



改进的多模型融合技术在客服问答系统上的应用

王广敏¹, 王尧枫²

(1. 中国电信股份有限公司上海研究院, 上海 200122;

2. 华东师范大学计算机科学与软件工程学院, 上海 200062)

摘要: 随着人工智能技术的发展,越来越多的公司采用机器客服代替人工客服。但若采用传统关键词模型,则机器客服准确率难以提高;若采用深度学习模型进行训练,则又面临用户问题是短文本时,模型训练和预测效果不佳的问题。针对这些问题,通过深入研究和多次试验,提出一种融合关键词模型和基于字向量的深度学习模型的算法。最后实现了模型的训练和预测,在与传统算法的准确率对比方面展现了优势。

关键词: 问答系统;深度学习;人工智能

中图分类号: TP391

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2018308

Application of improved multi-model fusion technology in customer service answering system

WANG Guangmin¹, WANG Yaofeng²

1. Shanghai Research Institute of China Telecom Co., Ltd., Shanghai 200122, China

2. College of Computer Science and Software Engineering, East China Normal University, Shanghai 200062, China

Abstract: With the development of artificial intelligence(AI), more and more companies use machine customer service instead of manual customer service. However, if the traditional keyword model is adopted, the accuracy of the machine customer service is difficult to improve. If the deep learning model is used, the predict result is poor when the user problem is short text. Aiming at these problems, an algorithm combining keyword model and deep learning model based on word vector was proposed. The training and prediction of the model was realized, and the advantages were shown in the comparison with the accuracy of the traditional algorithm.

Key words: question and answer system, deep learning, AI

1 引言

对话机器人历史悠久,1966年美国麻省理工学院的精神治疗师机器人 Eliza 到现在已有半个世纪。近些年,人工智能算法、硬件产业链的整

体发展,尤其是资本市场的进入,不断推动客服机器人行业竞争升级,“智能化”成为竞争要点^[1]。目前,在中国智能客服市场上既有科大讯飞、小i机器人等第三方智能客服企业,又有百度、阿里、网易、腾讯等提供智能客服解决方案和产品的国

内互联网巨头,更有谷歌、苹果、微软、亚马逊等国外互联网巨头不断渗透,中文客服机器人市场充满竞争活力。

但这些产品在追求智能化的过程中仍存在两个问题。第一,若采用关键词模型进行训练,由于中文词汇量非常多,同样的词在不同语境中的语义差别巨大,模型的通用性往往较差,且后期维护需要同义词库、用户字典、停用词表等。第二,若采用基于循环神经网络的深度学习模型,则对训练样本有较高的要求,但在通常的客服问答系统中,大量的文本为词语量小于或等于 3 个短文本问题,深度学习模型擅长把握文本上下文语义的优势难以显现。

本文首先从关键词模型和循环神经网络的原理出发,通过真实的用户问题数据集对比两种模型的优势及劣势,从而提出一种能取长补短的融合模型算法,最终展现算法的测试结果及现网运行情况。

2 客服问答系统关键技术

客服问答系统针对服务的行业通常需要先设计或整理出问答知识库,问答知识库包含多条用户的问题及其对应的标准答案。无论是关键词模型或深度学习模型其实质均是一个分类器,即把用户的问题通过这个分类器匹配最合适的答案。

2.1 基于关键词的文本分类技术

基于关键词的文本分类主要分为 3 个步骤:文本预处理、关键词特征提取和构建分类模型。

2.1.1 文本预处理

文本预处理主要包括文本分词和去停用词两个阶段。中文分词不同于英文有天然的空格间隔,需要设计复杂的分词算法。传统算法主要有基于字符串匹配的正向/逆向/双向最大匹配;基于理解的句法和语义分析消歧;基于统计的互信息/CRF 方法。而停止词是文本中一些高频的代词/连词/介词等对文本分类无意义的词,通常维护一个停用词表。

