Journal of Computer Applications ISSN 1001-9081 2018-08-10
计算机应用,2018,38(8):2170-2174 CODEN JYIIDU http://www.joca.cn

文章编号:1001-9081(2018)08-2170-05 doi:10.11772/j.issn.1001-9081.2018010190

聊天机器人中用户就医意图识别方法

余 慧[1],冯旭鹏[2],刘利军[1],黄青松[1,3*]

(1.昆明理工大学 信息工程与自动化学院,昆明 650500; 2.昆明理工大学 教育技术与网络中心,昆明 650500;
3.云南省计算机技术应用重点实验室(昆明理工大学),昆明 650500)

(*通信作者电子邮箱 kmustailab@hotmail.com)

摘 要: 传统的聊天机器人中用户意图识别一般采用基于模板匹配或人工特征集合等方法,针对其费时费力而且扩展性不强的问题,并结合医疗领域聊天文本的特点,提出了基于短文本主题模型(BTM)和双向门控循环单元(BiGRU)的意图识别模型。该混合模型将用户就医意图识别看作分类问题,使用主题特征,首先通过BTM对用户聊天文本逐句进行主题挖掘并量化,然后送入BiGRU进行完整上下文学习得到连续语句最终表示,最后通过分类完成用户就医意图识别。对爬取的语料进行实验,BTM-BiGRU方法明显优于传统的支持向量机(SVM)等方法,其F值更是高出目前较好的卷积长短期记忆组合神经网络(CNN-LSTM)近1.5个百分点。实验结果表明,在本任务上该混合模型重点考虑研究对象的特点,能有效提高意图识别的准确率。

关键词: 就医意图识别;医疗聊天文本;短文本主题模型;双向门控循环单元;模板匹配

中图分类号: TP391.1 **文献标志码:** A

Identification method of user´s medical intention in chatting robot

YU Hui[1], FENG Xupeng[2], LIU Lijun[1], HUANG Qingsong[1,3*]

(1. *Faculty of Information Engineering and Automation, Kunming University of Science and Technology, Kunming Yunnan* 650500, *China*;
2. *Educational Technology and Network Center, Kunming University of Science and Technology, Kunming Yunnan* 650500, *China*;
3. *Yunnan Provincial Key Laboratory of Computer Technology Applications (Kunming University of Science and Technology), Kunming Yunnan* 650500, *China*)

Abstract: Traditional user intention recognition methods in chatting robot are usually based on template matching or artificial feature sets. To address the problem that those methods are difficult, time-consuming but have a week extension, an intention recognition model based on Biterm Topic Model (BTM) and Bidirectional Gated Recurrent Unit (BiGRU) was proposed with considering the features of the chatting texts about health. The identification of user´s medical intention was regarded as a classification problem and topic features were used in the hybrid model. Firstly, the topic of user´s every chatting sentence was mined by BTM with quantification. Then last step´s results were fed into BiGRU to do context-based learning for getting the final representation of user´s continuous statements. At last, the task was finished by making classification. In the comparison experiments on crawling corpus, the BTM-BiGRU model obviously outperforms to other traditional methods such as Support Vector Machine (SVM), even the F value approximately increses by 1.5 percentage points compared to the state-of-the-art model combining Convolution Neural Network and Long-Short Term Memory Network (CNN-LSTM). Experimental results show that the proposed method can effectively improve the accuracy of the intention recognition focusing on characteristics of the study.

Key words: identification of medical intention; medical chatting text; biterm topic model; bidirectional gated recurrent unit; template matching

0 引言

近年来,人工智能发展迅猛,逐渐融入各个领域,其中在医疗领域的应用正引起学术界和工业界广泛关注。不少在线医疗平台开始使用聊天机器人来提供健康咨询服务,这时聊天机器人不仅充当客服角色,更多起到一个健康咨询师的作用。在和“健康咨询师”聊天过程中,用户会产生大量数据,不仅包含其健康信息,还包含其他相关信息。如果能够利用这些信息提前判断出用户就医倾向,则可以为下一步给用户提出合理治疗建议以及推荐相应科室作好准备[1]。

