工程科学学报, 第 47 卷, 第 1 期: 79−90, 2025 年 1 月
Chinese Journal of Engineering, Vol. 47, No. 1: 79−90, January 2025
https://doi.org/10.13374/j.issn2095-9389.2023.11.05.002; http://cje.ustb.edu.cn

基于渐进机器学习的中文问句匹配方法

贺学剑[1)], 陈安琪[2)], 郭志强[1)], 王致茹[3)], 陈　群[2,3)]✉

1) 河南林业职业学院, 洛阳 471002　2) 西北工业大学软件学院, 西安 710072　3) 西北工业大学计算机学院, 西安 710072
✉通信作者, E-mail: chenbenben@nwpu.edu.cn

摘　要　问句匹配旨在判断不同问句的意图是否相近. 近年来, 随着大型预训练语言模型的发展, 利用其挖掘问句对在语义层面隐含的匹配信息, 取得了目前为止最好的性能. 然而, 由于基于独立同分布假设, 在真实场景中, 这些深度学习模型的性能仍然受制于训练数据的充足程度和目标数据与训练数据之间的分布漂移. 本文提出一种基于渐进机器学习的中文问句匹配方法. 该方法基于渐进机器学习框架, 从不同角度提取问句特征, 构建融合各类特征信息的因子图, 然后通过迭代的因子推理实现从易到难的渐进学习. 在特征建模中, 设计并实现了两种类型特征的提取: (1) 基于 TF-IDF (Term frequency-inverse document frequency) 的关键词特征; (2) 基于 DNN (Deep neural network) 的深度语义特征. 最后, 通过通用的基准中文数据集 LCQMC 和 BQ corpus 验证了所提方法的有效性. 实验表明, 相比于单纯的深度学习模型, 基于渐进机器学习的方法可以有效提升问句匹配的准确率, 且其性能优势随着标签训练数据的减少而增大.

关键词　自然语言理解; 中文问句匹配; 渐进机器学习; 自然语言预训练模型; 因子图推理
分类号　TG319

Question-matching approach based on gradual machine learning

HE Xuejian[1)], *CHEN Anqi*[2)], *GUO Zhiqiang*[1)], *WANG Zhiru*[3)], *CHEN Qun*[2,3)]✉

1) Henan Forestry Vocational College, Luoyang 471002, China
2) School of Software, Northwestern Polytechnical University, Xi'an 710072, China
3) School of Computer Science, Northwestern Polytechnical University, Xi'an 710072, China
✉Corresponding author, E-mail: chenbenben@nwpu.edu.cn

ABSTRACT　Question matching attempts to determine whether the intentions of two different questions are similar. Recently, with the development of large-scale pretrained DNN (Deep neural network) language models, state-of-the-art question-matching performance has been achieved. However, due to the independent and identically distributed assumption, the performance of these DNN models in real-world scenarios is limited by the adequacy of the training data and the distribution drift between the target and training data. In this study, we propose a novel gradual machine learning (GML)-based approach for Chinese question matching. Beginning with initially labeled instances, this approach gradually labels target instances in order of increasing hardness *via* iterative factor inference on a factor graph. The proposed solution first extracts diverse semantic features from different perspectives and then constructs a factor graph by fusing the extracted features to facilitate gradual learning from easy to hard. In feature modeling, we extract and model two complementary types of features: 1) TF-IDF-based keyword features, which can capture the shallow semantic similarity between two questions; 2) DNN-based deep semantic features, which can capture the latent semantic similarity between two questions. We model keyword features as unary factors in a factor graph, which define their influence on the matching status of the two questions. The DNN-based features contain global and local features, where the global features correspond to a question pair's matching probability as estimated by a DNN model, and the local features correspond to the semantic similarity between two neighboring question pairs

收稿日期: 2023−11−05
基金项目: 国家自然科学基金面上资助项目 (62172335)

estimated by their vector representations in a DNN's embedding space. To facilitate gradual inference, we model the DNN-based global and local features as unary and binary factors, respectively, in a factor graph. Finally, we implement a GML solution for question matching based on an open-sourced GML inference engine. We validated the efficacy of the proposed approach through a comparative study on two open-sourced Chinese benchmark datasets, LCQMC and the BQ corpus. Extensive experiments demonstrate that compared with pure deep learning models, the proposed solution effectively improves the accuracy of question matching, and its performance advantage generally increases with a decrease in labeled training data. Our experiments also demonstrate that the performance of the proposed solution is very robust w.r.t key algorithmic parameters, indicating its applicability in real-world scenarios. In addition, our work on the GML solution is orthogonal to existing deep learning-based question-matching algorithms because our solution can easily accommodates and leverages other deep language models.

KEY WORDS natural language understanding; chinese question matching; gradual machine learning; natural language pretraining model; factor graph inference

问句匹配任务的研究对象为篇幅短且信息密集的问句, 核心任务是判断问句对的意图是否相似或者相同. 问句匹配技术在自然语言处理中具有广泛的应用. 例如, 在搜索引擎上, 问句匹配技术可以在用户输入问题时匹配到相似问题[1], 帮助用户更准确地表述查询关键词以获取更准确的反馈答案; 在短视频应用 (App) 和电商购物平台上, 智能推荐算法根据用户的行为特点推导用户的个人喜好, 从而精准推送符合用户需求的目标产品[2]; 其推荐原理一般是通过计算新产品和用户历史购买或浏览数据的匹配程度, 以判断是否推送. 另外, 在广泛使用的 FAQ (Frequently asked questions) 问答系统中[3], 问句匹配可以帮助系统快速完成相似问题的匹配, 从而将对应的标准答案反馈给用户[4]. 一方面, 问句是进行语义理解、意图分析的典型研究对象, 在自然语言处理领域具有很高的研究价值和关注度; 另一方面, 因为实用性强, 商业价值高, 国内外知名互联网公司如谷歌, 百度和美团等, 纷纷投入重金研究问句匹配技术.

问句匹配技术和自然语言处理技术的发展历程几乎同步. 传统的问句匹配技术从统计概率的角度提取文本特征, 直观地计算问句相似度[5]. 然而, 这些传统算法特征提取能力较弱, 仅能够从字词等层面提取特征, 忽略掉了意图等至关重要的语义层面信息; 同时就语言理解本身而言, 相同字词在不同上下文中可能存在多种含义, 且不同字词也可能具有相似的含义[6]. 随着深度学习技术的兴起, 对问句匹配技术的研究逐渐转向以深度神经网路, 特别是大型预训练语言模型为基础, 如 Word2vec[7]、BERT[8] 和 RoBERTa[9] 等. 这些预训练模型文本表征能力强大, 对语义信息的解析更加全面准确, 大大提升了问句匹配的准确率. 深度学习模型虽然可以取得更高的准确率, 但在一定程度上损失了可解释性和易操作性, 毕竟其内部的计算过程精密而复杂, 对于大多数研究者而言类似“黑盒”. 因此, 近年来有些研究者尝试把传统匹配模型和深度学习模型结合起来, 希望保留二者各自优点的同时弥补相互之间的不足. 例如, 文献 [10] 中采用 TF-IDF(Term frequency-inverse document frequency)传统算法区分文本中词汇的重要程度, 采用深度学习模型表示的词向量提取文本特征, 计算相似度时, 将 TF-IDF 值作为词向量权重, 带权求和表征总体文本信息; 文献 [11] 使用词性和 TF-IDF 对词向量进行加权, 以提升句向量表征的效果.