2.1.2 关键词特征提取

向量空间模型的文本表示方法的特征提取对应特征项的选择和特征权重计算两部分。特征选择的基本思路是根据某个评价指标独立地对原始特征项进行评分排序,从中选择得分最高的一些特征项,过滤掉其余的特征项。常用的评价有文档频率、互信息、信息增益等。特征权重主要是经典的 TF-IDF^[2]方法及其扩展方法。

词频 (term frequency, TF) 指的是某一个给定的词语在该文件中出现的频率,即某词在文章中出现次数/该文章总词数:

$$tf_{i,j} = \frac{n_{i,j}}{\sum_k n_{k,j}} \quad (1)$$

逆向文件频率 (inverse document frequency, IDF) 度量一个词语的普遍重要性,由总文件数目除以包含该词语之文件的数目,再将得到的商取对数得到,具体如下:

$$idf_i = \lg \frac{|D|}{|\{j:t_i \in d_j\}|} \quad (2)$$

其中, $|D|$ 指语料库中的文件总数, $|\{j:t_i \in d_j\}|$ 指包含词语 t_i 的文件数目。

$$tfidf_{i,j} = tf_{i,j} \times idf_i \quad (3)$$

式 (3) 意味着如果一个词在某一特定文件内出现的频率很高,并且该词语在整个文件集合中其他文件出现的频率很低,可以产生出高权重的 TF-IDF。因此,TF-IDF 倾向于过滤掉常见的词语、保留重要的词语。

在对关键词特征提取后,需要对训练样本进行特征筛选,并生成特征矩阵。若提取的特征数量为 M , 最终的分类数为 C , 则特征矩阵大小为 $C \times M$, 其中每行代表每个类别对应的特征是否出现,出现置 1, 不出现置 0; 每列代表该特征是否在每个类别中出现。

2.1.3 分类模型构建

当一个用户问题经过文本预处理及特征筛选



后, 形成一个长度为 M 的特征向量。此时调用一个分类算法即可对用户问题进行文本分类, 选取概率最大的结果对应的标准答案作为输出。常用的分类算法有朴素贝叶斯分类算法^[3-4] (Naïve Bayes)、KNN、SVM 和决策树等。

2.2 基于循环神经网络文本分类方法

基于循环神经网络进行模型训练的过程一般分为文本预处理、词向量构建和训练模型。

2.2.1 文本预处理

同关键词模型预处理一样, 首先需要对文本进行分词, 但不需要去除停用词。关键词模型中去停用词是为了减少特征提取时的无效词数量, 而在神经网络中, 所有的词都是对上下文关联非常重要的, 故一般不去除任何停用词。

2.2.2 词向量构建

第 2.1 节介绍的关键词特征提取的算法在神经网络中存在的问题: 特征矩阵是高维度高稀疏的, 特征表达能力很弱, 神经网络不擅长对此类数据的处理。因此需要将文本转化为一种稠密的、连续的实数向量, 这就是词向量的设计思想。

词向量的构建方法有很多, 如 one-hot encoding、Word2Vec 等, 本文不作赘述。

2.2.3 模型训练

循环神经网络 (RNN) 与其改进之后的版本长短期记忆神经网络 (LSTM) 在自然语言文本处理中应用非常广。

传统的循环神经网络结构如图 1 所示。

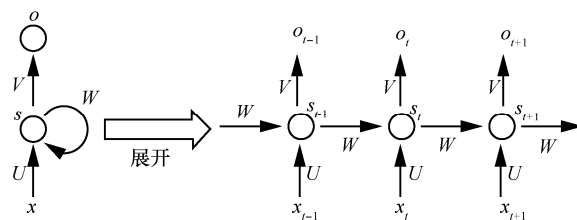


图 1 循环神经网络示意图

x 是一个向量, 表示输入层的值; s 是一个向量表示隐藏层的值; U 是输入层到隐藏层的权重矩阵; o 也是一个向量, 表示输出层的值; V 是隐藏层到输出层的权重矩阵。循环神经网络的隐藏层的值 s 不仅仅取决于当前这次的输入 x , 还取决于上一次隐藏层的值 s 。权重矩阵 W 就是隐藏层上一次的值作为这一次的输入权重。循环神经网络的输出值, 是受前面历次输入值影响的, 这就是为什么循环神经网络适合在自然语言处理中应用^[5]。