准确识别用户意图有助于了解用户潜在需求,辅助事件预测以及判断事件走向[2]。虽然目前聊天机器人中用户意图识别的研究工作还处于起步阶段,但由于互联网蓬勃发展,互联网用户意图识别的研究正如火如荼进行,因此可以借鉴这些相关领域的研究。比如基于搜索引擎的查询意图识别,研究者主要通过分类Query来确定用户的搜索内容[3]。在消费意图挖掘研究中有学者利用模板的思想来抽取和泛化用户的消费意图,Ramanand等[4]提出基于规则和图的方法来获取

收稿日期:2018-01-23;修回日期:2018-03-28;录用日期:2018-03-28。 基金项目:国家自然科学基金资助项目(81360230,81560296)。

作者简介:余慧(1993—),女,四川成都人,硕士研究生,主要研究方向:自然语言处理、医疗信息服务; 冯旭鹏(1986—),男,河南郑州人,助理实验师,硕士,主要研究方向:信息检索; 刘利军(1978—),男,河南新乡人,讲师,硕士,主要研究方向:医疗信息服务; 黄青松(1962—),男,湖南长沙人,教授,硕士,主要研究方向:智能信息系统、信息检索。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

意图模板; Chen 等[5] 考虑到消费意图语料的匮乏,在消费意图表达具有相似性的假设下提出了跨领域迁移学习(transfer learning) 的消费意图检测方法。这些传统的意图识别方法一般是基于模板匹配或人工特征集合,费时费力、扩展性不强。

针对上述问题,本文把聊天机器人中用户就医意图识别看作文本分类问题,即明确没有就医意图、极小可能就医、可能就医、极大可能就医和明确具有就医意图; 同时考虑医疗领域聊天文本的特点,即长度短、包含多轮回话的上下文信息和领域专有词,比如医院、医生等,构建了基于短文本主题模型(Biterm Topic Model,BTM) 和双向门控循环单元(Bidirectional Gated Recurrent Unit,BiGRU) 的意图识别模型(BTM-BiGRU) 来进行用户就医意图识别。针对领域文本特点,本文使用主题作为文本特征,首先通过 BTM 对用户聊天文本进行主题挖掘,相较于狄利克雷分布(Latent Dirichlet Allocation,LDA) 主题模型,BTM 对短文本的适应效果更好[6] 。然后结合深度学习方法,将上述 BTM 得到表示连续语句的特征向量集送入 BiGRU 中进行完整上下文学习,最后通过 Softmax[7] 输出分类结果实现本文任务。与支持向量机(Support Vector Machine, SVM) [8] 等传统机器学习方法相比,BiGRU 能够更好地对用户多轮对话进行建模,充分利用上下文信息来提取用户聊天文本特征。实验结果证明该混合模型结合了 BTM 和 BiGRU 的优点,相比传统方法,能更有效进行用户就医意图识别。

1　相关研究

1.1　词向量

在使用机器学习进行自然语言处理时,第一步肯定是将实际的文本内容变成计算机能识别的表示形式,即将要处理的信息数学化[9] 。向量空间模型是目前自然语言处理中的主流模型,其中词向量则是最基础和重要的。词向量常见的一种表达方式是 one-hot representation,它的向量维度是整个语料库中词个数,每一维代表语料库中一个词,其中 1 代表出现,反之为 0。很明显,one-hot representation 不仅存在维度灾难,而且最大问题是它只表达词本身是否出现,而没有表达词与词之间的关联。为此,便有了通过目标词上下文来预测目标词从而得到词向量的方法,称为 distributed representation。相应地,句向量、段向量等也就可以在此基础上得到。如今最常用的词向量训练方式是 word2vec[10] ,本文也使用它来训练词向量。

