

文章编号: 1000-5641(2019)05-0074-11

基于社区问答数据迁移学习的 FAQ 问答模型研究

邵明锐, 马登豪, 陈跃国, 覃雄派, 杜小勇

(中国人民大学 信息学院, 北京 100872)

摘要: 基于FAQ(Frequent Asked Questions)问答技术构建智能客服系统, 是当前业界普遍采用的技术方案. 基于 FAQ 构建的问答系统, 其返回的结果具有稳定、可靠、质量高的优点; 但因受限于人工标注的知识库规模, 识别能力有限, 容易遇到瓶颈. 为了解决 FAQ 数据集规模有限的问题, 给出了数据层面和模型层面的解决方法: 在数据层面, 利用百度知道爬取相关数据并挖掘语义等价问题, 保证了数据的相关性和一致性; 在模型层面, 提出了一种面向迁移学习的深度神经网络 transAT, 该模型融合了 Transformer 强大的特征抽取能力和注意力机制, 适用于句子对之间的语义相似度计算. 实验表明, 该方法可以显著提升模型在 FAQ 问答任务中的效果, 在一定程度上解决了 FAQ 数据集规模有限的问题.

关键词: 迁移学习; 深度神经网络; FAQ 问答

中图分类号: TP391 **文献标志码:** A **DOI:** 10.3969/j.issn.1000-5641.2019.05.006

Transfer learning based QA model of FAQ using CQA data

SHAO Ming-rui, MA Deng-hao, CHEN Yue-guo, QIN Xiong-pai, DU Xiao-yong
(Information college, Renmin University of China, Beijing 100872, China)

Abstract: Building an intelligent customer service system based on FAQ (frequent asked questions) is a technique commonly used in industry. Question answering systems based on FAQ offer numerous advantages including stability, reliability, and quality. However, given the practical limitations of scaling a manually annotated knowledge base, models often have limited recognition ability and can easily encounter bottlenecks. In order to address the problem of limited scale with FAQ datasets, this paper offers a solution at both the data level and the model level. At the data level, we use Baidu Knows to crawl relevant data and mine semantically equivalent questions, ensuring the relevance and consistency of the data. At the model level, we propose a deep neural network with transAT oriented transfer learning, which combines a transformer network and an attention network, and is suitable for semantic similarity calculations between sentence pairs. Experiments show that the proposed solution can significantly improve the impact of the model on FAQ

收稿日期: 2019-07-27

基金项目: 国家自然科学基金(U1711261, 61432006)

第一作者: 邵明锐, 男, 硕士研究生, 研究方向为自然语言处理与语义搜索.

E-mail: dhucstsmr@163.com.

通信作者: 陈跃国, 男, 教授, 博士生导师, 研究方向为语义搜索与知识图谱.

E-mail: chenyeuguo@ruc.edu.cn.

datasets and to a certain extent resolve the issues with the limited scale of FAQ datasets.

Keywords: transfer learning; deep neural network; FAQ(frequent asked questions) question-answering

0 引言

目前企业问答系统常见的有基于检索式的、基于生成式的、混合式的这3种方法. 给定一个问题, 基于检索的方法会从问答知识库中检索到最相似的问题并返回与之对应的答案; 基于生成的方法, 会通过训练的生成模型, 生成对应的答案; 而混合的方法, 会将基于检索式的、基于生成式的有效地结合起来. FAQ常问问题集, 通常由用户提出的历史问题标注得到. 基于FAQ的问答是检索式问答的重要组成部分, 其返回的结果具有稳定、可靠、质量高的特点. 这种技术方法得到了广泛应用, 如电商、打车、外卖等互联网服务业的客服系统, 银行、通信、医疗等传统行业的信息服务系统, 以及农作物病虫害防治、法律知识普及等经济社会的各个领域.

基于FAQ的问答首先需要人工构建高质量的知识库, 这通常由企业资深的客服人员通过挖掘客服日志标注, 或由各个行业的专家在整理大量的行业资料后得到. 当接收到用户提出的问题, 只需要从知识库中找到与之最相似的问题, 并将其对应的标准答案返回给用户即可. 因此, 基于FAQ问答的技术关键是如何从数据集中找到与当前问题最为相似的问题, 也就是如何计算两个句子之间的语义相似度. 目前主流的方法是采用深度学习技术. 但是由于FAQ需要依赖于人工标注的知识库, 所以其数据规模通常比较小, 在这种情况下, 进行语义计算的深度神经网络会陷入过拟合或遭遇性能瓶颈. 为了提升FAQ问答模型的效果, 减少FAQ数据集训练语料不足的问题, 本文提出了一种迁移学习方法, 主要贡献如下.

(1) 提出了一种面向迁移学习的语义相似度计算的深度神经网络 transAT, 该模型可以适用于不同规模的数据集, 有效地处理FAQ问答任务.

(2) 利用社区问答数据作为源域数据集, FAQ数据集作为目标域数据集, 基于数据层面和模型层面设计并实现了可行的迁移学习方案, 减少了源域数据集中噪声的影响.

