预处理来使得其符合 BERT 的输入限制. 首先我们对数据进行字粒度的切分, 并将问题和文章进行拼接, 并固定输入的序列长度为 512. 若输入的数量超出这个长度, 则利用 128 的滑动窗口来切分为多份数据.

而对于文本特征, 我们则利用已有的模型对实验数据进行了词性标注、命名实体识别以及依存分析的预处理. 接着以 BIO 规则进行标注, 以适应字符级别的输入粒度.

2.2 评价指标

对于抽取式机器阅读理解模型, 我们需要评估答案预测值和真实答案之间的字面匹配程度, 本文采用了文献 [1] 中的 EM 和 $F1$ 两个指标. EM 为模型预测的验证数据集中的答案与真实答案完全一致的百分比, 而 $F1$ 为机器学习中常用的指标, 是精确率与召回率的调和平均. 在本文场景下, 将答案的预测字符串与真实值各自按字符切分后, 分别视作词袋, 并计算二者的 $F1$ 值来粗粒度地评估它们的匹配程度.

2.3 实验设置

算法模型的搭建使用深度学习框架 PyTorch[23] 实

现, 其中的基准模型使用针对中文语料进行预训练的 Chinese-roBERTa-wwm-ext 模型[24], 相较于最初的中文预训练模型 BERT-base-Chinese[6], 该模型将掩码语言模型 (masked language model) 的训练策略由遮盖单个字变更为遮盖整个中文词, 且使用了更大规模的中文语料, 其在相关下游任务上有更强的表现.

我们使用最后一层的输出作为上下文表示特征, Aken 等人[15] 和 Cai 等人[25] 的工作也分别展示了在机器阅读理解任务上, BERT 中越高层的编码输出越有效. 模型的主要参数设置如下: batch size 设置为 4, 学习率为 3E–5, 并采用学习率预热的策略[26], dropout 设置为 0.2, 使用训练集微调两个轮次后, 在验证集上取得了不错的基准效果.

接下来我们分别尝试将 BERT 的输出与文本特征的输入进行交互. 共设置了 5 组实验, 分别基准模型的实验, 在基准模型基础上分别融合词性特征、命名实体特征和依存分析特征的实验以及融合全部特征的实验. 如表 1 所示.

对于每组实验, 我们分别设置了 5 个随机种子进行多次实验, 使用 5 次不同随机种子实验中性能的最佳结果以及平均结果作为该组模型实验的最终结果, 以排除一些训练过程中的随机性.

表 1 对照实验组别与模型

实验组	模型	特征
1	BERT (baseline)	无
2	BERT+POS	词性特征
3	BERT+NE	命名实体特征
4	BERT+SEP	依存分析特征
5	BERT+POS+NE+SEP	词性特征+命名实体特征+依存分析特征

2.4 实验结果与分析

(1) 模型阅读理解能力. 在数据集 CMRC2018 上的实验结果见表 2, 其中加粗行分别是添加单特征的最优实验结果和添加全部特征的实验结果.

表 2 在数据集 CMRC2018 上的实验结果 (%)

模型	5次平均		5次最佳	
	*F*1	EM	*F*1	EM
BERT (baseline)	85.56	66.84	86.21	67.85
BERT+POS	**85.85 (+0.29)**	**67.46 (+0.62)**	**86.39 (+0.18)**	**68.72 (+0.87)**
BERT+NE	85.77	67.13	86.18	68.22
BERT+DEP	85.61	67.37	86.28	68.58
BERT+POS+NE+DEP	**85.91 (+0.35)**	**68.27 (+1.43)**	**86.58 (+0.37)**	**69.41 (+1.56)**

表 2 展示了基准模型 Chinese-roBERTa-wwm-ext 经微调后, 最高可以达到 86.21% 的 *F*1 匹配率和 67.85% 的精确匹配率. 在此基础上分别添加词性特征、命名实体特征、依存句法特征进行实验, 实验结果表明每一种特征的融合都能够带来模型的精度的提升, 且对于 EM 值上的提高要明显高于 *F*1 值. 其中词性特征带来的匹配率提升效果最为显著, 可以达到 85.85% 的平均 *F*1 匹配率和 67.46% 的平均 EM 匹配率, 相较于基准模型分别可以提升 0.3% 和 0.6%, 而最优轮次的 EM 相较于基准模型提高 0.87%. 实体特征和依存分析特征也同样在两个评估标准上相较基准模型有一定的提高, 但较词性特征而言并不显著.

同时添加 3 项特征后, 实验结果可以达到 85.91% 的平均 *F*1 匹配率和 68.27% 的平均 EM 匹配率, 相较于基准模型分别可以提高 0.35% 和 1.43%, 相较于只融合单特征的实验结果, EM 值得到了接近一个百分点的提升. 而我们的最优模型达到了 86.58% 的 *F*1 匹配率和 69.41% 的 EM 匹配率, 相较于只使用预训练 BERT 模型, 分别可以得到 0.37% 和 1.56% 的提升.