LSTM 区别于 RNN 的地方, 主要就在于它在算法中加入了一个判断信息有用与否的处理器。一个处理器中被放置了 3 扇门, 分别叫作输入门、遗忘门和输出门。一个信息进入 LSTM 的网络中, 可以根据规则来判断是否有用。只有符合算法认证的信息才会被留下, 不符的信息则通过遗忘门被遗忘。在自然语言处理中, LSTM 可以选择性地使用上下文中有用的词的信息, 而过滤掉“没有用”的词, 使得模型预测更加准确, 如图 2 所示。

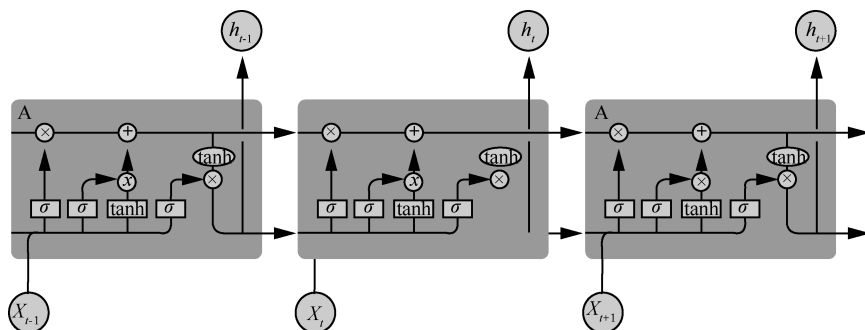


图 2 LSTM 示意图

神经网络的训练过程相对比较简单, 将处理后组成的词向量与其对应的输出类别编号输入模型中进行训练即可得到模型。

2.3 目前的分类方法在客服问答系统中存在的问题

通过对上述两节对两种最常使用的文本分类方法的介绍, 结合真实客服问答数据的特点进行分析, 可以得到两种模型的优缺点, 见表 1。

从对比中可以发现, 基于关键词的文本分类技术主要针对用户提问中的关键词进行答案的选取, 无论是选取单一答案还是多个标准问题组成的建议问题回答, 均无法脱离用户问题中的关键词。故关键词的特征选取就尤为重要, 在实际应用中, 很难选取一个合适的关键词特征列表, 往往选得较多时就会包含较多的无用词或干扰词, 选得较少就会造成信息的丢失。所以关键词模型准确率的提升随着训练集的增加慢慢降低, 最终趋于不变。

基于循环神经网络文本分类方法对训练集及其敏感。主要可以归纳为以下两点: 第一, 是训练样本的质量较优时, 模型的训练结果也会相应的提升, 这也是目前深度模型优化的一个方向; 第二, 训练样本数量的大小, 当训练样本中某一类样本的数量过少时, 会限制深度学习模型在该类上的学习效果, 导致在该类上的表现很差。

本文针对上述两种模型存在的问题, 设计了一种新的融合模型, 其主要优化内容为:

(1) 针对深度学习模型在训练中短文本过多的问题, 将词向量转化为字向量进行训练, 从而能获取到文本的上下文语义;

(2) 修改关键词文本分类模型为关键词模糊匹配模型, 使得其通用性更强;

(3) 根据用户输入的问题的特征 (如词的数量) 以及两种模型输出的结果对应的概率设计了一套融合算法, 使得融合模型倾向于对短文本提问、语义含糊的提问给予关键词模糊匹配模型的结果, 而在其他的情况下倾向于给予深度学习模型的结果。

3 采用多模型融合算法的设计与实现

本文采用的问答模型框架由模型训练和模型预测两个部分组成, 如图 3 所示。

在模型训练中, 采用多模型单独训练。关键词模糊匹配模型通过对用户问题进行预处理后形成宽表, 采用关键词模型通过对用户问题中关键词的模糊匹配精度对用户问题进行归类, 给出该类对应的业务知识答案。循环神经网络模型使用基于字向量的 LSTM 模型, 通过深度学习自我

