



基于自然语言学习的智能云导诊技术

汤人杰¹, 江涛², 杨巧节¹

(1. 中国移动通信集团浙江有限公司, 浙江 杭州 310016;

2. 浙江省卫生计生信息中心, 浙江 杭州 310006)

摘要: 分析了目前导诊系统存在的主要问题, 根据目前学术界研究的智能导诊系统现状, 创造性地提出了基于海量患者病情自述, 利用多种机器学习算法, 形成了病情自述自学习体系, 实现了针对病人自述的智能导诊。该平台利用网络爬虫技术获取了国际国内主流医疗机构科室设置以及海量的病人病情描述, 形成了基于国际标准的病理知识库及病情自述知识库, 为后续的智能文本识别奠定了基础。同时在算法上进行了创新, 构建了注意力(attention)与文本卷积(TextCNN)组合模型, 提升了导诊的准确性。

关键词: 智能导诊; attention 模型; TextCNN

中图分类号: TP183

文献标识码: A

doi: 10.11959/j.issn.1000-0801.2019078

Intelligent cloud guidance technology based on natural language learning

TANG Renjie¹, JIANG Tao², YANG Qiaojie¹

1. China Mobile Communications Group Zhejiang Co., Ltd., Hangzhou 310016, China

2. Health and Family Planning Information Center of Zhejiang Province, Hangzhou 310006, China

Abstract: The main problems existing in the current guidance system were analyzed. According to the intelligent guidance system researched in the current academic circles, a self-learning system for the disease which based on a large number of patient condition descriptions using a variety of machine learning algorithms was creatively proposed. The platform used Web crawler technology to obtain setting of the international and domestic medical institution departments and a large number of patient descriptions, which form a pathological knowledge base and a disease self-reported knowledge base. At the same time, the algorithm was innovatively constructed to combine the attention model with TextCNN to improve the accuracy of the consultation.

Key words: intelligent guidance, attention model, TextCNN

1 引言

随着信息化技术的发展, 越来越多患者开始

采用互联网手段进行在线预约与挂号, 但是由于患者自身医疗知识有限, 很难准确地判断自己要挂号的科室, 因此在预约挂号中经常会出现科室

收稿日期: 2019-03-22; 修回日期: 2019-04-10

通信作者: 江涛, 32037122@qq.com



选择错误等问题,这给患者带来了很大的不便,急需导诊系统来帮助患者提高挂号的准确性。但现有的智能导诊系统无论在推荐的准确性还是在使用的便利性上都难以满足病人实际的要求,需要新型的、更加准确的、普适性更强的导诊系统来支持病患挂号。

2 目前导诊服务现状

目前现有的导诊服务分为以下3种。

(1) 就医目录指南

就医目录指南是最普遍的导诊服务形式,这类服务主要通过设置疾病目录或者症状目录给患者提供就医科室、医生专家的推荐。目前大多数医院官网和第三方医疗卫生服务平台都采用这样的形式。就医目录在一定程度上能够实现导诊功能,但是需要患者对自己所患的疾病有明确的认识,显然这样的要求对于患者来讲是很难实现的,尤其是对初次患病的患者。

(2) 人工在线咨询

人工在线咨询是很多民营医疗机构经常采用的导诊服务形式。该服务通过微信、QQ、在线聊天工具、电话交流等形式接收患者咨询,并向患者提供就诊建议。这类服务的人性化、专业性比较好,但服务质量受咨询服务人员专业水平的影响比较大,特别是在缺乏监管的情况下,会出现欺诈问题,很多就医人工咨询服务普遍存在诱导就医、夸大病情等问题。另外,人工咨询服务运营成本高、服务能力比较有限,很难大规模地为患者提供导诊服务。

(3) 智能导诊服务

智能导诊服务是采用人工智能技术识别患者描述的病情并给出就医指导。这类服务效率高,不会受到人为因素的影响,更加客观。但是这类服务对于算法和数据资料的质量要求比较高。对于患者自我病情描述的理解水平直接影响后续疾病的判断,同时推理算法的准确率和效率也影响