1.2　主题模型

近几年,许多需要大规模文本分析的领域都成功应用了主题模型[11] ,包括自然语言处理、数据挖掘、商业智能、信息检索等。首先关于主题,它是一个概念、一个方面,表现为一系列相关的词语。例如一个文档如果涉及“医院”这个主题,那么“医院”“医生”等词语便会以较高的频率出现。用数学语言描述,主题就是词汇表上词语的条件概率分布,与主题关系越密切的词语,它的条件概率越大,反之则越小[12] 。而主题模型作为语义挖掘的利器,则是一种对文字隐含主题进行建模的方法。主题模型中最具代表性的是 Hofmann[13] 提出的基于概率潜在语义分析(Probabilistic Latent Semantic Analysis,PLSA) 模型和 Blei 等[14] 提出的 LDA 模型。而本文所采用的 BTM 则是 Cheng 等[15] 提出的针对短文本学习的主题模型,该模型通过词对共现模式加强主题学习。

2　基于 BTM-BiGRU 的用户就医意图识别

2.1　BTM

针对短文本,由于其数据稀疏,如果根据传统的词共现方式来进行主题挖掘,效果将很不理想。为此,Cheng 等[15] 利用词对共现来代替词共现,提出一种短文本主题模型(BTM) 。BTM 结构如图 1 所示,其中各参数的含义如表 1 所示。

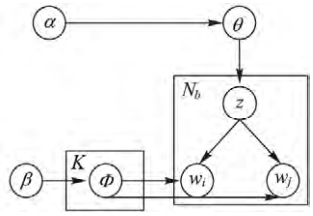

图 1　BTM 结构

Fig. 1　Structure of BTM

表 1　BTM 各参数含义

Tab. 1　Meanings of BTM's parameters

参数	含义
K	主题数目
N_b	语料中所有词对的数目
α	主题的狄利克雷先验参数概率分布
β	主题词的狄利克雷先验参数概率分布
θ	文档-主题概率分布
Φ	主题-词概率分布
z	主题

每篇文档的主题分布产生过程如下:

假设文档的主题概率等于此文档生成的词对的主题概率的期望值,如式(1) 所示:

$$P(z \mid d) = \sum_b P(z \mid b)\, P(b \mid d) \tag{1}$$

其中: z 表示主题, d 表示文档, b 表示词对。

通过贝叶斯公式能够得到词对的主题概率 $P(z \mid b)$,如式(2) 所示:

$$P(z \mid b) = \frac{P(z)\, P(w_i \mid z)\, P(w_j \mid z)}{\sum_z P(z)\, P(w_i \mid z)\, P(w_j \mid z)} \tag{2}$$

其中: $P(z) = \theta_z$, $P(w_i \mid z) = \phi_{i \mid z}$, w_i、w_j 表示某个词对 b 中的两个不同的词, θ_z 表示从主题分布 θ 中抽取一个主题 z, $\phi_{i \mid z}$ 表示主题-词分布 ϕ_z 中词 w_i 对应值。

用文档中词对的经验分布估计 $P(b \mid d)$,如式(3) 所示:

$$P(b \mid d) = \frac{n_d(b)}{\sum_b n_d(b)} \tag{3}$$

其中, $n_d(b)$ 为文档 d 中词对 b 的出现次数。

2.2　基于门控循环单元的双向循环神经网络

2.2.1　门控循环单元

针对普通循环神经网络(Recurrent Neural Network, RNN) [16] 存在的两大问题,即长距离依赖和梯度消失或梯度爆炸,Hochreiter 等[17] 提出了长短期记忆(Long-Short Term Memory, LSTM) 模型。相比传统 RNN,LSTM 的重复神经网络模块更复杂,增加了门结构,即遗忘门、输入门以及输出门。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

三个门的计算也造成了 LSTM 训练时间较长，而门控循环单元(Gated Recurrent Unit，GRU)[18]作为 LSTM 的一个变体，在保持其学习效果的同时又使结构更加简单，节省训练时间。GRU 只有重置门 r_t 和更新门 z_t，其中更新门由遗忘门和输入门合成，其工作流程具体如下：