本文第1节阐述基于FAQ问答的国内外研究现状; 第2节定义FAQ问答的问题本质, 并且介绍迁移学习数据层面的方案; 第3节介绍迁移学习模型层面的深度语义网络模型——FAQ问答模型; 第4节通过与其他参照模型的实验对比, 证明本文提出的 transAT 模型和迁移学习方法的有效性; 第5节总结本文工作并展望未来工作.

1 国内外研究现状

基于FAQ问答技术的关键是如何计算文本之间的相似度, 从该点出发可以将现有的工作分为3个阶段: ①基于传统信息检索的方法; ②基于机器学习的方法; ③基于深度学习的方法. 下面具体加以说明.

信息检索领域的文本相似度计算以向量空间模型^[1]为代表, 它首先把文本表示为向量空间中的向量, 通过向量间的距离表示文本相似度, 其中文本特征权重的计算包括TF-IDF(Term Frequency-Inverse Document Frequency)、BM(Best Matching)^{25[2]}, 以及TF-IDF基于Ngram的变种等模型. 向量空间模型可以有效捕捉浅层的语义特征, 但是对于词序、句法结构、深度语义却无法有效表征. 文献[3]应用集合中的词频去改善BM25中TF-IDF的值, 与其他方法相比, 这种方法有着更好的效果.

基于机器学习的方法是将文本相似度计算转变为分类问题、回归问题或者排序问题. 人工定义的特征工程是传统机器学习方法的核心. 除了 TF-IDF 值外, 还会深入挖掘文本的长度、词序的变化、句法描述等方面的特征. 文献 [4] 通过定义词对齐特征、问题的文法特征, 应用 learn-to-rank 方法基于 FAQ 训练了一个排序模型.

基于深度学习的方法是当前业界的主流方法, 它在文本相似度计算领域取得了显著的成果. 从预训练的词向量 Word2vec^[5]、Glove^[6], 到深度语义特征提取器 CNN(Convolutional Neural Networks)^[7]、RNN(Recurrent Neural Network)^[8]、LSTM(Long-Short Term Memory)^[9]、Transformer^[10], 再到迁移学习等理论在 NLP(Natural Language Processing)领域的应用^[11-13], 在大数据量的情况下, 深度学习已被证实超过信息检索、传统机器学习技术在 NLP 域的表现. PWIM^[14](Pairwise Word Interaction Modeling)充分融合了 LSTM、注意力机制和 DCNN(Deep CNN)的特性, 在包括 WikiQA 等多个 NLP 任务上都显著提升了效果. BERT^[12](Bidirectional Encoder Representations from Transformers)通过在大规模通用语料上预训练, 然后到具体的 NLP 任务中做微调, 就有超过 11 项 NLP 任务中的最佳效果. 文献 [13] 中提出了一种混合模型框架, 可以学习到从源域数据空间到目标域数据空间共有的特征表示和语义关联, 通过迁移学习的方法提升 FAQ 任务的效果, 模型被应用到 AliExpress 在线系统, 线上效果得到了显著提升. 文献 [14] 尝试不同的迁移学习任务, 发现 NLI(Natural Language Inference)任务更适合迁移到其他 NLP 任务中. 文献 [15] 中通过在答案提取阶段改进迁移学习方法从而提升了传统问答系统的效果.

通过分析以上工作, 本文发现深度学习面临的主要问题在于, 基于深度神经网络的方法缺乏足够规模的标注数据, 且相应的迁移学习方法需要有与目标域数据分布相似的一定规模的源域数据集. 基于此, 本文提出了利用社区问答数据作为源域数据集, 运用迁移学习的方法提升 FAQ 问答模型的效果; 为降低噪声的影响, 分别在数据层面和模型层面做了相应的设计.

2 问题定义和数据层面设计

2.1 问题定义

借鉴前人的工作, 本文将基于 FAQ 的问答任务转变为 PI(Paraphrase Identification)任务^[16], 即语义等价任务, 判断两个句子是否语义等价.

给定集合 $Q = \{((S_1, T_1), y_1), ((S_2, T_2), y_2), \dots, ((S_n, T_n), y_n)\}$, 目标是

$$\arg \max_{\theta} = \left\{ \sum_{(S_i, T_i, y_i) \in Q} \log(p(y_i | S_i, T_i; \theta)) \right\},$$

其中, S_i 代表 Q 中标准问题, T_i 代表 Q 中的扩展问题, $y_i \in 0, 1$ 是 S_i 和 T_i 语义等价的标签, 最大化的目标是已知 S_i, T_i , 分类正确的对数几率, 模型的参数 θ 通过语义等价问题集的极大似然估计进行优化. 首先, 需要构造数据集, 本文的目标域数据集是基于问答日志人工标注的语义等价数据集, 源域数据集从百度知道爬取并且整理得来. 为了避免在构建源域数据集过程中引入过多的噪声数据, 本文在数据层面给予了针对性的设计.