咨询的质量。

近年来国内外在医疗智能导诊方面的研究很多,López-Nores 等^[1]引入了一种新的过滤策略,这种算法对于有特定疾病的患者效率很高。Pattaraintakorn 等^[2]提出了一种使用粗糙集和规则分析的医疗导诊系统,主要目标是将病人自身的身体数据作为条件属性,推荐临床检查方式。徐守坤等^[3]提出一种医生资源均衡推荐算法,使用最佳结果优先的算法,解决了医疗资源使用过度集中而产生的问题。

3 现有智能导诊系统存在的问题

目前业界针对智能导诊的研究,主要是根据患者自身的病情描述,通过分词的方法,形成病情的关键字,通过病情的关键字与某类疾病的医学解释核心词进行相似度匹配,根据相似度匹配的高低进行疾病的匹配,进而推荐相应的科室。

以上智能导诊系统会存在以下问题。

(1) 患者病情描述模糊,病理关键词不清晰。病人在描述病情的时候往往难以做到很准确地描述,仅从有限的病情关键词上很难进行有效的判断。

(2) 相似度推荐问题难以解决病症相似的病情。通过病情关键词与病理进行相似度的匹配,往往会出现相似病症的多个科室的推荐,如感冒、鼻炎、早期哮喘、乙肝早期这些疾病和感冒的症状比较类似,因此很容易形成错误的推荐。

(3) 现有的智能导诊系统更多地从文本关键词的角度来分析病情,而没有更多地去考虑病人本身的因素,如年龄、性别、地域、病史等影响患者发病的因素,同样的症状在不同特征的人群中代表的疾病大相径庭,因此仅从文字本身确定病情有很大的局限性。

王志宏等^[4]提出了通过人工智能技术构建智能服务的体系,从理论上证实了人工智能技术在智能推荐领域的应用,为智能导诊的发展提出了

新的方向。

4 智能导诊系统的建设思路

基于目前人工智能领域的研究,针对现有智能导诊系统存在的问题,本文从大数据的角度出发,利用网络爬虫工具爬取的获得 200 多万条患者的问诊数据、患者个人信息和地域以及其所对应挂号科室,形成导诊训练数据,再利用 BERT 模型和 TextCNN 模型,训练形成病症分类器。另外,为了提高导诊的准确性,针对判别概率较小的病例,系统会要求患者增加病情的描述,提高导诊的准确性,以区分相似病症的不同疾病,提升导诊的准确性。

本文的导诊系统有以下特点。

(1) 算法上进行了创新,从巨量的患者自述病情来学习挂号的科室,摒弃了传统通过文本分解通过关键词相似度来方法,提升了算法的准确性和适应性。

(2) 导诊系统中充分考虑了病人自身的年龄、性别、生活环境等影响病症的重要因素,使最终的推荐能够因人而异,更加准确地判断患者的病症。

(3) 导诊系统与挂号相结合,不仅能实现挂号科室的推荐,同时也能够根据病人的情况推荐合适的医生与医院,并能够实现在线挂号,全面提升了患者的便利性,真正实现患者智能导诊。

(4) 模型直接应用于挂号 App,可以形成“数

据—建模—模型应用—新的数据—模型更新……”闭环,实现模型的不断完善和提升,进而形成自学习过程。

5 云智能导诊系统整体框架

云智能导诊采用交互的方式,患者从 App 或微信公众号进入智能导诊模块,选择自己导诊或给他人问诊,他人导诊需补充患者年龄、性别等基本信息。AI 机器人通过患者的语音或者文字描述病症,推荐相应的科室,若推荐科室置信度小于 50%,则要求患者进一步补充症状,患者选择相应科室之后,系统从挂号平台中根据医院等级和距离等进行综合排名,推荐最合适的就诊医院,并对接至科室中的医生平台智能提供导诊科室。

云智能导诊流程如图 1 所示。

整体系统由移动端、导诊后台、挂号平台和语音服务平台四大部分构成。

- 移动端:作为与用户会话的入口,负责语音与文本的互转,调用导诊后台做语音或文字应答,并在会话后调用和连接挂号平台。
- 导诊后台:根据会话文本及上下文,反复应答,直至给出挂号建议。
- 挂号平台:接受用户的挂号请求,并定时向导诊后台提供动态的医院/科室/医生/信息;并定期向导诊后台同步实际挂号信息。
- 语音服务平台:提供语音转换接口,将语音转文本,并回显到 App 端,同时 App 后

