基于 Fasttext 的人机对话意图系统

张根源，李 晓

浙江树人学院信息科技学院 浙江杭州

【摘要】对话意图识别是问答系统研究热点之一，意图识别是问答系统给出答案的前提。本文首先构建 Fasttext 意图识别模型，并进行模型的优化，进行多重分类研究算法效果。为了减小短文本，词语重复率较低，词共现并不适用，语言规范性不强的文档集的关键词抽取问题，在 Fasttext 模型输入阶段使用 TextRank 算法和 TF-IDF 算法提取文本的关键子句输入训练模型,同时采用 TF-IDF 提取文本的关键词作为特征补充,提高文本分类的效果。获取用户问题，输入到 Fasttext 模型进行意图分类，返回最终答案给用户。

【关键词】TextRank; Fasttext; 卷积神经网络方法

【基金项目】2020 年度大学生创新创业训练计划项目（202011842005）

Human-machine dialogue intention system based on Fasttext

Genyuan Zhang, Xiao Li

School of Information Technology, Zhejiang Shuren University, Hangzhou, Zhejiang

【**Abstract**】Dialogue intent recognition is one of the research hotspots in question answering systems, and intent recognition is the premise for question answering systems to give answers. This paper firstly builds the Fasttext intent recognition model, optimizes the model, and conducts multi-classification to study the effect of the algorithm. In order to reduce short text, the word repetition rate is low, the word co-occurrence is not suitable, and the keyword extraction problem of the document set with low language norm is used in the input stage of the Fasttext model to use the TextRank algorithm and the TF-IDF algorithm to extract the key. The clause is input to the training model, and TF-IDF is used to extract the keywords of the text as a feature supplement to improve the effect of text classification. Get the user question, input it into the Fasttext model for intent classification, and return the final answer to the user.

【**Keywords**】TextRank; Fasttext; Convolutional Neural Network Method

1 引言

问答系统是人工智能发展的必然产物。人们希望机器能够智能地理解人们输入的自然语言并满足用户的不同需求。问答系统的关键任务就是对用户输入的问句进行处理，然后以自然语言的方式准确地反馈给用户所需的信息。因此，问句处理模块对于问答系统的性能有着非常重要的作用。问句处理得当不仅能简化答案搜索还可以提高答案的准确性，对后续问答系统的实现提供帮助。问句处理模块一般包含以下几个步骤：（1）判定用户输入的问题类型，（2）通过问题类型确定对应输出答案的类型，（3）抽取问题关键词作为答案检索的查询输入。确定用户输入的问题类型一般采用问句分类来实现。问答系统用户输入地问句为自然语言而且语句较短，属于短文本，因此问句分类可以使用文本分类的方法。但是在文本分类的准确率上和时间上却没同时有好的效果。

近年来，国内外研究人员就上述问题开展了大量研究[2]。用户输入的问句我们可以将其分为闲聊类和垂直任务类。垂直任务类具有主题明确，易于检索的特点，而闲聊类具有范围较广泛、语句较短、口语化严重等特点，需要进一步的提取关键词才能

作者简介：张根源，男，河南，浙江树人大学信息科技学院计算机科学与技术专业 2019 级本科生。

更好地搜索答案。目前研究者们提出了多种针对文本结构的关键词抽取方法。但是问答系统问句语料库的语言结构和特点决语料库的关键词抽取方法仍不完善，导致了目前用于文本分类的机器学习和神经网络在文本分类方面虽然可以取得很好的分类准确率，但是分类时间一直是一个需要解决的问题。

针对上述缺陷，本文聚焦问答意图分类系统的时间及准确率，提出改进关键词提抽取和基于 Fast text 结合的问答系统。系统分为两部分：由改进关键词抽取和意图识别模型生成两部分组成；接收用户的自然语言问句，通过意图识别模型分类，为用户返回精准答案。构建基于 Fasttext 用户意图识别模型

1.1 数据处理

（1）数据来源

本文采用 SMP2017 中文人机对话的数据集.数据集包含闲聊类和面向任务类（垂直类），垂直类又包括 30 个小类别。

（2）数据预处理

本实验运用 jieba 中文分词系统,进行中文文本分词。每个词用空格分开，每条问句文本为一行,每一行的开头用 Fasttext 固定格式“_lable_+标签”的形式进行处理。

（3）构建过程

本实验将问句分为两层处理，第一层对问句进行二分类处理将问句分为闲聊类和垂直任务类两大类。将第一层分为垂直任务类的文本进行第二层分类，第二层分类将文本进行三十分类，分为 30 个小类，分别对应 30 个垂直领域。

1.2 Fasttext 模型

Fasttext 方法包含三部分，模型架构，层次 Soft Max 和 N-gram 子词特征。用户在输入问句后可通过模型和训练参数获取答案。本文采用 Fasttext 模型，通过输入层，隐藏层，输出层获得预测值。