2.2 数据层面设计

首先, 以目标域数据集作为种子, 从百度知道爬取相关的问答对, 然后进行语义等价问题挖掘工作. 在此之前, 首先需要定义语义等价问题的标准. 在问答系统中, 答案需要满足用

户的信息需求, 而问题反映了用户的查询意图. 因此, 本文采用以下标准: 如果两个问题可以被相同的答案回答, 那么认为这两个问题是语义等价的^[17]. 在具体量化时, 首先考虑答案的相似性. 但因为答案来源于网络社区, 可能质量不高, 所以, 会在计算时加上查询相似性. 整个计算过程可以定义为, 给定问答对集合 $M = \{(q_1, a_1), (q_2, a_2), \dots, (q_n, a_n)\}$, 两个问题之间的语义相关性为

$$\text{sim}(q_i, q_j) = \alpha R(q_i, q_j) + (1 - \alpha) R(a_i, a_j), \quad (1)$$

$$R(q_i, q_j) = \gamma \text{BM25}(q_i, q_j) + (1 - \gamma) \cos \langle W(q_i), W(q_j) \rangle, \quad (2)$$

其中, α 是权重系数, 用来调节查询的文本相似度和答案文本相似度在结果中的占比; R 代表文本相似度计算函数, 这个函数包含两个部分, 一部分是用 BM25 模型计算的结果, 另一部分是基于 Word2vec 词向量表示的句向量, 计算余弦距离得到的结果; γ 是相应的权重系数, 用来调节不同模型在结果中的占比. 假设 $q_i = \{w_1^{(i)}, w_2^{(i)}, \dots, w_n^{(i)}\}$, 共由 n 个词组成, 则

$$W(q_i) = \sum_{k=1}^n h_k^{(i)}, \quad (3)$$

其中, $h_k^{(i)} \in \mathbf{R}^d$ 是 $w_k^{(i)}$ 根据训练的 Word2vec 模型得到词向量, d 是词向量的维度, 此处 $d = 200$. 通过以上方法, 可以计算出问答集 M 中任意两个问题的语义相关分数. 然后设定分数阈值, 就可以得到语义等价问题集. 最后, 通过人工随机抽检 n 个问题进行复核, 不断调整阈值的大小和相关的参数. 经过多次复核操作, 确定最优的阈值和参数. 最终, α 的值取为 0.6, 因为答案往往文本过长, 噪音过大, 所以问题的质量相对高一些; γ 的取值为 0.7, 受限于词向量的质量, BM25 计算的结果会更可信, 所以阈值也更高一些.

3 FAQ 问答模型设计

3.1 语义计算网络

在第 2.1 节中, 本文将基于 FAQ 的问答任务转变为 NLP 领域中的句子分类任务. 因此, 语义计算网络的输入是一个句子对, 输出则是两个句子文本是否语义等价的标签. 借鉴前人的网络设计理念, 以及 BERT 的设计思想, 本文提出了一种新的神经网络结构 transAT. transAT 整个网络被分为 4 层, 分别是输入层、嵌入层、注意力层、输出层, 具体如图 1 所示.

输入层主要负责文本的编码工作, 输入的内容包括文本信息和位置信息这两部分, 这是由嵌入层采用的 Transformer 神经网络结构决定的. 在嵌入层, Transformer 分别对文本信息和位置信息做嵌入和编码的操作, 然后经过 Transformer 编码器获得句子的深度语义特征. 经过嵌入层后, 每个词被映射到语义空间中长度为 l 的向量. 假设输入的句子为 $S = \{s_1, s_2, \dots, s_n\}$ 和 $T = \{t_1, t_2, \dots, t_m\}$, 其中, s_i 和 t_j 分别代表组成句子的单词. 在经过嵌入层后, $E_S = \{E_{s_1}, E_{s_2}, \dots, E_{s_n}\}$, $E_T = \{E_{t_1}, E_{t_2}, \dots, E_{t_m}\}$, $E_{s_i}, E_{t_j} \in \mathbf{R}^l$, E_S 和 E_T 分别为 S 和 T 对应的嵌入向量集合. 在注意力层中, 希望得到每个单词相对于另一个句子的重要程度, 从而得到对应的特征表征. 因此, 此处需要由词向量得到句向量, 常见的操作有 pooling、concatenate、sum 等操作, 此处采用 sum, 即用句子对应词向量的和 H_1 表示句向量, H_1 的计算公式为

$$H_1 = \sum_{i=1}^n E_{s_i}. \quad (4)$$

H_2 的计算方式与 H_1 相同. 式 (4) 中, H_1 为 S 所对应的句向量, 同理, H_2 为 T 所对应的句向量. 在注意力层采用 scale dot attention, 分别计算两句话相应的权重和表征, 即

$$e_{s_i} = E_{s_i} \cdot H_2, \quad (5)$$

$$\alpha_{s_i} = \frac{e_{s_i}}{\sum_{j=1}^n e_{s_j}}, \quad (6)$$

$$H'_2 = \sum_{i=1}^n \alpha_{s_i} e_{s_i}, \quad (7)$$

同理可得 H'_1 . H'_1 和 H'_2 分别是 S 和 T 注意力机制得到的句向量. 输出层采用余弦距离计算文本之间的相似度, 损失函数采用 MSE, 计算公式分别为