表 1 两种分类方法的对比

问答系统文本特点	基于关键词的文本分类技术	基于循环神经网络的文本分类方法
短文本过多, 用户往往只输入不到 3 个词	比较擅长, 可以直接使用用户输入的关键词, 通过分类器寻找概率排名较高的一个或多个标准问题组成建议问题, 这些建议问题通常都包含用户提问的关键词, 能大概率解决用户问题	表现较差, 在训练过程中难以把握短文本素材前后文的语义, 造成训练效果不佳。同时在模型预测时经常会出现答非所问的情况
同一种类问题有多重不同的问法	表现较差, 中文针对一个问题有多种问法, 如果每种问法都有不同的关键词, 则关键词模型预测效果不佳。往往需要对每类问题维护一个庞大的扩展问的列表, 扩展问题包含用户对同一个问题的多种问法, 运维成本很高	表现较好, 仅需将该类问题各种问法的训练语料输入模型, 模型在预测中往往能定位到正确结果
问题中出现错别字	表现较差, 需要针对不同的词语维护一个庞大的同义词列表, 该列表中不但包括一个词的同义词, 也包括了这个词可能出现的错别字	表现较好, 神经网络主要通过上下文来推断词的意义, 当出现错别字时, 只要上下文语境类似, 模型仍能推断出与正确的词很类似的语义
含糊、笼统的提问, 不知道用户想表达的意思	表现一般, 关键词模型倾向于给出用户问题中出现的关键词组成的建议问题	表现依赖于训练集, 时好时坏。由于不了解用户的真实意图, 在进行训练集标注的过程中将该类问题标注在哪里直接决定了实际应用时的预测结果

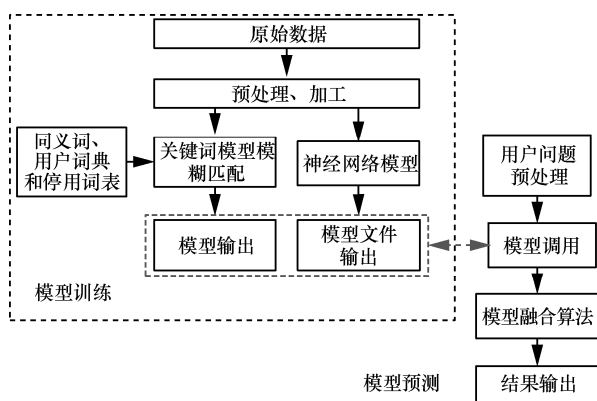


图3 问答模型框架

特征学习的方式来表达和理解用户问题，然后输出训练后的模型。

模型预测包括用户问题预处理模块、模型调用以及多模型融合模块。用户问题预处理模块对用户的问题进行清洗、分字；分别调用关键词模糊匹配模型和基于字向量的 LSTM 模型，结合模型输出和用户问题特征选取最佳答案。

3.1 数据预处理

数据预处理阶段包括专业知识库预处理和对历史用户数据预处理。专业知识库的数据预处理主要是对知识库的问题进行整理、答案回填、标准问题列分词以及同义词扩充等预处理，最终形成一张包含业务标准问题、标准答案和同义词词组的数据表，为传统的特征值匹配方法（下面提到的关键词模型）做数据准备。同时为用户标准问题和标准答案进行标签化处理，为后续关联和神经网络模型训练做准备。

对历史数据集进行数据填补和清洗，主要包括特殊符号的去除、上文为建议问的数字序号问题的回填以及用户答案的回填等，通过与标准知识库根据标准问题进行关联，扩充样本集中的标准答案和标准问题及其对应的标签。然后对用户问题进行归一化处理，即只保留文字和数字，接着对用户问题进行分词/分字，利用词性、停用词表以及用户词典，筛选关键词，组成关键词列，最终形成一张满足所用模型所需的训练样本格式