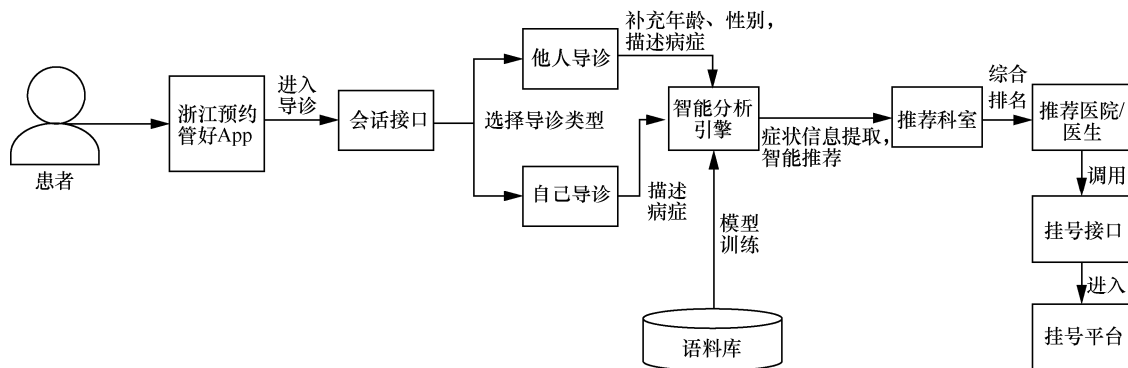


图 1 云智能导诊流程



台将文本发送给导诊后台处理。

本系统核心模块智能导诊系统主要分离线深度学习模型构建、在线客户输入解析、在线基于深度学习的科室分类、在线用户挂号科室推荐 4 个部分。

离线深度学习模型构建模块是基础模块，为其他 3 个线上模块提供基础能力：为在线的客户输入解析提供词典；为在线基于深度学习的科室分类模块提供深度学习模型；为在线用户挂号科室推荐模块提供科室字典。智能导诊模块架构如图 2 所示。

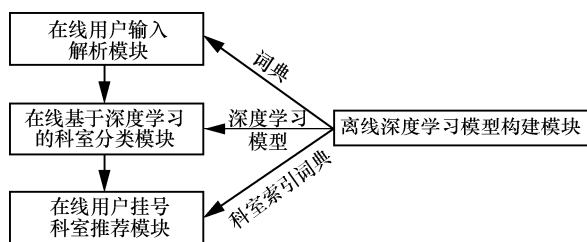


图 2 智能导诊模块架构

6 云智能导诊建模过程与模型说明

6.1 离线深度学习模型构建

6.1.1 数据准备

利用爬虫等方式获取 200 多万条涉及 50 多个二级科室的问诊数据，作为训练数据。

6.1.2 数据处理

数据处理主要有以下步骤。

步骤 1 建立科室与国标科室的对应关系。

步骤 2 创建国标科室索引。

步骤 3 利用分词工具，对问诊内容数据进行分词，并去停词，加入医学专有词。

步骤 4 统计词频建立词典，例：{'Unkown': 0, ' ': 1, '不适': 2, '术后': 3, '复查': 4, '疼痛': 5, '伴': 6, '复诊': 7, '左': 8, '咳嗽': 9, ...}。

步骤 5 问诊数据词袋化。

6.1.3 数据建模过程

数据建模过程主要有以下步骤。

步骤 1 对处理后的问诊数据进行词向量化，即 Embedding，这样做能有效降低运算量，同时使高维的稀疏表示，映射到低维的向量表示，保存了更多的语义信息。

步骤 2 在模型中加入了谷歌刷新了 NLP 多项记录的 BERT 模型中的核心模块多头注意力 (multi-head attention) 模块。该模块在导诊任务中，聚焦于辨识性更高的词汇，增加患者有效症状的权重，使得科室分类更加准确。