$$y^p = \text{mean}(\cos\langle H'_1, H'_2 \rangle + \cos\langle H_1, H_2 \rangle), \quad (8)$$

$$\text{Loss} = \frac{1}{k} \cdot \sum_{i=1}^k (y_i - y_i^p)^2. \quad (9)$$

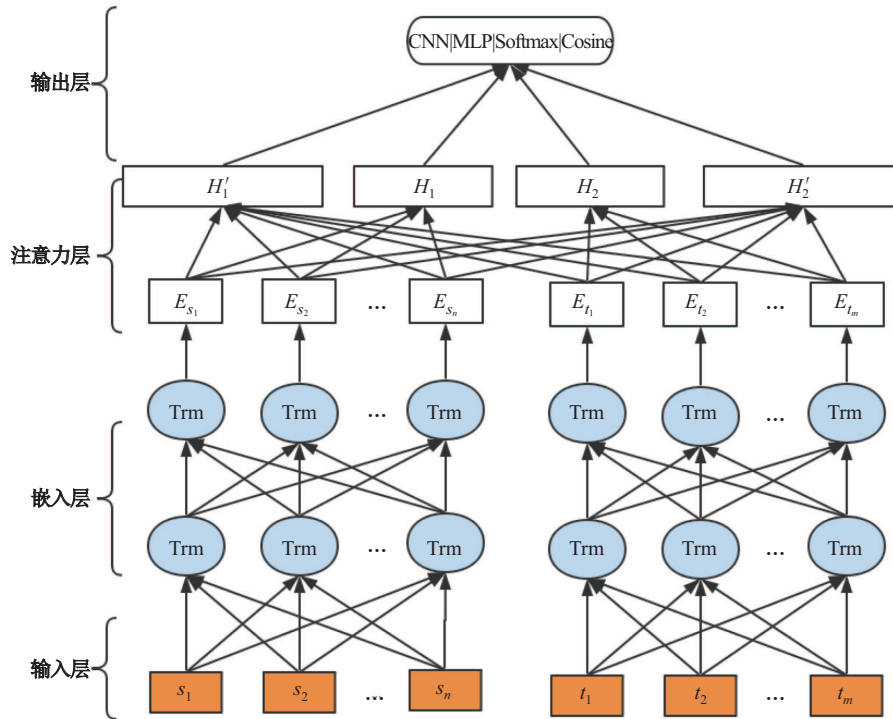


图1 语义计算神经网络结构

Fig. 1 Structure of a semantic computing neural network

3.2 迁移学习策略

针对“预训练+微调”这种迁移学习方式, 常用的迁移学习策略有局部微调和全局微调这两种策略. 本文分别应用这两种迁移学习策略, 针对 transAT 神经网络设计了相应的迁移学习方案, 并在实验中加以对比.

3.2.1 局部微调

通常随着神经网络层数的递增, 所捕获到的特征也将更抽象、更高级. 底层的神经网络所学习到的通常是低级特征, 而底层特征的复用性相比高级特征来说是更强的. 因此, 在局部微调时, 往往会复用整个神经网络模型底层的参数, 而对顶层的参数进行微调. 在微调时, 通常的做法是, 对网络层的模型参数做继续训练, 使模型参数可以自适应目标域数据集的特点. 另一种做法是, 对模型参数重新初始化进行训练, 这种方式可以有效屏蔽来自源域数据集噪声的影响, 加快模型的收敛速度.

针对本文的 transAT 神经网络模型, 冻结 Transformer 层及以下部分的模型参数, 对其以上部分做微调操作, 如图 2 所示. 尝试不同的微调方法, 对比实验结果, 最后确定最优的迁移学习方式.