基于上述实验结果以及分析, 我们的方法可以在预训练模型的基础上得到 1.5% 左右的 EM 匹配率提升, 证明了提出方法的有效性, 并且在 3 种特征中, 词性特征更加能够帮助阅读理解模型进行预测. 实验结果也验证了在 BERT 等预训练模型中引入显式的语言学知识同样能够帮助机器进行阅读理解. 至于 EM 值的提升如此显著, 我们分析认为显式的语言特征本身就是更加结构化的特征, 因此能够更有效地帮助机器归纳总结出更加精确的答案起始位置.

(2) 效率. 我们对“对预训练模型进行改进”与“引入显示语言学特征”这两种方法的细节进行比较, 包括训练的数据规模与算力成本, 以及各自在 CMRC2018 数据集上带来的性能提升百分比, 对比结果见表 3. 其中, B 表示 10 亿.

表 3 对预训练模型改进与引入显示语言学特征两种方法的比较

特点	预训练模型改进		引入显式语言学特征
具体方法	中文全词覆盖, 大规模数据集, 任务改进		引入词法与句法特征
代表模型	BERT-wwm-ext	RoBERTa-wwm-ext (base)	本文模型
单词数量	5.4 B	5.4 B	<<0.1 B
数据集大小	30 GB	30 GB	107 MB
训练步数	10E+6	10E+6	10E+2
训练时间	数周 (TPU v3, 128 GB)	数周 (TPU v3, 128 GB)	20 min (3080Ti, 20 GB)
原始模型	BERT-wwm	BERT-wwm	RoBERTa-wwm-ext
EM/*F*1提升 (%)	+0.8/+0.1	+1.1/**+1.6**	**+1.56**/+0.37

预训练模型改进: Cui 等人[24] 提出的 BERT-wwm-ext 在 BERT 的基础上, 将词掩码方式设置为中文全词覆盖, 并引入了包括百科、新闻以及问答页面的训练文本, 词量高达 5.4 B. 训练步数的数量级也高达百万, 在 TPU v3 上通常需要数周. 最终在 CMRC2018 验证集上分别可以得到 0.8% 和 0.1% 的 EM 和 *F*1 指标提升. 而 RoBERTa-wwm-ext 进一步移除了下句预测任务, *F*1 和 EM 分别提升了 1.1% 和 1.6%.

引入显式语言学特征: 本文方法引入词法和句法特征, 在远小于 0.1 B 词量的数据集上利用现有模型进行标注, 融合了分词结果和各类特征的训练集大小为 107 MB, 以 4 为 batch size 在 3080Ti 上训练 2 个轮次, 共需要 20 min, 而 EM 和 *F*1 指标最高可以提升 1.56% 和 0.37%.

本文方法相较对预训练模型进行改进, 使用远少于后者的数据与算力成本, 在阅读理解数据集上获得了持平甚至更优的指标提升. 可见本文方法较为高效, 同时也证明了引入显式的词法句法等语言特征能够为特定的下游任务带来较大的性能提升.

3 结论与展望

本文提出一种融合多种特征的抽取式机器阅读理解模型, 显式地引入包括词性、命名实体的词法特征以及依存分析的句法特征, 同时设计了基于注意力机制的自适应特征选择模块, 进一步提升了机器阅读理解模型的性能. 在抽取式机器阅读理解数据集 CMRC2018 的实验上表明, 本文提出的机器阅读理解模型能够通过极低的算力成本, 在 *F*1 和 EM 指标上取得 0.37% 和 1.56% 的提升.

实验结果验证了我们方法的有效性. 对于阅读理解模型而言, 词性特征相较命名实体特征和句法依存分析特征更能够帮助模型理解文本. 同时也说明了对于机器阅读理解这类难度较高的自然语言处理任务, 尽管 BERT 等预训练模型带来的表征能力是突破性的, 但是语言本身的一些特征也具有不可忽视的作用. 在未来的研究中包括但不限于词法、句法等各类语言学特征同样值得更多的关注, 它们在与预训练模型的结合中究竟起到了怎样的作用以及这些特征的重要程度都是值得关注的研究课题, 同时在二者的结合中也可以进一步帮助我们了解 BERT 等预训练模型对于语言的理解机制.

参考文献

1 Rajpurkar P, Zhang J, Lopyrev K, *et al*. SQuAD: 100,000+ questions for machine comprehension of text. Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing. Austin: Association for Computational Linguistics, 2016. 2383–2392.

2 Radford A, Narasimhan K, Salimans T, *et al*. Improving language understanding by generative pre-training. https://cdn.openai.com/research-covers/language-unsupervised/language_understanding_paper.pdf.