步骤 3 TextCNN 提取局部特征。多头注意力模块构建了加权后的全部信息，但不能提取局部特征。TextCNN 可以更好地捕捉局部特征。

步骤 4 Dropout 正则化，提高模型的泛化能力。

步骤 5 全连接层，对前一层的输出特征进一步抽象化。

步骤 6 Dropout 正则化，提高模型的泛化能力。

步骤 7 分类。

6.1.4 自学习标签修正

数据本身存在 20%~30% 的脏数据 (即错误标签数据)。为克服这个难题，通过自学习的方式，对数据标签进行重标注，如图 3 所示，主要步骤如下。

步骤 1 对带有错误标注的全量数据建模。

步骤 2 用生成的模型去预测全量数据的标签。

步骤 3 将预测置信度大于 90% 的数据抽取出来，标签为预测标签。

步骤 4 用新的数据重复步骤 2 的模型训练步骤。

步骤 5 直到模型的精度满足需求。

6.2 在线客户输入解析

客户在线输入患者的基本症状：

步骤 1 利用分词工具 jieba 对输入进行分词处理；

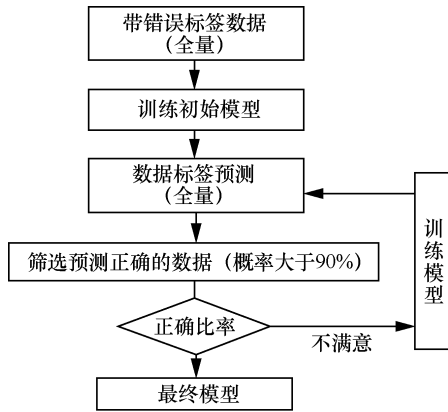


图3 自学习修正体系

步骤 2 将标点符号等与症状无关的符号和部分单词从输入中删除；

步骤 3 对步骤 1 和步骤 2 生成的输入构建索引；

步骤 4 生成索引后的词袋向量。

6.3 在线基于深度学习的科室分类

线上采用了 Python 的 Flask 作为后端的 Web 服务引擎，利用 Gunicorn 作为负载均衡的工具。Flask 服务启动时加载深度学习模型。服务模型接收在线客户输入解析模块的输出作为输入，输出国标科室的索引。

6.4 在线用户挂号科室推荐

将在线基于深度学习的科室分类模块的输出索引映射为标准的国标科室，同时向客户推荐最合适的科室。

7 核心模型说明

7.1 注意力模型

注意力 (attention) 模型是由编码器 (encoder) 和解码器 (decoder) 两个模块以及编码器和解码器堆叠的自注意力 (self-attention) 模块和全连接层 (fully-connected layer) 所构成的。模型共包含 3 个注意力成分，分别是编码器的自注意力、解码器的自注意力以及连接编码器和解码器的注意力。3 个注意力模块都是多头注意力的形式，输入都是 Q 、 K 、 V 3 个元素， Q 、 K 、 V 的取值是不同

的。 Q 矩阵里面装的就是全部的词嵌入， K 、 V 矩阵也是一样的情况。记词嵌入的“尺度”为 d 模型， Q 的就是 $n \times d$ 模型， K 、 V 也一样，第 i 个词的嵌入为 v_i ，该词的注意力应为：

$$\text{Attention}(Q_i, K, V) = \text{softmax}(v_i[v_1^T, v_2^T, \dots, v_n^T])$$

$$\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} = \text{softmax}(Q_i, K^T)^V \quad (1)$$

其中， Q 为编码器对输入句子的词进行编码所形成的向量，也就是将输入句子进行非线性变换。 K 为解码器对输出句子的词进行编码，也就是将输入句子进行非线性变换。 V 为编码器和解码器共同对输入句子的词进行的编码， v 为单个词的值。同时做全部所有词的注意力，则为：

$$\text{Attention}(Q, K, V) =$$

$$\text{softmax} \left(\begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} [v_1^T, v_2^T, \dots, v_n^T] \right) \begin{bmatrix} v_1 \\ v_2 \\ \vdots \\ v_n \end{bmatrix} =$$

$$\text{softmax}(QK^T)^V \quad (2)$$

多头注意力是缩放点注意力的叠加。首先对 Q 、 K 、 V 做线性变换，接着对新生成的 Q' 、 K' 、 V' 计算注意力，重复操作 h 次，然后把 h 次的结果做连接组合，并进行一次线性变换，得到多头注意力输出：

$$\text{MultiHead}(Q, K, V) =$$

$$\text{Concat}(\text{head}_1, \dots, \text{head}_h)W^O \quad (3)$$

$$\text{where head}_i = \text{Attention}(QW_i^Q, KW_i^K, VW_i^V)$$

其中， head 为一次注意力计算， W 为各个注意力的权重。

7.2 文本积卷神经网络模型

文本积卷神经网络 (TextCNN) 是利用卷积神经网络对文本进行分类的算法，包含了嵌入层、卷积池化层 (convolution and pooling) 两大模块。