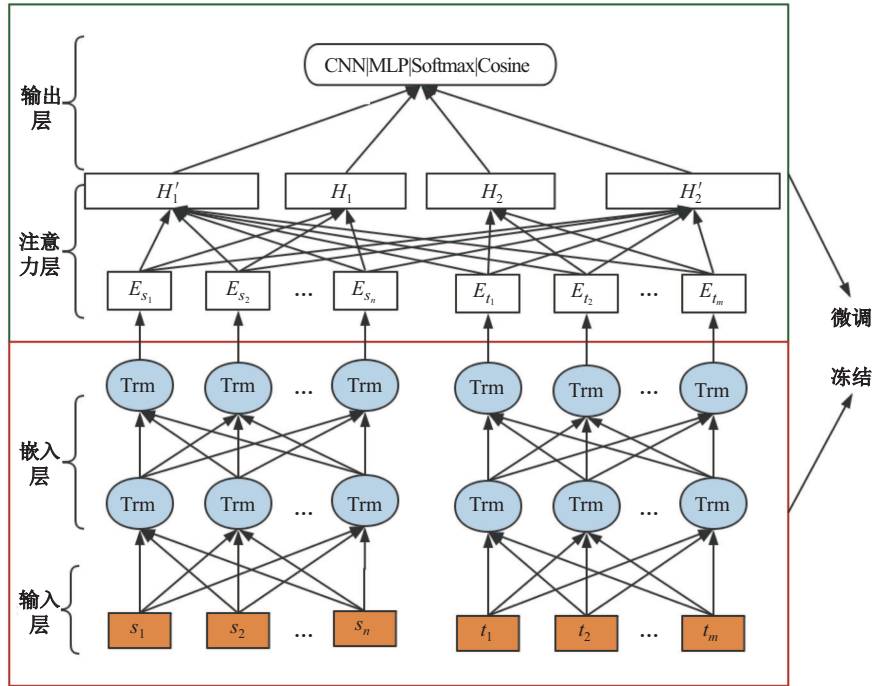


图2 transAT神经网络局部微调方案

Fig. 2 Local fine-tuning scheme of a transAT neural network

3.2.2 全局微调

与局部微调相比, 全局微调会在微调时, 修改整个神经网络的全部参数, 使得神经网络可以更好地适应目标域数据集. 考虑到输出层对训练模型的影响, 本文针对 transAT 神经网络模型设计了 3 种全局微调的迁移学习方案: ① transAT 模型在预训练和迁移学习时, 拥有相同的神经网络结构, 在微调时, 对所有参数做更新操作; ② transAT 模型在预训练时, 输出层采用全连接层, 损失函数采用二分类交叉熵, 在微调时, 输出层变为余弦距离, 损失函数为 MSE; ③ 同方案 ② 相反, 预训练时采用“余弦距离+MSE”, 微调时采用全连接层.

4 实验结果

4.1 实验设置

本文设计了两部分实验: 一部分是语义等价任务实验, 因为在前文提到, 受限于标注数

据形式, 将 FAQ 问答任务转变为 NLP 中的 PI 任务, 即语义等价任务; 另一部分是 FAQ 问答任务实验, 模拟线上流程, 测试模型在离线数据上的效果. 经过前期数据预处理, 通过爬取社区问答数据, 获得有效语义等价问题对 404 354 个, 这部分作为源域数据集; 人工标注的 FAQ 的数据集包含有效问题对 43 556 个, 这部分作为目标域数据集. 对源域数据和目标域数据进行比对发现, 虽然在数据处理阶段, 通过设置查询种子和挖掘语义等价问题等方式, 保证了源域数据和目标域数据在内容上的相关性和形式上的一致性, 但实际结果是, 它们之间仍然存在一些差异. 从两个数据集随机抽取 5 个标准问和扩展问, 得到的结果如表 1 所示.

表 1 源域和目标域数据对比

Tab. 1 Comparison of data from the resource and target domain		
	标准问	扩展问
源域	支付宝里的钱怎么花掉	支付宝怎么改变付款方式
	华为 mate7 严重发烫怎么回事	手机发烫怎么办
	手机受限后收不到验证码怎么办	无法接收到短信息
	荣耀 6 的网络位置怎么耗电这么快	小米 note2 耗电快怎么办
	华为自带手机浏览器如何删除	华为手机浏览器卸载会如何
目标域	商品有质量问题怎么办	收到货外观破损怎么办
	怎么查询不到物流信息	我买的商品, 物流一点动静都没有
	如何确认收货	收到邮件如何操作确认收货
	评价上说的都是真的吗	为何有差评
	如何办理退货	可以退货吗

通过对比发现, 源域和目标域数据在表达方式、用词规范, 以及涉及的领域都有一定的差别. 因此, 源域数据中是存在一些噪音的. 比如表 1 中, “小米 note2 耗电快怎么办”与“与荣耀 6 的网络位置怎么耗电这么快”两个问题存在语义差距, 但是因为源域数据来自网络社区, 必然会有存在差异, 所以本文在微调时尝试了不同的策略, 以弱化这种影响. 本文所有实验的平台为 Linux GPU 服务器, 其中, CPU 为 Intel(R) Xeon(R) CPU E5-2650 v4 @ 2.20 GHz×2 共 48 核, GPU 为 Tesla P40, 内存为 200 GB.

在语义等价任务实验中, 首先需要构造负例, 此处采用随机构造负例法构造负例. 分别对源域数据集和目标域数据集构造负例, 并且以 8:2 划分训练集和测试集, 得到的结果如表 2 所示.

表 2 语义等价任务实验数据集划分结果

Tab. 2 Semantic equivalent experimental data partitioning results		
	社区问答数据集	FAQ 数据集
总量	808 708	87 112
训练集	646 966	69 690
测试集	161 742	17 422

在语义等价任务实验中, 选取 LSTM、BCNN、PWIM 模型作为参照模型, 其中, LSTM 模型采用基础的架构, 即 Word Embedding+LSTM+全连接层实现, BCNN 和 PWIM 分别应用 Github 上的开源项目实现.