这些深度学习模型的有效性依赖于独立同分布假设. 然而, 在真实场景中, 标注大量的标签数据通常需要较高的人力成本, 因此不可轻易获取; 使情况更糟的是, 目标数据和训练数据通常都存在一定程度的分布漂移, 这使得一个模型即使在训练数据上得到充分训练, 其在目标数据上的表现依然可能不如预期. 为了解决以上不足, 本文提出了一种基于渐进机器学习的问句匹配方法. 渐进机器学习最早由陈群团队提出[12]. 作为一种通用的机器学习框架, 渐进机器学习不同于传统的机器学习框架(如深度学习), 不是基于独立同分布假设, 而是从一些证据标签数据开始, 以从易到难的顺序渐进地标注目标数据. 除实体解析任务外, 渐进机器学习也已被应用于情感分析等自然语言处理任务[13]. 在渐进机器学习中, 渐进学习是通过因子图的渐进推理实现的, 而标注数据和未标注数据之间的知识传递是通过共享特征实现.

在本文中, 针对问句匹配任务的要求, 本文设计并实现了两种类型的特征以实现渐进学习: (1) 基于 TF-IDF 的关键词特征; (2) 基于 DNN(Deep neural network)的深度语义特征. 关键词特征旨在

实现浅层语义的知识传递, 而 DNN 特征旨在实现深层语义的知识传递. 基于 DNN 的特征种类包含两个互补性的特征: (1) 基于 DNN 的预测概率特征; (2) 基于 DNN 的最近邻特征. 预测概率特征是从全局的层面推测问句匹配的概率, 而最近邻特征从局部层面推测匹配的概率. 对问句匹配任务而言, 不同的深度学习模型即便绝对性能差异不大, 其提取的特征也有一定的互补性. 因此, 本文提出的算法利用多个深度神经网路模型提取多样化的特征表达, 以提升渐进学习的效能.

1 相关工作

最早解决问句匹配问题的思路是基于字词统计信息的文本匹配, 其中使用频率较高的算法有编辑距离[14](Edit distance)、N 元模型[15](N-gram)、杰卡德相似系数[16](Jaccard similarity coefficient)和 TF-IDF[17] 等. 这些算法优点在于原理简单且易操作, 可解释性强, 但缺点在于仅仅从字形词频等表层信息分析文本匹配度, 维度单一, 无法获取深层语义的信息. 深度学习模型的出现为提取深层语义信息提供了新思路, 有效弥补了传统文本匹配技术的局限性. 相关发展最早可追溯至 1986 年, Hinton 提出用 Distributed representation 的概念表示词[18], 即用向量代表句子中的每一个词, 将每个词映射为特征空间中的某一点, 据此计算词和词之间的物理距离, 来代表句子之间的相似度. 2003 年, NPLM 模型(Neural probabilistic language model)的诞生标志着人们初次尝试将深度学习模型应用于自然语言处理领域[19]. 2006 年起, 陆续出现了众多针对自然语言处理的深度学习模型, 如 Word2vec[7,20−21]、LSTM[22−23]、RNNLM[24]、LSTM-RNNLM[25]、BiRNN[26] 和 BiLSTM[27] 等.

随着 BERT 模型[28] 的提出, 关于深度学习模型的研究转向大规模预训练语言模型. BERT 通过两阶段的迁移学习方式, 包括预训练阶段和微调阶段, 实现了很高的通用性. 在预训练阶段, BERT 利用 Masked language model(MLM)和 Next sentence prediction(NSP)两个类型的任务最小化组合损失函数, 得到预训练模型. 在微调阶段, BERT 进一步根据下游任务的特点微调模型; 其具体训练过程和预训练阶段一致, 最主要的差异是在最后一层 Transformer 的输出序列顶部添加分类层以获取信息. 应用于问句匹配任务时, BERT 在经过训练获取到向量表征后, 取 [CLS] 并通过 Softmax 分类函数即可获得最终的相似度[29]. 作为自然语言处理领域里程碑式的模型, BERT 在自然语言处理上已得到广泛使用, 各种基于 BERT 的微调模型层出不穷, 例如: Roberta[9]、ALBERT[30] 和 XLNet[31] 等.

在国内, 针对中文问句匹配领域的研究虽然起步较晚, 但同样发展迅速[32]. 相对于英文, 中文除字形结构和英文本质上不同以外, 在组词造句时, 词语间不存在分隔符也是中英文之间较为显著的差异. 同时, 中文词语所蕴含的含义更为丰富, 字组词和词组句的组合方式十分灵活. 以上情况给中文问句匹配技术的研究带来了许多新挑战[33−34]. 最早的一些方法尝试直接将基于词频等统计信息的文本匹配算法用于中文, 例如杰卡德系数[35] 或改进的 TF-IDF 算法[36]. 这些算法虽然取得一定效果, 但总体性能仍不理想. 随着深度学习技术的兴起, 不少结合中文特点的深度学习匹配模型被陆续提出. 2016 年, 复旦大学在 LSTM 的基础上提出 DF-LSTM 模型[37]; 中科院同年提出了计算中文文本相似度的 MV-LSTM 模型[38]; 2019 年, 哈工大讯飞联合实验室提出基于 BERT 的中文预训练模型 BERT-wwm[32]. 同年, 百度飞桨提出的 ERNIE 模型通过持续学习海量文本中词汇语法知识, 在当时 40 多个中文自然语言处理任务中取得最佳成绩[39]. 香侬科技也于 2019 年提出 Glyce 预训练模型, 利用中文字形提升 BERT 的表征能力, 在序列标注、句对分类和单句分类等任务上取得优异的性能[40].

近年来, 对问句匹配的研究转向在预训练语言模型基础上的进一步优化提升. 比如, 周丽杰等[36] 研究如何利用检索系统里积累的大量相似问题来提升问题匹配的精度, 提出了一个 RSEN(Relation-aware semantic enhancement network)模型用于捕捉问句之间的关系. 类似地, Huang 等[41] 提出一种 IFE(Interaction feature extractor)结构来捕捉问句之间的关系, 并将捕捉的信息融合到判定模型中用以提升判定精度. 与此同时, Ying 等[42] 提出融合词语层面和句子层面的特征表达来提升问句之间的语义相似度匹配精度. Faseeh 等[43] 则提出利用不同的深度神经网络来分别提取问句中的短期和长期语义依赖, 并综合考虑它们提取的特征来判定问句之间的相似度. 赵云肖等[44] 提出利用中文形音义等多元的知识表达来提升问句匹配的精度. 还有一些学者针对某些具体应用领域提出优化的问句匹配技术. Guo 等[45] 研究了金融邻域的问句匹配问题, 通过一种 FinKENet(Financial knowledge enhanced network)结构来捕捉金融知识, 并利用其

来提升问句的语义表达精度. 徐若卿等[46]针对医疗领域的问句匹配, 提出了利用医疗知识图谱来提升问句的语义表达精度. 张劲桉等[47]针对旅游问题的问句匹配, 提出综合利用不同角度的相似度度量来提升匹配精度.

渐进机器学习最早由陈群团队[12]提出, 并应用于实体解析任务, 主要解决在只有少量甚至没有标签数据的情况下如何实现分类的难题. 渐进机器学习技术能基于初始少数的证据实例, 通过由易到难的渐进推理实现准确的分类标注[48−51]. 作为一种通用的机器学习框架, 渐进机器学习后来也被应用于情感分析问题[13,49]和图像分类[52], 取得了超越单纯深度学习模型的性能. 在本文中, 将其应用于中文问句匹配问题.

2　基于渐进机器学习的算法框架

渐进机器学习的总体思路是将已标注实例和未标注实例建模在因子图模型中, 依据在已标注实例上学习到的知识渐进推理未标注实例的标签.