的宽表，为模型训练准备高质量数据集。历史数据的预处理主要是为传统的特征值匹配方法做准备，降低异常值或者影响因子较少的值对模型结果的影响。

3.2 模型设计

3.2.1 关键词模糊匹配模型

本文采用的传统的特征值提取的方法主要是利用关键词进行模糊匹配。关键词模糊匹配模型如图4所示。

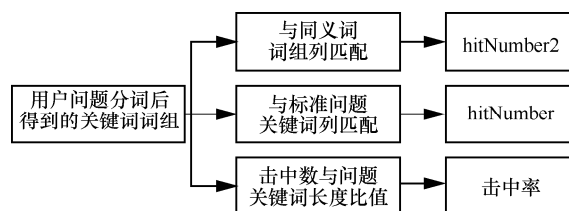


图4 关键词模糊匹配模型

关键词模糊匹配模型的核心思路是：对用户问题进行分词去停用词和词性筛选，形成一组关键词列表（KeyWordList），用该组关键词组中的每个关键词分别与标准知识库中的标准问题关键词列和同义词词组列进行关键词匹配，然后分别计算每个标准问题匹配到的总长度，各自形成击中数（hitNumber 与 hitNumber2）列，同时计算击中数与关键词列表长度的比值，作为命中率列。

关键词模型的输出策略为：首先依次筛选击中数列匹配数最大的标准问题组成子集；然后在该子集中选取命中率最大的标准问题组成的数据集作为最终的输出。若该数据集长度为1，则表示有唯一精确匹配的回答；若该数据集长度大于1，则表示可以将数据集中的标准问题组成建议问答复给用户。

3.2.2 基于字向量的 LSTM 模型

深度学习模型采用 LSTM 模型进行训练。在 LSTM 模型的数据准备阶段，汇总所有字组成字库。利用生成后的字库，对数据集中的用户问题进行数字化转换，便于计算机识别和计算；同时形成标签库，利用标签库对类别进行 one-hot 格式

变换。数据处理完毕后经过 LSTM 模型进行训练, 训练完成后会生成模型文件。

在模型预测过程中, 需要用到模型文件、字库和标签库文件。由于字库和标签库序号在每次训练中均随机生成, 故每次训练结束后均需要保证模型文件、字库和标签库的一致性。

3.2.3 模型融合算法设计

为了更好地设计融合算法, 本文首先对上述两个模型进行了单独的训练与预测, 通过对两个模型的输出结果的对比可以发现以下特点。

(1) 针对用户经常会问到的问题, 由于训练样本较多且表述丰富, 因此 LSTM 模型能够发挥很大的优势。

(2) 针对比较生僻的问题, 如出现率小于千分之一的问题, 由于用户问题的多样性, LSTM 模型大概率上会将该类问题误判到其他类别中。在这种情况下, 考虑到该类问题一般都有明显区别于其他问题的专业术语, 故可以使用关键词模糊匹配模型, 来弥补 LSTM 模型的不足。

(3) 针对只有一个词组成的用户问题, 关键词模糊匹配模型无论是精准回答还是建议问, 均能大概率找到正确答案, 而 LSTM 模型却会出现答非所问的情况, 这时应尽量采用关键词模糊匹配模型。

(4) LSTM 模型前三大率对应的标准回答有 99% 的概率包含正确结果。故融合模型在神经网络输出的结果选取中, 保留前三大率对应的标准回答。

(5) LSTM 第一大概率超过 40% 时, 其对应的标准回答的准确率为 91.38%。故可以认为当 LSTM 第一大概率超过 40% 时, LSTM 模型基本能输出正确结果。

根据上述特征, 融合模型具体算法如下。

步骤 1 当关键词模型匹配到击中率等于 1 的标准问仅有 1 条, 且匹配到的总的标准问小于

或等于 2 条时, 输出关键词模型的答案。

步骤 2 步骤 1 条件不满足时, 若神经网络模型输出的答案的概率大于 40%, 输出神经网络答案。

步骤 3 若步骤 1、2 条件均未满足, 那么若关键词模型匹配到的标准问条数大于 1 条, 且小于 6 条, 则模型给出的答案为是否由关键词组成的建议问; 若关键词模型匹配到的标准问条数大于 6 条, 则模型给出的答案为是否由神经网络模型输出的前 3 个标准问题组成的建议问。