在 FAQ 问答任务中, 源域数据集划分与语义等价任务实验一致, 目标域数据集重新划分训练集和测试集, 保证问题对中的每个标准问题均匀地分配到训练集和测试集中, 得到训练集 39 074 个问题对, 测试集 4 075 个问题对. 然后, 针对训练集尝试用不同的负例构造方法和负例占比构造负例; 对于负例构造方法, 分别尝试用 BM25 构造负例和随机法构造负例. 对于负例占比, 分别生成负例占比为 1:1, 1:2, 1:3, 1:5 的训练集. 为了模拟场景, 在测试时, 首先用基于

Ngram 的 TF-IDF 模型做召回操作, 召回前 50 个候选项, 加快了实验效率, 避免了过多干扰项的影响.

4.2 实验结果

首先, 分别对源域数据集和目标域数据集进行语义等价任务实验, 各个模型在不同数据集上的结果如表 3、表 4 所示. 表 3、表 4 中, Precision 为精确率, Recall 为查全率, F1 表示 F1-score, Time cost 为模型从训练到收敛的时间.

表 3 各个模型在源域数据集的测试结果

Tab. 3 Test results for various models in the source domain datasets

模型	Precision	Recall	F1	Time_cost/s
LSTM	0.908 3	0.958 3	0.932 6	1 223
BCNN	0.870 0	0.970 0	0.920 0	4 200
PWIM	0.937 1	0.928 7	0.932 9	172 800
transAT	0.982 5	0.983 7	0.983 1	3 927

表 4 各个模型在目标域数据集的测试结果

Tab. 4 Test results for various models in the target domain datasets

模型	Precision	Recall	F1	Time_cost/s
LSTM	0.840 8	0.944 7	0.889 7	420
BCNN	0.890 0	0.940 0	0.920 0	1 188
PWIM	0.951 7	0.952 2	0.952 0	622 182
transAT	0.965 3	0.959 2	0.962 2	540

从表 3 可以看出, 本文提出的 transAT 模型在源域大规模数据集上, 超过了所有的参照模型, 达到了很好的结果, 并且在训练时间上远小于神经网络架构复杂的 PWIM 模型. 由此证明 transAT 模型在大规模语料上, 有着很强的学习能力, 并且在性能上, 有着更短的训练时间和更快的收敛速度(Time cost).

从表 4 可以看出, transAT 模型在目标域数据集上也超过了参照模型, 并且取得了很好结果, 有效避免了过拟合的发生.

因为参照模型没有相应的迁移学习方案, 此处, 只针对 transAT 模型对比不同的迁移学习方案, 设计以下 5 组迁移学习方案: ① 采用局部微调的方式, 微调时重新初始化参数再训练, 记为 fine_tune0; ② 采用局部微调的方式, 微调时接着原有的参数继续训练, 记为 fine_tune1; ③ 采用全局微调方式, 不改变网络结构, 记为 fine_tune_all; ④ 采用全局微调方式, 预训练时输出层用余弦距离+MSE, 微调时输出层为全连接, 记为 CSBC; ⑤ 采用全局微调方式, 预训练时采用全连接, 微调时采用余弦距离+MSE, 记为 BCCS. 不同的迁移学习方式, 在目标域数据集上的测试结果如表 5 所示.

从表 5 可以看出, 局部微调的效果很差. 分析其原因是, 当把 Transformer 层的参数冻结后, 可以调节的参数量很小, 模型会很快地收敛, 受源域数据集影响太大; 而全局微调的结果会好很多, 并且 BCCS 方式, 超过了未经过迁移学习训练的 transAT 模型效果, 这证明了迁移学习方式的有效性.

在 FAQ 问答任务实验中, 首先测试不同负例构造方式对模型结果的影响. 本文用精确率(Precision)来衡量效果, Precision@1(P@1)代表返回结果 top 1 的精确率, 这和实际的问答场景一致, 因为真实的问答系统只返回一个正确答案. 各个模型结果如表 6 所示. 从表 6 可以看出, 基于 BM25 模型构造负例的方法要优于随机(random)构造负例的方法, 究其原因, BM25 构造的

负例更具有区分性,可以帮助模型学习到不常见的特征,提升模型的区分能力.并且,在FAQ问答任务实验中,经过迁移学习训练的 transAT 模型(transAT (pretrain))要优于其他所有参照模型.