（1）模型架构

Fasttext 的架构和 word2vec 中的 CBOW 的架构类似，两种模型都是基于 Hierarchical Softmax，都是三层架构：输入层、隐藏层、输出层。

CBOW 的架构：输入的是 w（t）w（t）的上下文 2N 个词，经过隐藏层后，输出的是 w（t）。如图 1 所示，W（tN+1）……W（t-1）W（t_N+1）……W（t-1）是目标单词的上下文作为输入，W（t）是目标词汇。如图 1 所示。

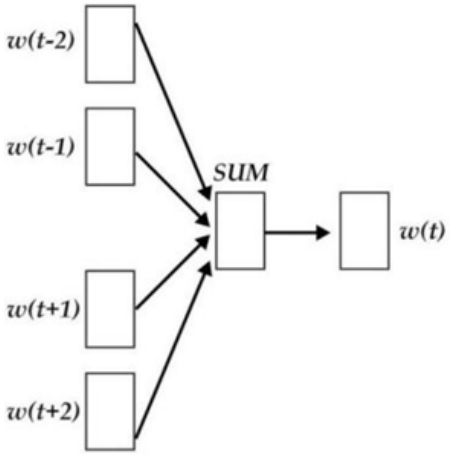

图 1 CBOW 模型

Fasttext 模型架构：这和前文中提到的 cbow 相似，cbow 用上下文去预测中心词，而此处用全部的 n-gram 去预测指定类别。Fasttext 模型架构和 word2 vec 中的 CBOW 很相似，不同之处是 Fasttext 预测标签而 CBOW 预测的是中间词，即模型架构类似但是模型的任务不同。如图 2 所示。

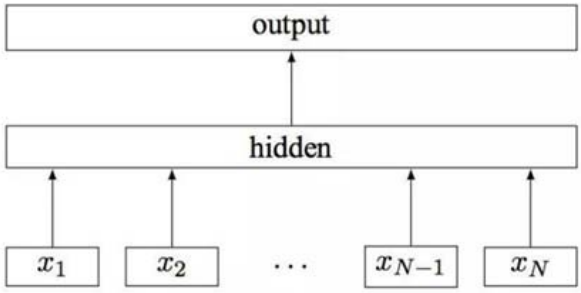

图 2 Fasttext 模型

Fasttext 模型输入一个词的序列（一段文本或者一句话)，输出这个词序列属于不同类别的概率。序列中的词和词组组成特征向量，特征向量通过线性变换映射到中间层，中间层再映射到标签。Fasttext 在预测标签时使用了非线性激活函数，但在中间层不使用非线性激活函数。

其中 x_1，x_2，...，x_{N-1}，x_N x_1，x_2，...，x_{N-1}，x_N 表示一个文本中的多个单词及其 n-gram 特征，每个特征是词向量的平均值。

（2）层次 SoftMax

对于有大量类别的数据集，Fasttext 使用了一个分层分类器（而非扁平式架构），不同的类别被整

合进树形结构中。为了改善运行时间，Fasttext 模型使用了层次 Softmax，对标签进行编码，能够极大地缩小模型预测目标的数量。另外，Fasttext 也利用了类别（class）不均衡这个事实，通过使用 Huffman 算法建立用于表征类别的树形结构。因此，频繁出现类别的树形结构的深度要比不频繁出现类别的树形结构的深度要小，这也使得进一步的计算效率更高。

（3）N-gram 子词特征

常用的特征是词袋模型，但词袋模型不能考虑词之间的顺序，因此 Fasttext 还加入了 N-gram 特征。例如“我爱她”这句话中的词袋模型特征是“我”，“爱”，“她”。这些特征和句子“她爱我”的特征是一样的。如果加入 2-Ngram，第一句话的特征还有“我-爱”和“爱-她”，这两句话“我爱她”和“她爱我”就能区别开来了。当然，为了提高效率，需要过滤掉低频的 N-gram。

1.3 模型训练

（1）基于卷积神经网络与 Fasttext 的模型对比

通过 TensorFlow 调用提供好的 API 接口来构建卷积神经网络模型卷积神经网络模型由 1 个词嵌入层、1 个卷积层和 1 个池化层组成，词向量维度为 150 维，池化层为 1-max pooling，卷积核窗口高度 3，4，5，卷积核为 150。实验数据使用上文中的数据集。

（2）模型效果对比

通过准确率、召回率、F1 值以及分类时间四个方面进行对比。Fasttext 模型和 Cnn 模型的对比数据如表 1 所示。

表 1　TextCnn 和 Fasttext 的效果对比

模型	准确率（%）	召回率（%）	F1 值	时间（s）
Fasttext	92.833333	90.566666	90.255555	6s
Cnn	82.835213	85.123141	83.223145	56s