和 LSTM 一样，GRU 的关键是元胞状态。首先决定要从旧元胞状态和当前输入中丢掉哪些信息，由重置门来控制其程度，如式(4)所示：

$$r_t = \sigma(W_r x_t + U_r h_{t-1}) \tag{4}$$

其中：σ 代表 Sigmod 非线性函数，x_t 代表当前输入，h_{t-1} 代表上一时刻隐层的输出。

接下来是决定将哪些新信息保存到元胞状态，具体分为两部分：

1) 更新门用来控制忘记之前信息和添加新信息的程度，如式(5)所示：

$$z_t = \sigma(W_z x_t + U_z h_{t-1}) \tag{5}$$

2) 由 tanh 层创造一个新的候选值 $\tilde{h}_t$，该值可能会加入到元胞状态中，如式(6)所示：

$$\tilde{h}_t = \tanh(W_h x_t + r_t \odot U_h h_{t-1}) \tag{6}$$

其中：tanh 代表双曲线正切函数，$\odot$ 代表元素级相乘运算。

最后，把这两个值组合起来用于更新旧元胞状态 h_{t-1} 到新元胞状态 h_t，如式(7)所示：

$$h_t = (1 - z_t) h_{t-1} + z_t \tilde{h}_t \tag{7}$$

2.2.2 双向门控循环单元

双向 GRU 是 GRU 的改进。由于 GRU 是单方向推进，往往忽略了未来的上下文信息，而双向 GRU 的基本思想是使用同一个训练序列向前向后各训练一个 GRU 模型，再将两个模型的输出进行线性组合，使得序列中每一个节点都能完整地依赖所有上下文信息。对于本任务，在每一个时间步上充分利用过去和未来的上下文，将有助于更好理解用户的就医意图。双向 GRU 的结构如图 2 所示。

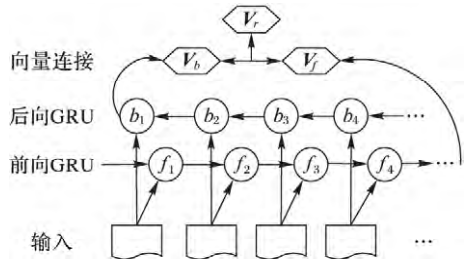

图 2 双向 GRU 模型结构

Fig. 2 Structure of BiGRU

2.3 BTM-BiGRU 混合模型

2.3.1 混合模型结构

该混合模型的结构如图 3 所示，由 BTM 层、Sentence embedding 层、BiGRU 层和最终的 Softmax 层组成。

1) BTM 层：按照聊天顺序将聊天语句逐句送入 BTM 层，利用 BTM 挖掘出每句的文档 - 主题概率分布 $P(z \mid d)$ 的最大值 $p = (z \mid d)_{\max}$ 对应主题下的主题-词分布前 N 个词。

2) 语句嵌入层：将 BTM 层得到的主题词作为特征项并用词向量表示，特征词对应权重由整个实验语料的 TF(Term Frequency)和 IDF(Inverse Document Frequency)确定，如式(8)所示，从而得到句子的特征向量。假设长度为 n 句话的用户聊天文本，令 $\boldsymbol{S}_i \in \mathbf{R}^k$ 代表聊天中第 i 句话的句向量，则整个用户聊天文本表示为 $\boldsymbol{CR} \in \mathbf{R}^{nk}$，如式(9)所示：

$$w_{ik} = TF_{ik} \times IDF_{ik} \tag{8}$$

其中，w_{ik} 代表第 i 句话的第 k 个主题特征词权重。

$$\boldsymbol{CR} = \boldsymbol{S}_1 \oplus \boldsymbol{S}_2 \oplus \cdots \oplus \boldsymbol{S}_n \tag{9}$$

其中，$\oplus$ 代表将句向量依次进行拼接操作。

假设经过滤得到某用户的聊天对话：$D = S1 + S2$，其中 $S1$ =“最近我这一直胸闷，想去医院看看”，$S2$ =“请问我该挂什么号”。通过 BTM 层后分别得到 $S1$ 的最大主题概率下的前 2 个主题词为“医院”“医生”，$S2$ 的为“挂号”“科室”。利用 word2vec 分别得到医院、医生、挂号、科室的词向量，假设为 $\boldsymbol{x}_1$、$\boldsymbol{x}_2$、$\boldsymbol{x}_3$、$\boldsymbol{x}_4$，其对应 TF-IDF 权重为 w_{11}、w_{12}、w_{21}、w_{22}，则 $S1$ 的句向量 $\boldsymbol{S}_1 = w_{11} \cdot \boldsymbol{x}_1 \oplus w_{12} \cdot \boldsymbol{x}_2$，$S2$ 的句向量 $\boldsymbol{S}_2 = w_{21} \cdot \boldsymbol{x}_3 \oplus w_{22} \cdot \boldsymbol{x}_4$。