表 5 不同迁移学习方式的迁移学习效果

Tab. 5 Results of various transfer learning methods

迁移学习方式	Precision	Recall	F1
transAT	0.965 3	0.959 2	0.962 2
transAT(fine_tune0)	0.840 2	0.801 2	0.820 2
transAT(fine_tune1)	0.841 3	0.800 9	0.820 6
transAT(fine_tune_all)	0.953 8	0.961 8	0.957 8
transAT(CSBC)	0.959 8	0.963 4	0.961 5
transAT(BCCS)	0.970 3	0.968 2	0.969 2

表 6 不同负例构造方式, FAQ 问答实验结果

Tab. 6 Results of FAQ QA experiment with different negative sample constructions

	P@1(random)	P@1(BM25)
LSTM	0.643 4	0.656 1
BCNN	0.633 1	0.655 2
PWIM	0.742 1	0.761 1
transAT	0.772 2	0.805 3
transAT(pretrain)	0.762 1	0.806 0

然后测试不同负例占比对模型结果的影响,实验结果如表 7 所示.从表 7 可以看出,在训练时,不同的负例占比对结果的影响很大.整体来看,随着负例占比数量的增加,模型的区分能力越来越强,但是当负例占比过高时,也会对模型的效果产生负面的影响.本文提出的 transAT 模型相比于参照模型,效果有显著的提升,并且经过迁移学习的 transAT (transAT(pretrain))取得了更优的效果.

表 7 不同正负例占比, FAQ 问答实验结果

Tab. 7 Results of FAQ QA experiment with different ratios of positive and negative

	P@1(1:1)	P@1(1:2)	P@1(1:3)	P@1(1:5)
BM25	0.601 2	0.601 2	0.601 2	0.601 2
LSTM	0.643 4	0.653 2	0.667 8	0.632 1
BCNN	0.633 1	0.634 5	0.645 8	0.620 0
PWIM	0.742 1	0.751 2	0.742 0	0.730 3
transAT	0.772 2	0.785 6	0.784 8	0.751 1
transAT(pretrain)	0.762 1	0.786 6	0.785 8	0.742 7

4.3 实验分析

在实验过程中,本文对神经网络的结构设计、模型的超参数、注意力层的影响等方面做了大量的实验分析,碍于篇幅限制,此处只介绍语义等价任务实验中注意力层的影响.

在语义等价任务实验中,为了直观感受注意力层对实验结果的影响,本文分别做了对比实验,并且对注意力层的输出结果进行了可视化操作.通过实验发现,在 transAT 网络中加入注意力层与不加入注意力层的差距非常显著,不加入注意力层在源域数据集上得到的 F1 为 0.972 4,小于最优结果 0.983 1,在目标域数据集上得到 F1 为 0.939 8 远小于最优结果 0.962 2.注意力层

的可视化结果如图 3 所示。

从图 3 可以看出, 发热与散热这两个词存在明显的语义间隔, 并且在图上有清晰的显式, 说明 transAT 中应用的注意力网络可以有效地学习词之间的语义关联, 帮助网络学到更好的语义特征。

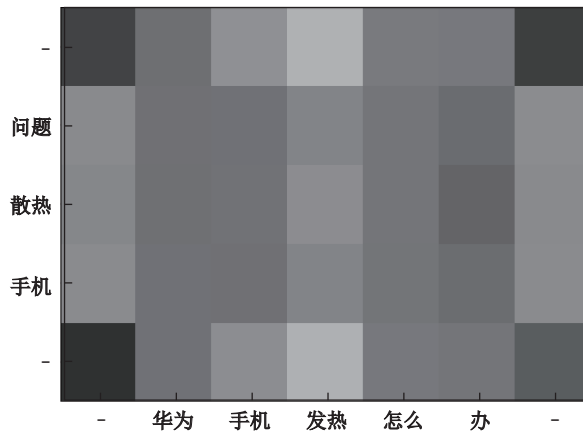


图 3 注意力机制权重图

Fig. 3 Visualization of attention weights

5 总结与展望

本文针对 FAQ 问答数据集规模有限、训练语料不足的问题, 提出了应用社区问答数据集作为源域数据集进行迁移学习训练. 为了有效屏蔽社区问答数据集中噪声的影响, 分别在数据层面和模型层面设计了相应的解决方法. 在数据层面, 通过爬取策略和语义等价问题挖掘, 保证了源域数据集和目标域数据集数据的一致性和相关性; 在模型层面, 提出了一种面向迁移学习的深度神经网络结构 transAT 和不同的迁移学习策略. 实验证实, transAT 网络可以有效适应不同规模的数据集, 并且优于 PWIM、BCNN 等参照模型, 取得了很好的效果. 通过实验证明, 基于本文提出的设计方案, 可以有效提升 FAQ 问答任务中模型的效果, 有效缓解数据集规模有限对模型造成的影响.

迁移学习被广泛地应用到 NLP 领域中, 并且取得了很好的效果. 在接下来的工作中, 需要进一步尝试不同的迁移学习方法, 进一步提升模型的效果.

[参 考 文 献]

- [1] TURNEY P D, PANTEL P. From frequency to meaning: Vector space models of semantics[J]. Journal of Artificial Intelligence Research, 2010, 37: 141-188.
- [2] ROBERTSON S, ZARAGOZA H. The probabilistic relevance framework: BM25 and beyond[J]. Foundations and Trends in Information Retrieval, 2009, 3(4): 333-389.
- [3] KATO S, TOGASHI R, MAEDA H, et al. LSTM vs BM25 for open-domain QA: A hands-on comparison of effectiveness and efficiency[C]//Proceedings of the 40th International ACM SIGIR Conference on Research and Development. ACM, 2017:1309-1312.
- [4] WANG Z G, ITTYCHERIAH A. FAQ-based question answering via word alignment[J]. arXiv:1507.02628v1 [cs.CL].
- [5] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality[C]//Proceedings of the 2013 Conference on Neural Information Processing Systems Association. NIPS, 2013: 3111-3111.