2.1　因子图模型

如图 1 所示, 因子图模型由观测变量、隐变量以及它们之间的因子构成. 观测变量对应已标注实例, 隐变量对应未标注实例; 在问句匹配问题中, 分别对应已标注的问句对和待标注的问句对. 问句匹配的因子图中, 每个变量的取值为 0 或者 1, 其中 1 意味着其对应的问句对“语义匹配”, 而 0 意味着“语义不匹配”. 变量之间的关系则通过因子节点来描述. 在问句匹配的因子图模型中, 因子可以是单因子或双因子. 单因子描述关联变量的标签状态与某一特征的关系; 双因子则直接描述两个变量的标签状态之间的关系. 一般来说, 单因子或双因子刻画的关系都满足统计单调性. 在单因子的定义中, 一个关联变量具有某个标签状态的概率会随着对应特征值的增大而增大; 在双因子定义中, 两个变量具有相同标签状态的概率会随着两个变量特征相似度的增大而增大.

因子图中, 所有变量的联合概率分布(P_w)可以表示为所有因子的连乘积, 具体如下所示:

$$P_w(V)=\frac{1}{Z_w}\prod_{v\in V}\left[\phi_{\mathrm{f}}(v)\prod_{\mathrm{f_u}\in F_{\mathrm{u}}^{v}}\phi_{\mathrm{f_u}}(v)\right]\prod_{\mathrm{f_u}\in F_{\mathrm{r}}}\phi_{\mathrm{f_r}}\left(v_i,v_j\right) \tag{1}$$

其中, v为因子图中的一个变量; ϕ为因子条件概率, Z 为规范常数, w 为因子权重, f、u、r 分别表示指示因子 (factor)、单因子 (unary factor)、指示双因子 (relational factor); V表示所有变量的集合; F_{k}^{v}表示变量v关联的所有单因子集合; F_{r}表示全部关系双因子集合. 给定一个因子图, 因子图推理通过优化以下目标函数, 学习与观测实例标签保持最大一致的因子权重($\hat{w}$):

$$\hat{w}=\arg\min_{w}-\ln\sum_{V_I}P_w(\Lambda,V_I) \tag{2}$$

其中, Λ表示与因子相连的所有已标注变量集合; V_I表示与因子相连的所有待标注变量集合, I 表示

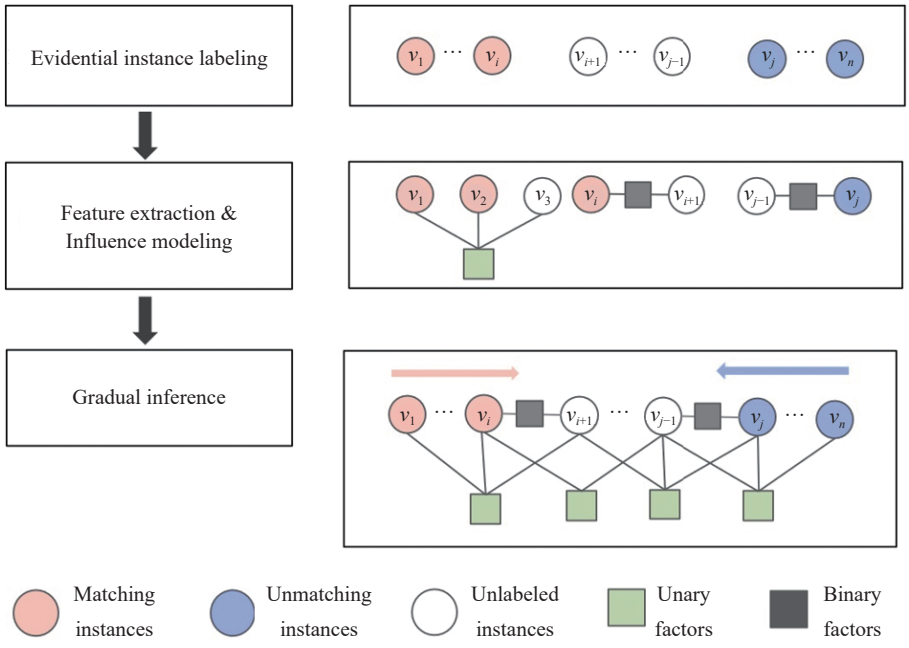

图 1　渐进机器学习算法框架

Fig.1　Framework of the gradual machine learning algorithm

待标注的变量.

2.2 渐进机器学习算法步骤

如图 1 所示, 与渐进机器学习的通用框架一致, 问句匹配的渐进机器学习算法也是由三个步骤组成: (1)初始证据实例标注; (2)特征抽取与因子建模; (3)因子图渐进推理.

2.2.1 初始证据实例标注

该步骤主要用于获取初始证据实例的数据信息, 标注为因子图中的可观测变量. 对于无监督任务, 初始证据实例标注通常可以通过专家规则或无监督方法等实现. 例如, 文本分类任务中, 通过无监督聚类方法获取各个类别的类中心, 所有靠近类中心的实例在统计上有比较大的概率与类中心有相同的标签, 因此可以标记为初始证据实例. 对于本文的问句匹配任务而言, 初始证据实例即为训练集中的所有标注数据, 无需额外标注.

2.2.2 特征抽取与因子建模

该步骤主要根据已标注实例和未标注实例之间的关系构建因子图, "关系"意味着它们之间有着共有特征, 共有特征能保证学习推理的过程中知识信息在实例之间有效传递. 对于问句匹配任务而言, 本文提取两种类型的特征: (1)基于 TF-IDF 的关键词特征; (2)基于 DNN 的深度语义特征.

在问句匹配中, 语义近似的两个问句一般都包含相同或类似的关键词. 与此相反, 不匹配的问句对中, 其中一个问句通常包含有另一个问句不包含的关键字. 如表 1 所示, Q2 中包含有关键词"时候", 而 Q1 中不包含与其语义相似的关键词, 所以 Q1 和 Q2 不匹配. 具体地, 提取关键词单现和共现两个特征来帮助问句匹配的推理. 关键词单现特征指的是一个关键词出现在一个问句中, 但不出现在另一问句中, 如表 1 中"时候"这一关键词; 而关键词共现特征指的是某一个关键词同时出现在两个问句中, 如表 1 中"礼物"这一关键词. 在算法实现中, 为了提高知识传递的效能, 不是简单提取某个关键词的单现和共现特征, 而是先把关键字按语义相似度聚类, 然后按照一组语义相似的关键词来提取.

表 1 关键字特征抽取示例

Table 1 Example of keyword feature extraction

Question identity	Question
Q1	过年给长辈送什么礼物比较好
Q2	过年前什么时候给长辈送礼物
Label	0(不匹配)

由于自然语言的复杂性, 通过关键词浅层语义一般不足以判断问句是否匹配. 因此, 本文同时利用大规模预训练语言模型(如 ERNIE 和 Glyce 等)来捕捉深层语义, 并借此判断问句之间的语义相关性. 一般地, 深度语言模型都是把问句映射到一个高维的特征嵌入空间中, 然后根据特征分布来判断问句的语义相似度. 在嵌入空间中, 语义相似程度较高的问句通常会被映射到距离相近的点. 所以, 利用深度语言模型生成问句对的向量表达, 然后提取两个特征: (1)全局的相似度特征, 即直接用问句对的向量表达判断匹配概率; (2)局部的 KNN (K nearest neighborhood) 关系特征, 即给定一个问句对, 在嵌入空间中寻找与其最近邻的问句对, 通过最近邻问句对的标签推理其本身的标签.