3) BiGRU 层：由前向 GRU 和后向 GRU 组成。分别将用户聊天文本向量 $\boldsymbol{CR}$ 顺序、逆序输入前向 GRU 和后向 GRU，得到两个方向的连续语句表示 F_CR' 和 B_CR'，将其拼接起来作为用户多轮会话的最终表示。

4) Softmax 层：本任务一共有 5 个输出，即该用户明确没有就医意图、极小可能就医、可能就医、极大可能就医和明确具有就医意图，因此混合模型的 Softmax 层输出维度为 5。Softmax 层的输出是判别类别的概率，即根据条件概率的值来判断聊天文本到底属于哪类用户意图。

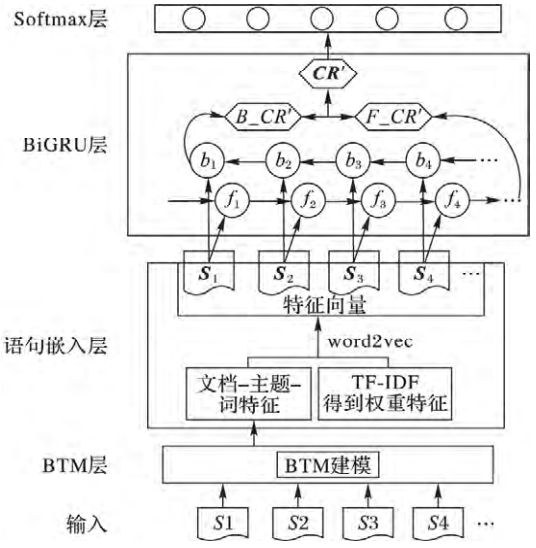

图 3 BTM-BiGRU 结构

Fig. 3 Structure of BTM-BiGRU

2.3.2 混合模型优点

本文综合 BTM 和双向 GRU 的优点，构建了 BTM-BiGRU 混合模型。正如前文分析所得，由于 BTM 采用词对共现模式代替传统的词共现模式，因此可以更好地对短文本进行主题特征挖掘。GRU 不仅保持 LSTM 可以对文本的上下文信息进行有效刻画，解决文本长期依赖问题的优势，而且只有两个门的计算，参数减少，节省了模型的训练时间。但不管是 LSTM 还是 GRU，都有一个问题，它是从左往右推进的，后面的输入会比前面的更重要。很明显对于本任务来说这是不妥的，因为聊天中各语句应该是平权的，因此本文采用双向 GRU 来完整捕捉上下文信息。

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

3　实验分析

3.1　实验数据及评价指标

由于目前公开的医疗聊天语料较少,因此本文使用的实验数据是通过爬虫程序在浙江大学附属第一医院爬取下来的用户聊天语料,共计 18 822 条。通过关键字及模板规则过滤掉对意图没有贡献的语句,再人工将同一用户的一句或多句话归为一组,共分了 5 660 组数据。然后由两名标注人员各自对其标注意图类别,通过匹配两名标注人员的标注结果,去掉不一致,最终得到标注结果一致的 5 425 组数据,如表 2 所示。其中,明确没有就医意图有 891 组,极小可能就医有 538 组,可能就医有 1 868 组,极大可能就医有 782 组,明确具有就医意图有 1 346 组。另外,为避免实验结果的偶然性,在进行对比实验时,采用 5 折交叉验证,即将数据集分成 5 等份,其中 1 份作为测试集,其余 4 份作为训练集,进行循环实验,得到 5 次分类结果。