3 Devlin J, Chang MW, Lee K, *et al*. BERT: Pre-training of deep bidirectional transformers for language understanding. Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). Minneapolis: Association for Computational Linguistics, 2019. 4171–4186.

4 Hermann KM, Kočiský T, Grefenstette E, *et al*. Teaching machines to read and comprehend. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal: MIT Press, 2015. 1693–1701.

5 Wang SH, Jiang J. Machine comprehension using match-lstm and answer pointer. arXiv: 1608.07905, 2016.

6 Seo MJ, Kembhavi A, Farhadi A, *et al*. Bidirectional attention flow for machine comprehension. arXiv:

1611.01603, 2016.
7 Yu AW, Dohan D, Luong MT, *et al*. QaNet: Combining local convolution with global self-attention for reading comprehension. arXiv: 1804.09541, 2018.
8 LeCun Y, Bottou L, Bengio Y, *et al*. Gradient-based learning applied to document recognition. Proceedings of the IEEE, 1998, 86(11): 2278–2324. [doi: 10.1109/5.726791]
9 Elman JL. Finding structure in time. Cognitive Science, 1990, 14(2): 179–211. [doi: 10.1207/s15516709cog1402_1]
10 Wiedemann G, Remus S, Chawla A, *et al*. Does BERT make any sense? Interpretable word sense disambiguation with contextualized embeddings. Proceedings of the 15th Conference on Natural Language Processing. Erlangen: KONVENS, 2019. 161–170.
11 Rajpurkar P, Jia R, Liang P. Know what you don't know: Unanswerable questions for SQuAD. Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers). Melbourne: Association for Computational Linguistics, 2018. 784–789.
12 Jawahar G, Sagot B, Seddah D. What does BERT learn about the structure of language? Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. Florence: Association for Computational Linguistics, 2019. 3651–3657.
13 Si CL, Wang SH, Kan MY, *et al*. What does BERT learn from multiple-choice reading comprehension datasets? arXiv: 1910.12391, 2019.
14 Albilali E, Altwairesh N, Hosny M. What does BERT learn from Arabic machine reading comprehension datasets? Proceedings of the 6th Arabic Natural Language Processing Workshop. Kyiv: Association for Computational Linguistics, 2021. 32–41.
15 van Aken B, Winter B, Löser A, *et al*. How does BERT answer questions? A layer-wise analysis of transformer representations. Proceedings of the 28th ACM International Conference on Information and Knowledge Management. Beijing: ACM, 2019. 1823–1832.
16 Zhang ZS, Wu YW, Zhao H, *et al*. Semantics-aware BERT for language understanding. Proceedings of the AAAI Conference on Artificial Intelligence, 2020, 34(5): 9628–9635. [doi: 10.1609/aaai.v34i05.6510]
17 Mikolov T, Sutskever I, Chen K, *et al*. Distributed representations of words and phrases and their compositionality. Proceedings of the 26th International Conference on Neural Information Processing Systems. Lake Tahoe: Curran Associates Inc., 2013. 3111–3119.
18 Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). Doha: Association for Computational Linguistics, 2014. 1532–1543.
19 Xia F. The part-of-speech tagging guidelines for the Penn Chinese Treebank (3.0). IRCS Technical Reports Series, 2000. 38.
20 Chang PC, Tseng H, Jurafsky D, *et al*. Discriminative reordering with Chinese grammatical relations features. Proceedings of the 3rd Workshop on Syntax and Structure in Statistical Translation. Boulder: Association for Computational Linguistics, 2009. 51–59.
21 Li YH, Wang NY, Liu JY, *et al*. Factorized bilinear models for image recognition. Proceedings of the 2017 IEEE International Conference on Computer Vision. Venice: IEEE, 2017. 2098–2106.
22 Cui YM, Liu T, Che WX, *et al*. A span-extraction dataset for Chinese machine reading comprehension. Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). Hong Kong: Association for Computational Linguistics, 2018. 5883–5889.
23 Paszke A, Gross S, Massa F, *et al*. PyTorch: An imperative style, high-performance deep learning library. Proceedings of the 33rd International Conference on Neural Information Processing Systems. Red Hook: Curran Associates Inc., 2019. 721.
24 Cui YM, Che WX, Liu T, *et al*. Pre-training with whole word masking for Chinese BERT. arXiv: 1906.08101, 2019.
25 Cai J, Zhu ZZ, Nie P, *et al*. A pairwise probe for understanding BERT fine-tuning on machine reading comprehension. Proceedings of the 43rd International ACM SIGIR Conference on Research and Development in Information Retrieval. ACM, 2020. 1665–1668.
26 Gotmare A, Keskar NS, Xiong CM, *et al*. A closer look at deep learning heuristics: Learning rate restarts, warmup and distillation. arXiv: 1810.13243, 2018.

(校对责编: 牛欣悦)