该步骤也需要把特征建模成因子图中的共享因子, 以实现渐进学习. 如图 1 所示, 本文将关键词特征和基于 DNN 的全局相似度特征建模成单因子, 把基于 DNN 的局部 KNN 关系特征建模成双因子. 与之前的影响力建模方法类似[12], 不管是单因子还是双因子, 都是通过单调的 sigmoid 函数来量化其对匹配概率的影响力.

2.2.3 因子图渐进推理

"渐进推理"是指整个推理过程并不是一次性推导出最终结果, 而是需要进行多次迭代, 每次迭代时仅选取证据确定性较高的一个或几个待标实例进行标注; 新标注的实例作为新的证据实例, 继续加入下一次迭代推理过程, 为剩余的待标实例提供更多的知识. 在反复迭代的过程中, 证据实例不断增多, 而待标实例的数量不断减少, 最终推理出整个因子图中所有实例的标签.

已知证据变量的概率分布, 经过因子图推理可获得整个因子图的联合概率分布, 再通过吉布斯采样即可得到单个待标注变量的边缘概率分布 $P(v)$. 在标注实例时, 需要选出当前概率分布表现最为极端, 也就是确定性最高的一个实例进行标注. 具体地, 通过使用熵 $H(v)$ 的倒数来衡量每个实例当前被标注的确定性. $H(v)$的计算公式如下:

$$H(v) = -\{P(v)\cdot\log_2 P(v) + [1-P(v)]\cdot\log_2[1-P(v)]\} \tag{3}$$

在具体的算法实现中, 因为因子图推理通常比较耗时, 与之前渐进机器学习在实体解析和情感分析的应用一样, 我们使用一种可扩展的高效方法实现渐进推理. 具体地, 如算法 1(表 2)所示, 该算法过程如下: (1)先选取待标实例中证据支持度最高的前 top-m 个实例; (2)然后在这 m 个候选

表 2　可扩展渐进推理算法

Table 2　The scalable gradual inference algorithm

Algorithm 1: gradual inference algorithm	
Input:	factor graph G; parameter m, k, n
Output:	All labels of unlabeled variables
1	while there exists any unlabeled variable in G do
2	$V' \leftarrow$ all the unlabeled variables in G;
3	for $v \in V'$ do
4	Measure the evidential support of v in G;
5	Select the top-m unlabeled variables with the most evidential support (denoted by V_m);
6	for $v \in V_m$ do
7	Approximately rank the entropy of v in$G$;
8	Select the top-k most promising variables in terms of entropy in V_m(denoted by V_k);
9	for $v \in V_k$ do
10	Compute the probability of v in G by factor graph inference over a subgraph of G;
11	Label the variable with the minimal entropy in V_k, with a probability greater than 0.5 marked as 1, otherwise marked as 0;

实例中, 选取近似熵最低的 top-k 个实例: (3)最后, 通过因子图推理一一计算这 k 个实例的匹配概率, 选取其中熵最低的实例进行标注. 需要指出的是, 针对问句匹配的渐进推理过程与之前针对情感分析的渐进推理过程基本相同, 都是基于开源的渐进机器学习引擎开发 (https://github.com/gml-explore/gml). 用户只需构建好因子图, 并预先标注好初始证据实例, 渐进推理过程可由开源引擎自动实现.

3　因子图建模

3.1　特征提取

本文方案提取两种类型的特征: (1)显式语义的关键词特征; (2)深层语义的 DNN 特征.

(1) 关键词特征. 关键词在问句匹配中发挥重要作用. 如表 1 所示, 两个问句的主要组成部分高度相似, 最主要的差异为关键词——“时候”. 然而, 该处差异代表了最关键的不匹配信息, Q1 中“送什么”询问的是物品, 而 Q2 中“什么时候”询问的是时间, 二者意图不一致, 因此被标注为“不匹配”. 基于上述观察, 本文提出了基于关键词的特征提取方法. 关键词提取的具体步骤为: 使用 Jieba 分词器对所有问句进行分词, 使用停用词库过滤掉所有无意义的词汇, 再通过 TF-IDF 方法分别计算问句中每个词汇的重要程度, 在获得了训练集和测试集中所有词汇对应的 TF-IDF 值后, 由大到小进行排序, 按顺序截取 TF-IDF 值较高的前 n 个词汇, 这些词汇即组成该实验数据集的关键词词库.

以例子来说明, 在表 1 的问句对示例中: 问句对在经过分词和排序筛选后, 最终选入词库的关键词为“过年”, “礼物”, “长辈”, “时候”, 其中“过年”, “礼物”和“长辈”在两个句子中共同出现, 因此记为“共现”关键词, 而“时候”仅在 Q2 中出现, 因此记为“单现”关键词. 以上关键词均可建模为因子加入因子图中, 以关键词的出现状态作为一类特征信息. 如同人类语言理解的习惯一样, “共现”关键词特征建模的因子为问句对的“匹配”提供正向支持, 而“单现”关键词特征建模的因子为问句对的“匹配”提供负向支持.

然而, 在实际测试时发现, 使用以上方式提取的关键词特征在因子图推理时能够提供的信息有限. 经过分析, 主要原因为: 因子图中大量的“共现”或“单现”关键词特征仅与个别实例相关, 不同的关键词特征之间完全独立, 信息传递被限制在小范围内, 无法有效实现大范围的知识传递. 为解决上述问题, 本文在上述构建的关键词词库基础上, 通过聚类算法将所有关键词划分为不同的类. 如表 3 所示, 划分标准为: 相同类别中关键词之间尽可能相似, 不同类别之间的关键词尽可能增大差异. 通过聚类操作将单个的关键词转换为词集合, 后续加入因子图中的特征信息由关键词单(共)现状态更换为关键词所属集合单(共)现状态. 在使用集合的类别标签代替关键词的情况下, 如果问句对中的两个问句都包含了同一类别的关键词, 则该实例具有“共现”类特征; 如果问句对中仅有一个问句包含了某一类别的关键词, 则该实例具有“单现”类特征. 因为基于大规模语言模型(如 RoBERTa 等)的聚类方法好于传统的 k-means 和 BIRCH 等, 因此我们在算法实现中利用 RoBERTa 实现聚类.

(2) DNN 特征. 通过关键字无法提取问句包含的深层语义, 而深层语义对判断复杂问句对的匹配至关重要. 因此, 本文利用在自然语言处理取得广泛成功的预训练语言模型(如 ERNIE 等)抽取问句对的深层语义关系, 并借此辅助判断问句对的匹配关系. 先利用深度语言模型获取问句对的高维向量表达, 然后基于向量表达分别提取全局特征和局部特征. 其中, 全局特征主要从总体数据的角度定义全部实例都共享的特征, 局部特征则是从更细粒度的角度定义实例所独有的特征. 基

表 3 关键字特征抽取示例

Table 3 Examples of keyword feature extraction

Category	Keyword clusters from the BQ (Bank question) corpus
1	['微商', '微啦贷', '微代粒会', '微信端', '微信开', '连微', '微信微', '微梨贷', '微众充', '进微众', '微拉能', '截微众', '微车', '微卡', '微银转', '事微众', '微信群', '微信货', '微立', '微镇', '微十星贷', '系微信', '微付', '微丽贷', '微路', '微力袋', '微理贷', '微立款', '微大', '微中']
2	['久等', '几秒钟', '几时', '下个星期', '前几日', '近几个月', '很久很久', '一个多', '几久', '这会儿', '近来', '久久', '有效期', '有效期限', '几钟', '多时', '何时', '好几年']
3	['后会降', '没升', '还会降', '会重', '会降', '会有', '会变']
4	['低价', '优惠活动', '复利', '打折', '还要', '小额', '赢近', '超限', '利后', '双倍', '高利', '特价', '力度', '绝对', '利近', '优惠政策', '利不', '超额']
5	['简介', '货介绍', '声明', '目录', '综合信息', '详细数据', '条例', '文件', '资料', '使用手册', '指引', '公司简介', '介绍']
...	...