本文的用户就医意图识别本质是多分类问题,因此采用文本分类器中常用的评估指标来对每种类型进行评价,即准确率 P、召回率 R 以及二者的综合评价 F-$measure$ 值,如式(10) ~ (12)所示:

$$P = TP/(TP + FP) \tag{10}$$

$$R = TP/(TP + FN) \tag{11}$$

$$F\text{-}measure = 2RP/(R + P) \tag{12}$$

其中: 属于类 A 的样本被正确分到类 A 中,记这一类样本数为 TP(True Positive);属于类 A 却被错误分到类 A 以外的其他类,记这一类样本数为 FN(False Negative);不属于类 A 被正确分到类 A 以外的其他类,记这一类样本数为 TN(True Negative);不属于类 A 却被错误分到类 A 中,记这一类样本数为 FP(False Positive)。

关于整体性能的评价,用准确率 P、召回率 R 和 F-$measure$ 的期望,其中每种类型的权重与其对应语料数量成比例。

表 2　实验语料举例

Tab. 2　Examples of the corpus

编号	用户聊天文本	意图
1	我妈有糖尿病,平时需要注意什么	明确没有
2	糖尿病除了尿多还有别的啥明显特征	极小
3	今早起来,肚子突然不舒服 没有拉肚子,还有点恶心 没有别的不舒服了	可能
4	上个月底到现在半个多月了,一直咳嗽 吃了药店拿的药也不见效果 会不会咳出肺炎了	极大
5	最近我这一直胸闷,想去医院看看 请问我该挂什么号	明确具有

3.2　参数设置

BTM-BiGRU 混合模型的参数设置中,BTM 部分,将主题数设为 $K = 8$,$\alpha = 50/K = 6.25$,$\beta = 0.01$,在不影响实验效果的情况下选取主题-词概率分布中前 $N = 2$ 个词作为文档特征。此基础上得到主题-词分布如表 3 所示。在 word2vec 时,设置词向量维度为 100。在 BiGRU 这部分设置输出维度为 128。另外,设置训练过程中 $batch\_size = 32$,为防止数据过拟合,还应使用 $dropout$ [19] 与 L2 正则化进行约束,其中 $dropout$ 应用于 BiGRU 层与 Softmax 分类层之间,并且 $dropout = 0.25$,而 L2 正则化则应用于最终的 Softmax 层。

表 3　主题-词分布

Tab. 3　Subject-terms distribution

主题序号	主题词	主题序号	主题词	主题序号	主题词
0	医院 医生	3	想 去	6	注意 小心
1	挂号 科室	4	有点 偶尔	7	可能 应该
2	平时 一般	5	非常 特别		

3.3　实验结果对比与分析

实验环境搭建在 VMware Workstation + Ubuntu + Linux 下。实验中,首先需要对上述实验数据进行清洗,去除数据中的杂质文本,然后采用 jieba 分词工具进行分词,之后使用 Google 开源提供的 word2vec 进行词向量模型的训练以量化文本。

3.3.1　不同方法效果比较

本文为与现有方法作更好对比,除基于 BTM 和 BiGRU 单模型外,还比较传统模板匹配、SVM 分类方法以及文献[20]中提到目前较好的基于卷积和长短期记忆网络(Convolution Neural Network and Long-Short Term Memory Network,CNN-LSTM)的方法。CNN-LSTM 利用 CNN 能够获取深层特征的优点,先对用户聊天文本提取局部代表性特征,然后为保证文本时序性,按照卷积先后顺序重新组合,再依次输入 LSTM 中进行上下文学习,最后得到较为理想的分类结果。不同方法下,整体性能表现如表 4 所示。

表 4　意图识别实验结果

Tab. 4　Experimental results of intention recognition

方法	P	R	F
模板匹配	0.803 3	0.701 5	0.749 0
SVM	0.856 3	0.743 2	0.804 1
CNN-LSTM	0.891 8	0.865 2	0.878 3
BTM	0.851 2	0.766 9	0.806 9
BiGRU	0.882 5	0.852 3	0.867 1
BTM-BiGRU	0.909 4	0.891 5	0.900 4