- [6] PENNINGTON J, SOCHER R, MANNING C D. Glove: Global vectors for word representation[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics. EMNLP, 2014: 1532-1543.
- [7] KIM Y. Convolutional neural networks for sentence classification[C]// Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). ACL, 2014: 1746-1751.
- [8] LIPTON Z C. A critical review of recurrent neural networks for sequence learning[J]. arXiv:1506.00019v1 [cs.LG].
- [9] SAK H, SENIOR A W, BEAUFAYS F. Long short term memory recurrent neural network architectures for large scale acoustic modeling[C]//Proceedings of the 2014 Conference of the International Speech Communication Association. INTERSPEECH, 2014: 338-342.
- [10] VASWANI A, SHAZEER N, PARMAR N, et al. Attention is all you need[C]// Advances in Neural Information Processing Systems 30 (NIPS 2017). 2017: 6000-6010.
- [11] DEVLIN J, CHANG M W, LEE K, et al. BERT: Pretraining of deep bidirectional transformers for language understanding[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies. 2019:4171-4186.
- [12] YU J F, QIU M H, JIANG J, et al. Modelling domain relationships for transfer learning on retrieval-based question answering systems in e-commerce[C]// Proceedings of the 11th ACM International Conference on Web Search and Data Mining. 2018: 682-690.
- [13] NIE Y X, BANSAL M. Shortcut-stacked sentence encoders for multi-domain inference[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing Association for Computational Linguistics, EMNLP 2017, 2017: 41-45.
- [14] CONNEAU A, KIELA D, SCHWENK H, et al. Supervised learning of universal sentence representations from natural language inference data[C]//Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing. ACL, 2017: 670-680.
- [15] KRATZWALD B, FEUERRIEGEL S. Putting question-answering systems into practice: Transfer learning for efficient domain customization[J]. ACM Transactions on Management Information Systems, 2019, 9(4): 15: 1-15: 20(Article No.15).
- [16] HE H, LIN J J. Pairwise word interaction modeling with deep neural networks for semantic similarity measurement[C]//Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics Human Language Technologies. 2016: 937-948.
- [17] CHEN Q, ZHU X D, LING Z H, et al. Enhanced LSTM for natural language inference[C]//Proceedings of the 2017 Annual Meeting of the Association for Computational Linguistics. ACL, 2017: 1657-1668.

(责任编辑: 李 艺)

(上接第 73 页)

- [2] BORKAR S, RAJESWARI K. Attributes selection for predicting students' academic performance using education data mining and artificial neural network[J]. International Journal of Computer Applications, 2014, 86(10): 25-29.
- [3] LAN A S, WATERS A E, STUDER C, et al. Sparse factor analysis for learning and content analytics[J]. Journal of Machine Learning Research, 2013, 15(1): 1959-2008.
- [4] 张嘉, 张晖, 赵旭剑, 等. 规则半自动学习的概率软逻辑推理模型[J]. 计算机应用, 2018, 38(11): 98-103.
- [5] 薛颖, 沙秀艳. 基于改进模糊聚类算法的灰色预测模型[J]. 统计与决策, 2017(9): 29-32.
- [6] 文传军, 詹永照. 基于样本模糊隶属度归 n 化约束的松弛模糊 C 均值聚类算法[J]. 科学技术与工程, 2017, 17(36): 96-104.
- [7] 赵琦, 孙泽斌, 冯文全, 等. 一种基于支持向量回归的建模方法[J]. 北京航空航天大学学报, 2017, 43(2): 352-359.
- [8] 张麒增, 戴翰波. 基于数据预处理技术的学生成绩预测模型研究[J]. 湖北大学学报(自然科学版), 2019, 41(1): 106-113.
- [9] 孙毅, 刘仁云, 王松, 等. 基于多元线性回归模型的考试成绩评价与预测[J]. 吉林大学学报(信息科学版), 2013, 31(4): 404-408.
- [10] 陈岷. 因子分析和神经网络相融合的体育成绩预测模型[J]. 现代电子技术, 2017(5): 138-141.
- [11] NÚÑEZ J C, SUÁREZ N, ROSÁRIO P, et al. Relationships between perceived parental involvement in homework, student homework behaviors, and academic achievement: Differences among elementary, junior high, and high school students[J]. Metacognition and Learning, 2015, 10(3): 375-406.
- [12] BUNKAR K, SINGH U K, PANDYA B, et al. Data mining: Prediction for performance improvement of graduate students using classification[C]//IEEE 2012 Ninth International Conference on Wireless and Optical Communications Networks (WOCN). New York: IEEE, 2012: 1-5.

(责任编辑: 张 晶)