于深度语言模型提取的问句全局特征和局部特征如下.

1) 全局特征: 问句对相似度特征. 在问句对匹配的问题中, 相似度度量是关于结果判断最直接有效的特征信息, 也是全部实例都拥有的对齐特征. 全局特征联通因子图中的所有实例, 保证了最基本的渐进知识传递. 本文先使用训练集微调预训练语言模型, 然后从微调后模型的最后一个全连接层获取问句对的向量表征, 输入 Softmax 分类函数计算后取得最终的匹配概率, 即为问句对的相似度特征.

2) 局部特征: 问句对的 k 近邻关系特征(简称关系特征). 深度语言模型通常会将有相同标签的问句对("匹配"或"不匹配")聚在一起, 而把不同标签的问句对尽量分开. 因此, 两个问句对之间最近邻关系意味着它们有很大概率具有相同的标签. 这种关系特征有利于问句对标签的渐进推理.

在具体实现中, 利用 ERNIE 和 Glyce 模型提取问句对的深度向量表征. BERT 模型自 2018 年诞生, 几乎成为所有基于交互的文本匹配深度学习模型中表征效果最优的代名词. ERNIE 和 Glyce 都是 BERT 的升级模型. 通过多次实验, 分别测试 RoBERTa、ERNIE、Glyce、ALBERT[30] 和 XLNet[31] 等预训练模型的文本匹配效果, 发现在相同实验条件下, 对比其他实验模型, ERNIE 和 Glyce 模型应用于本文实验方案的总体效果优异且表现稳定, 因此最终选择使用 ERNIE 和 Glyce 模型完成文本匹配向量的表征提取.

3.2 特征因子建模

特征影响力建模时, 如果特征仅定义于一个实例, 如关键词特征和基于 DNN 的相似度特征, 建模为单因子; 如果特征定义于两个实例, 如基于 DNN 的最近邻关系特征, 建模为双因子.

数学上, 把关键词特征和相似度特征定义为如下的单因子:

$$\phi_{\mathrm{f_u}}(v)=\begin{cases}\mathrm{e}^{w_{\mathrm{f_u}}} & \text{if} \quad v=1\\ 1 & \text{if} \quad v=0\end{cases} \tag{4}$$

其中, $v=0/1$表示待标注变量, 即问句对的匹配状态, 1 为匹配, 0 为不匹配; $w_{\mathrm{f_u}}$表示单因子的权重.

类似地, 我们把最近邻关系特征定义为如下的双因子:

$$\phi_{\mathrm{f_r}}\left(v_i,v_j\right)=\begin{cases}\mathrm{e}^{w_{\mathrm{f_r}}} & \text{if} \quad v_i=v_j\\ 1 & \text{otherwise}\end{cases} \tag{5}$$

其中, $w_{\mathrm{f_r}}$表示双因子的权重. 跟据之前的工作[12], 不管是单因子还是双因子, 都通过单调的 sigmoid 函数刻画特征对实例标签状态的影响力. 数学上, 我们正式定义因子的权重 ($w_{\mathrm{f}}(v)$) 为:

$$w_{\mathrm{f}}(v)=\theta_{\mathrm{f}}(v)\cdot\tau_{\mathrm{f}}\left[x_{\mathrm{f}}(v)-\alpha_{\mathrm{f}}\right] \tag{6}$$

其中, $\theta_{\mathrm{f}}(v)$表示特征影响力的置信度; $x_{\mathrm{f}}(v)$表示一个实例 d 的特征值; τ_{f}和α_{f}表示 sigmoid 函数的拟合参数, 曲线的中值和斜率. 在上述公式中, $\theta_{\mathrm{f}}(v)$通过回归误差直接估计, 参数τ_{f}和α_{f}则需要根据当前的证据实例的状态在渐进学习过程中动态学习.

如图 1 所示, 所有的关键词单现特征共享一个因子, 但不同的实例有不同的特征值; 同样地, 所有的关键词共现也共享一个单因子. 类似地, 所有的相似度特征共享一个单因子, 所有的关系特征共享一个双因子. 每个因子有自己独立的参数, 包括τ_{f}和α_{f}.

4 实验验证

4.1 实验设置

本文实验使用两个公开的中文基准数据集: (1)LCQMC 是哈工大发表的一个面向开放领域的超大规模中文问句匹配数据集, 借助搜索引擎从百度问答中抓取高频问题并整理标注而来. 其包

含 206608 个问句对, 其中训练集包含 238766 个问句对, 验证集包含 8802 个问句对, 测试集包含 12500 个问句对; (2) BQ(Bank question) corpus 是一个面向金融垂直领域的大规模中文问句匹配数据集, 从一个在线银行一年的客户服务日志中标记整理而来, 训练集包含 100000 个问句对, 验证集包含 10000 个问句对, 测试集包含 10000 个问句对.

在对比实验中, 先假定可以利用 100% 训练数据, 把渐进机器学习方法与现有的各种深度学习模型对比. 为了摸拟真实场景, 选取不同比例(10%、20% 和 50%)的训练数据, 对比它们之间的性能差异. 跟之前工作一样, 用两个指标, 准确率和 F1, 来衡量匹配性能.

基于开源的渐进机器学习引擎实现所提出的算法, 分别基于 ERNIE 和 Glyce 抽取相似度和 KNN 关系特征. 在抽取 KNN 关系特征时, 为了保证关系特征的准确率, 最近邻个数建议设置在一个小于 10 的值; 在算法实现中, 本文默认最近邻个数为 6, 即 n=6. 渐进推理参数中, 1) top-m 表示在计算完所有实例的证据支持后, 需要筛选出后续参与近似熵估计的实例数目; 因为测试数据集规模在 1 万 ~ 2 万条, 算法一般选取证据支持度最大的 10% ~ 20% 实例参与后续的近似推理, 默认设置 m=2000; 2) top-k 表示在计算完 top-m 个实例的近似熵后, 后续建立因子图参与真实推理的实例数目; 因为近似推理大多数情况下准确率较高且真实推理比较费时, 算法一般选取少量的实例进行真实概率推理, 默认设置 k=20; 3)算法使用随机梯度下降方法学习参数, 子图参数学习轮数和子图推理轮数均设为 50. 实验显示, 进一步提高参数学习和子图推理轮数对实验结果的影响微乎其微. 具体参数设置情况如表 4 所示.

表 4　渐进机器学习算法参数设置

Table 4　Parameter setting of GML algorithm implementation

Parameters	Default values	Suggested value ranges
n of KNNs in binary/relational factors	6	[4, 8]
m of the top-m evidential support in gradual inference	2000	[1000,3000]
k of the top-k minimum entropy in gradual inference	20	[10, 30]
i, the number of iterations for factor parameter optimization	50	≥50

在 4.4 小节的敏感度实验表明, 所提算法对因子近邻个数 n、最大证据支持度 top-m 和最小近似熵 top-k 等参数的设置不敏感; 只要设置在合理区间, 算法性能保持基本稳定. 在对比实验中, 报告的性能表现都是 5 次运行的平均值, 每次运行之间的性能差异很小.