实验结果表明,在聊天机器人中用户就医意图识别任务上,本文的 BTM-BiGRU 方法效果最好。对比基于 BTM、SVM 方法与本文方法的实验结果,说明了 BiGRU 的优势,它能够充分地利用上下文建模。对比基于 BiGRU、CNN-LSTM 方法与本文方法的实验结果,则证明本任务上使用 BTM 挖掘的主题特征作为循环神经网络的输入效果更好。虽然基于模板匹配方法的效果不错,但其需要事先获取意图模板,比较费时费力,而且一旦模板不准确,效果将很差。

3.3.2　不同方法时间复杂度比较

实际应用中,不仅考虑方法效果,也需要关注其时间复杂度。考虑到传统基于模板匹配的意图识别方法,其模板获取所花费时间远远多于其余五种方法,因此这里主要比较基于 SVM、CNN-LSTM、BTM、BiGRU 以及 BTM-BiGRU 的方法执行时间,如表 5 所示。

这四种方法训练模型时均涉及迭代,迭代次数不同所用时间也就不同,这里均以最终实验效果最好为准得到各自执行时间。结果表明,时间最短的是基于 SVM 的方法,时间最长的是 CNN-LSTM,这体现了 GRU 的优势,相比 LSTM 能够节省训练时间。因此无论从方法效果还是时间复杂度上考虑,

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

本文提出的 BTM-BiGRU 方法都是行之有效的。

表5 各方法执行时间

Tab. 5 Execution time of different methods

方法	执行时间/s	方法	执行时间/s
SVM	76	BiGRU	320
CNN-LSTM	481	BTM-BiGRU	392
BTM	89		

4 结语

在线医疗聊天机器人中，如果能在聊天过程中识别出用户就医倾向，将便于进一步为用户提出合理治疗建议以及推荐相应科室。本文将任务当作文本分类问题，并针对医疗聊天文本的特点，提出了基于短文本主题模型和双向门控循环单元的意图识别模型 BTM-BiGRU 来进行用户就医意图识别。实验结果表明，该混合模型的整体分类准确率优于传统的用户意图识别方法以及目前较好的 CNN-LSTM 方法。

本文在不同阶段利用了 BTM 和 BiGRU 提取特征，后续工作将在特征工程上作进一步研究，寻找更好方法进行医疗聊天文本特征提取，从而继续提高聊天机器人中用户就医意图识别的效果。

参考文献 (References)

[1] NIE L, WANG M, ZHANG L, et al. Disease inference from health-related questions via sparse deep learning [J]. IEEE Transactions on Knowledge & Data Engineering, 2015, 27(8): 2107 - 2119.

[2] 卢婷婷. 基于短文本的互联网用户意图识别方法及应用研究 [D]. 济南: 济南大学, 2016: 6 - 10. (LU T T. Research on short texts based Internet users' intention recognition and application [D]. Jinan: University of Jinan, 2016: 6 - 10.)

[3] LI X. Understanding the semantic structure of noun phrase queries [C] // ACL '10: Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2010: 1337 - 1345.

[4] RAMANAND J, BHAVSAR K, PEDANEKAR N. Wishful thinking: finding suggestions and 'buy' wishes from product reviews [C] // CAAGET '10: Proceedings of the NAACL HLT 2010 Workshop on Computational Approaches to Analysis and Generation of Emotion in Text. Stroudsburg, PA: Association for Computational Linguistics, 2010: 54 - 61.

[5] CHEN Z, LIU B, HSU M, et al. Identifying intention posts in discussion forums [C] // NAACL 2013: Proceedings of the 2013 Conference of the North American Chapter of the Association for Computational Linguistics. Stroudsburg, PA: Association for Computational Linguistics, 2013: 62 - 71.

[6] 汤秋莲. 基于 BTM 的短文本聚类 [D]. 合肥: 安徽大学, 2014: 17 - 19. (TANG Q L. The short text clustering based on BTM [D]. Anhui: Anhui University, 2014: 17 - 19.)