3.3 模型优化

本文通过知识库、融合算法、数据集 3 个方面的优化实现整体准确率的提升。

知识库优化主要是优化标准问题、合并相似问题或者同类问题, 同时也对标准答案过于生硬的话术进行优化, 前者是为了减少分类个数, 提升模型性能; 后者则是为了增强用户体验。

融合算法优化将不断地通过新的数据优化模型参数, 同时提炼出更多的用户问题的特点, 进行场景细分, 归纳出新的算法。

数据集优化通过人工标注数据集的方法, 主要是对训练集根据优化过的知识库进行重新筛选, 并将分类错误的问题正确归类, 避免训练集数量上损失。

3.4 测试结果及分析

本文使用了某运营商 6 个月的客服问答数据^[6], 共计 215 万条原始数据, 通过数据清洗后可用训练的数据集为 10.3 万条, 利用人工对这 10.3 万条数据进行了标注, 根据知识库中每个标准问题对应的样本进行训练集、验证集及测试集的等比例划分, 划分结果如下。

- 训练集: 7.5 万条。
- 验证集: 1.4 万条。
- 测试集 1: 1.4 万条。
- 测试集 2: 根据知识库中的标准问题, 人为设计的测试样本, 共 788 条。
- 知识库大小: 共 221 个分类。



问答系统的数据规模见表 2。

表 2 问答系统的数据规模

	训练集	验证集	测试集 1	测试集 2
数据量	7.5 万条	1.4 万条	1.4 万	788 条
神经网络模型	95%	91.38%	85%	88%
关键词模型	/	73.1%	70%	85%
融合模型	/	/	91%	86%

从表 2 中可以看出, LSTM 模型在测试样本 2 上的表现要优于测试集 1, 这是因为测试集 2 是根据知识库的标准问题人为设计的测试问题, 而测试集 1 是真实场景下发生的用户问题。而关键词模型的准确度也具有同样的趋势, 这是因为相比测试集 2, 真实场景下发生的用户问题表述更加丰富, 使得关键词模型难以匹配特定的词组。同时, 融合模型在两种测试集中均有不错的表现, 在实际场景下的效果更佳。

本文设计的系统已在某运营商某产品的客服部门上线使用, 目前平均每月处理 30 万条用户提问, 用户转人工率为 22%。系统稳定运行, 响应时间小于 1 s。

4 结束语

文本针对传统客服问答使用的两类模型存在的问题, 在深入研究和对比了基于关键词的文本分类技术和基于循环神经网络的文本分类技术后, 提出了一种适用于客服问答系统的融合算法, 并开发了相应的系统, 通过测试数据及现网数据的验证, 取得良好的效果。

客服问答系统虽然以解决用户问题为主, 但其闲聊功能是其“智能性”的最好表现。增加客服问答系统的闲聊功能是本系统下一步的工作重点, 相关的领域也涌现出不少新的技术, 如 SeqGan 网络等。

参考文献:

- [1] YU Z T. Chinese question classification based on support vector machine[J]. Journal of South China University of Technology, 2005, 33(9): 25-27.
- [2] LI X, ROTH D. Learning question classifiers: the role of semantic information[J]. Natural Language Engineering, 1998, 12(3): 229-249.
- [3] YUAN X J, SHI J X, NING H. Feature selection using syntactic and semantic information in question classification[J]. Computer Engineering & Applications, 2008, 44(33): 144-147.
- [4] SUN J G, CAI D F, DE-XIN L V, et al. HowNet based Chinese question automatic classification[J]. Journal of Chinese Information Processing, 2007, 21(1): 90-95.

[作者简介]



王广敏(1991-), 女, 中国电信股份有限公司上海研究院工程师, 主要从事 NLP 算法研究和数据挖掘方面的工作。



王尧枫(1997-), 男, 华东师范大学在读, 主要从事数据分析和计算机网络方面的研究工作。