4.2　对比实验结果

在利用全部训练数据的情况下, 对比实验结果如表 5 和表 6 所示. 可以看出, 在两个数据集上, 基于 BERT 的预训练模型(RoBERTa、ERNIE 和 Glyce)的性能在所有指标上都明显优于之前的深度神经网路模型. 基于 BERT 的不同模型之间, 性能总体上差异不大. 基于渐进机器学习的算法在 F1 和准确率两个指标上都优于基于 BERT 的模型. 以上结果显示, 通过有效融合传统机器学习方法和深度神经网路提取的特征, 渐进机器学习相比深度学习具有稳定的性能优势.

对比渐进机器学习算法与深度学习模型在利用不同比例训练情况下的性能. 由于基于 BERT 的 ERNIE 模型在深度学习模型中相对表现最好, 本文把 GML 与 ERNIE 模型进行比较, 结果如图 2 所示. 可以看出, 在不同比例下, GML 的性能都稳定地优于 ERNIE. 特别需要指出的是, 在两个数据集上, GML 的性能优势随着训练数据比例的减少而增大. 实验结果表明, 相比单纯的深度学习模型, 渐进机器学习算法在训练数据不足的情况下

表 5　LCQMC 数据集上实验对比结果(100% 训练数据)

Table 5　Experimental comparison results on LCQMC (100% training data)

Model	Precision/%	Recall/%	F1/%	Accuracy/%
Text-CNN	69.62	71.84	70.71	70.12
BiLSTM	76.54	73.76	75.12	74.96
BiMPM	80.28	91.18	85.38	83.55
DIIN	79.88	93.46	86.13	85.11
BERT	80.34	95.24	87.16	85.90
RoBERTa	80.34	95.77	87.38	86.17
ERNIE	80.66	96.37	87.83	86.56
Glyce	80.88	95.66	87.65	86.38
GML	81.67	97.42	88.89	88.76

表 6 BQ corpus 数据集上实验对比结果(100% 训练数据)

Table 6 Experimental comparison results on the BQ corpus (100% training data)

Model	Precision/%	Recall/%	F1/%	Accuracy/%
Text-CNN	67.77	70.64	69.17	68.52
BiLSTM	75.04	70.46	72.68	73.51
BiMPM	82.28	81.18	81.73	81.85
DIIN	81.58	81.14	81.36	81.41
BERT	82.98	85.36	84.24	84.17
RoBERTa	83.65	85.24	84.44	84.29
ERNIE	83.88	86.75	85.30	84.79
Glyce	85.76	84.45	85.09	84.46
GML	86.64	87.82	87.20	87.06

表现更健壮, 性能有稳定的优势.

4.3 消融实验结果

在消融实验中, 本文在渐进机器学习算法的实现中分别去掉一种特征, 然后比较其与原来利用全部特征的算法的性能差异. 详细的对比结果如表 7 所示. 可以明显看出, 去掉任何一种特征, GML 的性能都有不同程度的下降. 其中, 相比关键词特征, 去掉基于 DNN 的相似度特征或 KNN 特征, 性能下降的更为明显; 但是, 关键词特征仍然能在一定程度上提升性能. 实验结果显示, 本文提出的方案中, 各种特征之间具有一定的互补性, 综合建模能有效提升问句匹配的准确率.

4.4 参数敏感度测试实验

在参数敏感度测试实验中, 选取渐进推理算法的三个关键参数, 即关系因子最近邻个数 n、最大证据支持度个数 top-m 和最小近似熵个数 top-k, 测试 GML 的性能对这三个参数设置的敏感度. 之前的实验中, KNN-n、top-m 和 top-k 的默认取值为 6、2000 和 20; 在敏感度测试实验中, 如表 4 所示, KNN-n 的取值范围为 [4, 8]、top-m 的取值范围为 [1000, 300]、top-k 的取值范围为 [10, 30], 测试 GML 的性能变化. LCMQC 数据集上的测试结果如表 8 ~ 10 所示, 在 BQ corpus 数据集上的测试结果类似, 因此在此忽略.

表 9 和表 10 的实验结果显示, 随着 top-m 和 top-k 参数取值的变化, GML 算法的性能只是小幅波动. 敏感度实验结果表明, GML 的算法性能对参数不敏感, 证明了其表现的健壮性.

5 结论及展望

本文提出了一种基于渐进机器学习的中文问句匹配方法. 方法抽取浅层和深层语义特征, 通过因子模型实现从易到难的渐进学习. 在基准数据集上的实验表明: 1)本文所提的渐进机器学习方

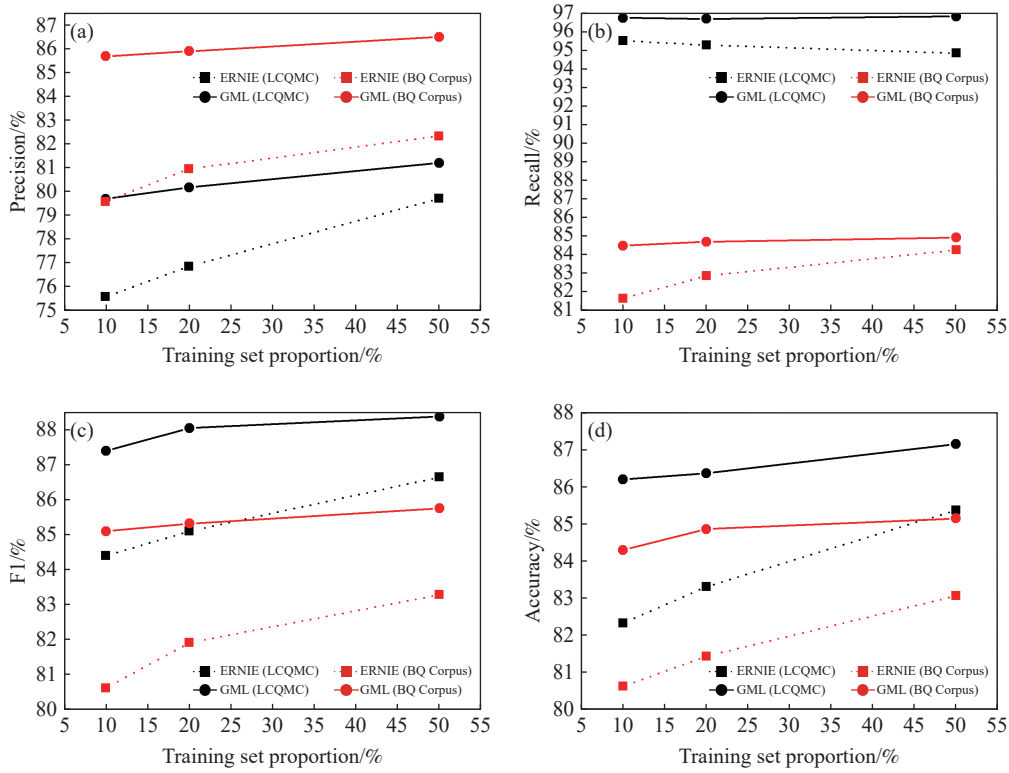

图 2 不同训练数据比例下的实验对比结果 (10%/20%/50%). (a)查准率; (b)召回率; (c)F1 值; (d)准确率

Fig.2 Experimental comparison results under different training data ratios (10%/20%/50%): (a) precision; (b) recall; (c) F1; (d) accuracy

表 7　消融实验测试结果

Table 7　Test results of ablation experiments

Dataset	Factor combination method	Precision/%	Recall/%	F1/%	Accuracy/%
LCQMC	All-factor	81.67	97.42	88.89	88.76
	Without keyword	81.34	96.10	88.12	87.80
	Without Similarity	79.30	93.42	85.78	84.85
	Without KNN double factor	80.56	94.60	87.02	85.42
BQ corpus	All-factor	86.64	87.82	87.20	87.06
	Without keyword	85.54	87.10	86.33	86.20
	Without Similarity	82.40	83.82	83.10	82.95
	Without KNN double factor	83.68	84.58	84.12	84.16