[7] 阳馨, 蒋伟, 刘晓玲. 基于多种特征池化的中文文本分类算法 [J]. 四川大学学报(自然科学版), 2017, 54(2): 287 - 292. (YANG X, JANG W, LIU X L. Chinese text categorization based on multi-pooling [J]. Journal of Sichuan University (Natural Science Edition), 2017, 54(2): 287 - 292.)

[8] LI C, XU Y. Based on support vector and word features new word discovery research [M] // ISCTCS 2012: Proceedings of the 2012 International Conference on Trustworthy Computing and Services. Berlin: Springer, 2013: 287 - 294.

[9] 周昭涛. 文本聚类分析效果评价及文本表示研究 [D]. 北京: 中国科学院计算技术研究所, 2005: 32 - 36. (ZHOU Z T. Quality evaluation of text clustering results and investigation on text representation [D]. Beijing: Institute of Computing Technology, Chinese Academy of Sciences, 2005: 32 - 36.)

[10] MIKOLOV T, SUTSKEVER I, CHEN K, et al. Distributed representations of words and phrases and their compositionality [J/OL]. arXiv Preprint, 2013, 2013: arXiv: 1310.4546 (2013-10-16) [2017-11-06]. https://arxiv.org/abs/1310.4546.

[11] 王李东, 魏宝刚, 袁杰. 基于概率主题模型的文档聚类 [J]. 电子学报, 2012, 55(4): 77 - 84. (WANG L D, WEI B G, YUAN J. Document clustering based on probabilistic topic model [J]. Acta Electronica Sinica, 2012, 55(4): 77 - 84.)

[12] 高章敏, 何祥, 刘嘉勇, 等. 基于主题模型的中文词义归纳 [J]. 四川大学学报(自然科学版), 2016, 53(6): 1269 - 1272. (GAO Z M, HE X, LIU J Y, et al. Chinese word sense induction based on topic model [J]. Journal of Sichuan University (Natural Science Edition), 2016, 53(6): 1269 - 1272.)

[13] HOFMANN T. Probabilistic latent semantic indexing [C] // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. New York: ACM, 1999: 50 - 57.

[14] BLEI D M, NG A Y, JORDAN M I. Latent dirichlet allocation [J]. The Journal of Machine Learning Research, 2003, 3: 993 - 1022.

[15] CHENG X, YAN X, LAN Y, et al. BTM: topic modeling over short texts [J]. IEEE Transactions on Knowledge and Data Engineering, 2014, 26(12): 2928 - 2941.

[16] LeCUN Y, BENGIO Y, HINTON G E. Deep learning [J]. Nature, 2015, 521(28): 436 - 444.

[17] HOCHREITER S, SCHMIDHUER J. Long short-term memory [J]. Neural Computation, 1997, 9(8): 1735 - 1780.

[18] CHO K, VAN MERRIËNBOER B, GULCEHRE C, et al. Learning phrase representations using RNN encoder-decoder for statistical machine translation [C] // EMNLP 2014: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing. Stroudsburg, PA, USA: Association for Computational Linguistics, 2014: 1724 - 1734.

[19] SRIVASTAVA N, HINTON G, KRIZHEVSKY A, et al. Dropout: a simple way to prevent neural networks from over-fitting [J]. Journal of Machine Learning Research, 2014, 15: 1929 - 1958.

[20] 钱岳, 丁效, 刘挺, 等. 聊天机器人中用户出行消费意图识别 [J]. 中国科学: 信息科学, 2017, 47(8): 997 - 100. (QIAN Y, DING X, LIU T, et al. Identification method of the user's travel consumption intention in chatting robot [J]. SCIENTIA SINICA Informationis, 2017, 47(8): 997 - 100.)

This work is partially supported by the National Natural Science Foundation of China (81360230, 81560296).

YU Hui, born in 1993, M. S. candidate. Her research interests include natural language processing, medical information service.

FENG Xupeng, born in 1986, M. S., assistant experimentalist. His research interests include information retrieval.

LIU Lijun, born in 1978, M. S., lecturer. His research interests include medical information service.

HUANG Qingsong, born in 1962, M. S., professor. His research interests include intelligent information system, information retrieval.

(C)1994-2021 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net