通过分析可以看出，Fasttext 在测试数据集得到不错的结果，分类准确率、召回率以及 F1 值都高于卷积神经网络的分类结果，而且时间上，Fasttext 的分类器训练和测试的时间要明显少于卷积神经网络所构建的分类器。

2 改进 Fasttext 模型

（1）基本思想

通过提取长文本关键特征的方法保留关键特征，同时减少无关词语的占比。长文本的特征可以从关键子句提取，关键子句可以有效的保留文本的中心特征句和特征句子词之间的联系。使用 Text Rank 算法提取文本的关键子句，再将使用 TF-IDF 算法提取文本的关键子句抽取的结果进行并集处理作为输入数据使用。关键子句容易丢失关键子句外的关键词信息，因此采用关键词词组对特征进行补充。

使用 TF-IDF 算法提取每个文档中相对于整体文档区分度高的词，既考虑词频又考虑了逆文档频率，并为该文档标记分类标签，与关键子句一同作为输入数据使用。

（2）模型架构

本文在输入层中添加关键词提取模型，先使用 TextRank 算法和 TF-IDF 提取输入文本的关键子句做并集处理，同时使用 TF-IDF 筛选文本特征词词组作为输入数据的特征补充；之后将得到的关键子句和特征词词组标记为当前文本的分类标签分别送入隐藏层计算。改进 FastText 的模型结构如图 3 所示。

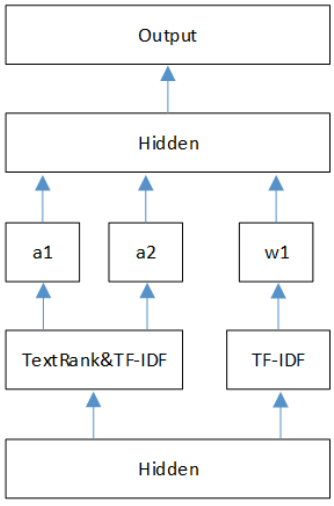

图 3　改进 Fasttext 模型

a1、a2、a3 是文本经过 TextRank 和 TF-IDF 方法提取的关键子句，w1 是通过 TF-IDF 提取的文本关键词词组，二者均标记为当前文本的标签类型，作为 FastText 训练模型的输入。输入的关键子句和关键词词组在输入到隐藏层前会被转换为各自对应词序列的特征向量，特征向量通过线性变换映射到隐藏层，该隐藏层通过求解最大似然函数后进行层次 Softmax 计算，得到最终的输出。

3 系统设计

（1）系统结构设计

本文研制的人机意图识别系统包括用户端和管理端。用户端登录系统后可以进行文字识别、语音识别、图片识别等操作。管理端登录后可以进行用户端的识别效果的校准等操作。

（2）系统功能验证

用户输入文字，先进行数据处理，将语句进行拆分，利用改进后的 textrank 进行关键词的提取，将提取的关键词加入已经训练好的 Fasttext 分类器的模型，进行预测。运行界面如图 4 所示。

图 4　系统运行界面

4 结语

人工智能时代，越来越多的智能问答系统出现在人们的生活中，问答系统能够快速精准地返回人们所提出问题的答案，因此意图识别系统不仅是科研领域的研究焦点更是工业界研究的热点。为了能使意图识别系统信息检索效率更高，本文提出用户意图多级分类思想，针对不同的用户意图制定不同的答案搜索策略。在对用户闲聊类语料库进行分析的基础上，本文发现闲聊类问句具有主题宽泛，中心思想不明，内容较短，口语化严重等问题，直接进行答案检索效率不高。本文通过对其进行关键词抽取来提高信息搜索效率。对 Fasttext 的输入阶段进行优化处理，使其更快更准确的返回给用户答案。

参考文献

[1] 陈永平，杨思春，毛万胜，etal.中文问答系统中基于主题和焦点的问题理解[J]. 计算机系统应用，2011,20(6): 56-60.

[2] 张芳芳,马敬东,王小贤,卢乃吉,夏晨曦.国外医学领域自动问答系统研究现状及启示[J].医学信息学杂志,2017,38(03):2-6,25.

[3] 文勖.中文问答系统中问题分类及答案候选句抽取的研究[D].哈尔滨：哈尔滨工业大学，2006.

[4] 郑实福，刘挺，秦兵，etal.自动问答综述[J].中文信息学报，2002, (06): 46-52.

[5] 毛成勇,高慧敏.基于 OSEK 的任务调度算法改进及实现[J].计算机工程,2010,36(04):233-235.

收稿日期：2022 年 7 月 10 日
出刊日期：2022 年 8 月 15 日
引用本文：张根源，李晓，基于 Fasttext 的人机对话意图系统[J],科学发展研究, 2022, 2(3) : 61-64
DOI: 10.12208/j.sdr.20220078

检索信息：RCCSE 权威核心学术期刊数据库、中国知网（CNKI Scholar）、万方数据 （WANFANG DATA）、Google Scholar 等数据库收录期刊