表 8　KNN-n 参数敏感度测试实验结果(LCQMC, top-m=2000 & top-k=20)

Table 8　Experimental results of sensitivity testing for the top-m parameters (LCQMC, top-k = 20)

KNN-n	Precision/%	Recall/%	F1/%	Accuracy/%
4	80.82	96.80	88.09	87.82
6	81.67	97.42	88.89	88.76
8	81.88	97.26	88.91	88.82

表 9　top-m 参数敏感度测试实验结果(LCQMC, n=6 & top-k=20)

Table 9　Experimental results of sensitivity testing for the top-m parameters (LCQMC, top-k = 20)

top-m	Precision/%	Recall/%	F1/%	Accuracy/%
1000	81.94	96.56	88.22	88.57
2000	81.67	97.42	88.89	88.76
3000	80.87	97.88	88.17	88.32

表 10　top-k 参数敏感度测试实验结果 (LCQMC, n=6 & top-m=2000)

Table 10　Experimental results of sensitivity testing for the top-k parameters (LCQMC, top-m = 2000)

top-k	Precision/%	Recall/%	F1/%	Accuracy/%
10	80.34	97.86	88.33	88.27
20	81.67	97.42	88.89	88.76
30	81.53	97.18	88.78	88.35

案能够取得比目前微调预训练语言模型的主流方案更好的性能, 且性能提升的幅度随着训练数据的减少有所提升, 测试数据和目标数据之间分布有差异, 渐进机器学习的算法能够在一定程度上有效克服这种差异. 随着训练数据的减少, 训练数据对于测试数据的分布代表性就越弱, 预训练语言模型的性能随之下降; 而渐进机器学习的算法因为本身就是为分布差异的场景而设计的, 所以其表现相对更优异, 其性能虽然也有相应下降但幅度相对较小; 2)在渐进机器学习算法里同时融合文本统计模型和预训练模型提取的特征, 相比只利用它们之中任何一种特征, 能取得更好的性能表现, 对于中文问句匹配问题而言, 目前的预训练语言模型虽然相比传统的统计语言模型有明显的性能优势, 但是并不能完全刻画和描述传统语言模型能够抽取的信息; 因此, 就语义表达而言, 大规模预训练语言模型和传统的文本统计模型仍具有一定的互补性, 融合这两种模型能够有效提升问句语义匹配的精度; 3)本文所提方案的性能对于渐进机器学习算法的关键参数不敏感, 只要把它们设置在合理范围, 性能很稳定. 渐进机器学习算法的参数鲁棒性, 增强了本文方案在真实场景的落地可行性.

展望未来研究方向: 1)目前本文方案利用现有的预训练语言模型抽取深度语义特征, 但这些模型都是通过基于独立同分布假设的深度学习框架来微调; 虽然实验表明将其抽取的特征直接应用于渐进推理上有一定的效果, 但这种抽取特征的方式仍有待优化提升. 如何为不基于独立同分布假设的渐进机器学习设计一种专门的深度神经网络结构以及如何训练它以便于提取特征是未来值得研究的方向之一; 2)随着大规模预训练语言模型的发展, 越来越多的自然语言处理任务只需很少的标注数据即可精准完成. 虽然本文提出的渐进机器学习算法相比于微调预训练语言模型有一定的性能优势; 但其性能表现仍然很大程度上依赖于预训练语言模型, 其在问句匹配问题上的表现又在很大程度上依赖一定数量的标注数据. 如何在只有很少甚至没有任何标注数据情况下设计渐进机器学习算法以实现准确的渐进问句匹配, 是另一个未来值得研究的问题; 3)目前的自然语言处理大模型, 如英文的 ChatGPT、LlaMa 和中

文的文心一言、混元大模型等, 在很多下游的自然语言处理任务上表现出不错的性能, 如何利用这些大模型来进一步提升渐进机器学习算法的性能也是一个非常有趣的研究方向.

参 考 文 献

[1] Kwok C C T, Etzioni O, Weld D S. Scaling question answering to the Web // *Proceedings of the* 10*th International Conference on World Wide Web*. Hong Kong, 2001: 150

[2] Wang Q, Qian L. Analysis on the development trend of personalized recommendation service in e-commerce under big data environment. *Commer Res*, 2014(8): 150
(王茜, 钱力. 大数据环境下电子商务个性化推荐服务发展动向探析. 商业研究, 2014(8): 150)

[3] Sultana T, Badugu S. A review on different question answering system approaches // *Advances in Decision Sciences, Image Processing, Security and Computer Vision*. Hyderabad, 2019: 579

[4] Juan Z M. An effective similarity measurement for FAQ question answering system // 2010 *International Conference on Electrical and Control Engineering*. Wuhan, 2010: 4638

[5] He B, Huang J X, Zhou X F. Modeling term proximity for probabilistic information retrieval models. *Inf Sci*, 2011, 181(14): 3017

[6] Eddington C M, Tokowicz N. How meaning similarity influences ambiguous word processing: The current state of the literature. *Psychonomic Bull Rev*, 2015, 22(1): 13

[7] Mikolov T, Chen K, Corrado G, et al. Efficient estimation of word representations in vector space [J/OL]. *arXiv preprint* (2013–01–16) [2023–11–05]. https://arxiv.org/abs/1301.3781

[8] Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need // *Proceedings of the* 31*st International Conference on Neural Information Processing Systems*. Long Beach, 2017: 6000

[9] Liu Y H, Ott M, Goyal N, et al. RoBERTa: A robustly optimized BERT pretraining approach [J/OL]. *arXiv preprint* (2019–07–26) [2023–11–05]. https://arxiv.org/abs/1907.11692

[10] Kenter T, de Rijke M. Short text similarity with word embeddings // *Proceedings of the* 24*th ACM International on Conference on Information and Knowledge Management*. Melbourne, 2015: 1411

[11] Yu C C. *Research and Improvement of Question Matching Method in Question Answering System Based on Word Vector* [Dissertation]. Harbin: Harbin Institute of Technology, 2020
(于翠翠. 基于词向量的问答系统中问句匹配方法研究与改进[学位论文]. 哈尔滨: 哈尔滨工业大学, 2020)

[12] Hou B Y, Chen Q, Wang Y Y, et al. Gradual machine learning for entity resolution. *IEEE Trans Knowl Data Eng*, 2022, 34(4): 1803

[13] Wang Y Y, Chen Q, Shen J Q, et al. Gradual machine learning for aspect-level sentiment analysis [J/OL]. *arXiv preprin*t (2019–06–06) [2023–11–05]. https://arxiv.org/abs/1906.02502

[14] Banerjee S S, El-Hadedy M, Bin Lim J, et al. ASAP: Accelerated short-read alignment on programmable hardware. *IEEE Trans Comput*, 2019, 68(3): 331

[15] Chidananda H T, Das D, Sagnika S. Sentiment analysis using N-gram technique // *Proceedings of International Conference on Computing Analytics and Networking* (*ICCAN*) 2017. Bhubaneswar, 2017: 359

[16] Real R, Vargas J M. The probabilistic basis of jaccard's index of similarity. *Syst Biol*, 1996, 45(3): 380

[17] Wu H C, Luk R W P, Wong K F, et al. Interpreting TF-IDF term weights as making relevance decisions. *ACM Trans Inf Syst*, 2008, 26(3): 1

[18] Hinton G E. Learning distributed representations of concepts // *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*. Amherst, 1986: 46

[19] Bengio Y, Schwenk H, Jean-Sébastien Senécal, et al. A neural probabilistic language model. *J Mach Learn Res*, 2003(3): 1137

[20] Le Q, Mikolov T. Distributed representations of sentences and documents // *Proceedings of the* 31*st International Conference on International Conference on Machine Learning*. Beijing, 2014: 1188

[21] Bojanowski P, Grave E, Joulin A, et al. Enriching word vectors with subword information. *Trans Assoc Comput Ling*, 2017, 5: 135

[22] Hochreiter S, Schmidhuber J. Long short-term memory. *Neural Comput*, 1997, 9(8): 1735

[23] Greff K, Srivastava R K, Koutnik J, et al. LSTM: A search space odyssey. *IEEE Trans Neural Networks Learn Syst*, 2017, 28(10): 2222

[24] Mikolov T, Karafiát M, Burget L, et al. Recurrent neural network based language model // *Proceedings of the* 11*th Annual Conference of the International Speech Communication*. Chiba, 2010: 1045

[25] Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modelling // *Proceedings of the* 13*th Annual Conference of the International Speech Communication Association*. Portland, 2012: 194

[26] Schuster M, Paliwal K K. Bidirectional recurrent neural networks. *IEEE Trans Signal Process*, 1997, 45(11): 2673

[27] Siami-Namini S, Tavakoli N, Namin A S. The performance of LSTM and BiLSTM in forecasting time series // 2019 *IEEE International Conference on Big Data* (*Big Data*). Los Angeles, 2019: 3285

[28] Devlin J, Chang M W, Lee K, et al. BERT: Pre-training of deep bidirectional transformers for language understanding *arXiv preprint* (2018–10–11) [2023–11–05]. https://arxiv.org/abs/1810.04805

[29] Wang Z G, Ng P, Ma X F, et al. Multi-passage bert: A globally normalized bert model for open-domain question answering// *Proceedings of the* 2019 *Conference on Empirical Methods in Natural Language Processing and the* 9*th International Joint Conference on Natural Language Processing* (*EMNLP-IJCNLP*). Hong Kong, 2019: 5878

[30] Lan Z Z, Chen M D, Goodman S, et al. ALBERT: A lite BERT for

self-supervised learning of language representations [J/OL]. *arXiv preprint* (2019–9–26) [2023–11–05]. https://arxiv.org/abs/1909.11942

[31] Yang Z L, Dai Z H, Yang Y M, et al. XLNet: Generalized autoregressive pretraining for language understanding // *Proceedings of the* 33*rd International Conference on Neural Information Processing Systems*. Vancouver, 2019: 5753

[32] Cui Y M, Che W X, Liu T, et al. Pre-training with whole word masking forhinesee BERT [J/OL]. *arXiv preprint* (2019–01–19) [2023–11–05]. https://arxiv.org/abs/1906.08101

[33] Sun Y. *Chinese Sentence Similarity Calculation Based on Convolutional Neural Network* [Dissertation]. Hefei: University of Science and Technology of China, 2019
(孙阳. 基于卷积神经网络的中文句子相似度计算[学位论文]. 合肥: 中国科学技术大学, 2019)

[34] Chen Y M, Chen E H, Zhang K, et al. A relation-aware representation approach for the question matching system. *World Wide Web*, 2024, 27(2): art No. 17

[35] Chen Z M, Huo Y. Optimization and implementation of the edit distance algorithm in Chinese similarity calculation. *J Shaoguan Univ*, 2015, 36(12): 8
(陈正铭, 霍英. 编辑距离算法在中文文本相似度计算中的优化与实现. 韶关学院学报, 2015, 36(12): 8)

[36] Zhou L J, Yu W H, Guo C. Research on text similarity algorithm based on improved TF-IDF strategy. *J Taishan Univ*, 2015, 37(3): 18
(周丽杰, 于伟海, 郭成. 基于改进的 TF-IDF 方法的文本相似度算法研究. 泰山学院学报, 2015, 37(3): 18)

[37] Wan S X, Lan Y Y, Guo J F, et al. A deep architecture for semantic matching with multiple positional sentence representations // *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. Phoenix, 2016: 2835

[38] Liu P F, Qiu X P, Chen J F, et al. Deep fusion LSTMs for text semantic matching // *Proceedings of the* 54*th Annual Meeting of the Association for Computational Linguistics* (*Volume* 1: *Long Papers*). Berlin, 2016: 1034

[39] Sun Y, Wang S H, Li Y K, et al. ERNIE: Enhanced representation through knowledge integration [J/OL]. *arXiv preprint* (2019–04–19) [2023–11–05]. https://arxiv.org/abs/1904.09223

[40] Meng Y X, Wu W, Wang F, et al. Glyce: Glyph-vectors for chinese character representations [J/OL]. *arXiv preprint* (2019–01–29) [2023–11–05]. https://arxiv.org/abs/1901.10125

[41] Huang S Z, Wu Y M, Lu J W, et al. Related questions detection model in stack overflow based on semantic matching // 35*th International Conference on Software Engineering & Knowledge Engineering*. San Francisco, 2023: 339

[42] Ying Y Y, Zhang Z Q, Wu H Y, et al. Er-EIR: A chinese question matching model based on word-level and sentence-level interaction features // *Chinese Conference on Computer Supported Cooperative Work and Social Computing*. Harbin, 2023: 108

[43] Faseeh M, Ali Khan M, Iqbal N, et al. Enhancing user experience on Q&A platforms: Measuring text similarity based on hybrid CNN-LSTM model for efficient duplicate question detection. *IEEE Access*, 2024, 12: 34512

[44] Zhao Y X, Li R, Li X J, et al. A text semantic matching model with Chinese Characters'Glyph, pinyin and sense-based multi-knowledge and label embedding. *J Chin Inf Process*, 2024, 38(3): 42
(赵云肖, 李茹, 李欣杰, 等. 基于汉字形音义多元知识和标签嵌入的文本语义匹配模型. 中文信息学报, 2024, 38(3): 42)

[45] Guo Y, Liang T, Chen Z, et al. FinKENet: A novel financial knowledge enhanced network for financial question matching. *Entropy*, 2023, 26(1): 26

[46] Xu R Q. Medical question answering system integrating knowledge graph and semantic matching. *Mod Electron Tech*, 2024, 47(8): 49
(徐若卿. 融合知识图谱和语义匹配的医疗问答系统. 现代电子技术, 2024, 47(8): 49)

[47] Zhang J A, Ren W, Wang S G. Tourism question identification via multiple similarity attentions. *J Chin Inf Process*, 2023, 37(6): 157
(张劲桉, 任伟, 王素格. 融合多相似度注意的神经网络旅游问题识别方法. 中文信息学报, 2023, 37(6): 157)

[48] Ahmed M, Chen Q, Wang Y Y, et al. DNN-driven gradual machine learning for aspect-term sentiment analysis // *Findings of the Association for Computational Linguistics: ACL-IJCNLP* 2021. Online, 2021: 488

[49] Wang Y Y, Chen Q, Ahmed M H M, et al. Supervised gradual machine learning for aspect-term sentiment analysis. *Trans Assoc Comput Ling*, 2023, 11: 723

[50] Hou B Y, Chen Q, Shen J Q, et al. Gradual machine learning for entity resolution // *The World Wide Web Conference* 2019. San Francisco, 2019: 3526

[51] Zhong P, Li Z H, Chen Q, et al. Attention-enhanced gradual machine learning for entity resolution. *IEEE Intell Syst*, 2021, 36(6): 71

[52] Chen N, Kuang X M, Liu F Y, et al. Few-shot image classification based on gradual machine learning. *Expert Syst Appl*, 2024, 255: 124676