(1) 嵌入层

文本卷积神经网络使用预先训练好的词向量作嵌入层 (embedding layer)。对于数据集里的所有词, 每个词都可以表征成一个向量, 因此可以得到一个嵌入矩阵 M , M 里的每一行都是词向量。 M 可以是静态的、固定不变的, 也可以是非静态的, 可以根据反向传播更新。

(2) 卷积池化层

• 卷积 (convolution)

输入一个句子, 首先对这个句子进行切词, 假设有 s 个单词。对每个词, 跟句嵌入矩阵 M , 可以得到词向量。假设词向量一共有 d 维, 对于该句, 便可以得到 s 行 d 列的矩阵 $A \in R_{s \times d}$ 。矩阵 A 看成是一幅图像, 使用卷积神经网络提取特征。由于句子中相邻的单词关联性很高, 因此可以使用一维卷积。卷积核的宽度是词向量的维度 d , 高度是可以设置的超参数。现假设有一个卷积核是一个宽度为 d 、高度为 h 的矩阵 w , 那么 w 有 $h \times d$ 个参数需要被更新。对于一个句子, 经过嵌入层之后可以得到矩阵 $A \in R_{s \times d}$ 。 $A[i:j]$ 表示 A 的第 i 行到第 j 行, 卷积操作可以用式 (4) 表示:

$$O_i = w \cdot A[i:i+h-1], i=1,2,\dots,s-h+1 \quad (4)$$

其中, O_i 表示中间层的值。再叠加上偏置 b , 再使用激活函数 f 激活, 得到所需的特征值, 具体如下:

$$c_i = f(o_i + b) \quad (5)$$

对一个卷积核, 可以得到特征 $c \in R_{s-h+1}$, 总共有 $s-h+1$ 个特征。可以使用更多高度不同的卷积核, 得到更丰富的特征表达。

• 池化 (pooling)

不同尺寸的卷积核得到的特征图 (feature map) 大小也是不一样的, 因此对每个特征图使用池化函数, 使它们的维度相同。最常用的就是 1-max 池化, 提取出特征图照片那个的最大值。因此每一个卷积核得到特征就是一个值, 对所有卷积核使用 1-max 池化, 再级联起来, 可以得到

最终的特征向量, 这个特征向量再输入 softmax 层做分类。此处可以使用退出机制 (dropout) 防止过拟合。

8 实验验证

针对“浙江预约挂号”App 用户进行了验证, 针对患者在挂号平台的病情进行科室推荐, 患者一次导诊正确率 63.10%, 二次交互导诊成功率为 96.35%, 具体见表 1。

表 1 科室推荐的实验验证

科室代码	科室名称	一次导诊正确率	二次交互导诊正确率
0	皮肤科	78.42%	95.14%
1	骨科	81.17%	95.63%
2	妇产科	58.20%	94.83%
3	妇科	49.67%	95.37%
4	眼科	90.74%	98.10%
5	泌尿外科	75.06%	96.45%
6	口腔科	85.79%	97.41%
7	耳鼻喉科	79.61%	97.25%
8	消化内科	65.18%	96.32%
9	内科	61.91%	96.32%
10	普通外科	50.54%	94.18%
11	生殖健康与不孕症	49.10%	93.76%
12	神经内科	58.37%	97.33%
13	精神科	56.17%	96.21%
14	神经外科	80.62%	98.26%
15	内分泌	63.92%	96.16%
16	心血管内科	56.11%	95.80%
17	胸外科	59.36%	96.96%
18	肿瘤科	52.10%	95.66%
19	呼吸内科	48.32%	95.39%
20	肛肠科	57.85%	96.84%
21	免疫学	60.92%	97.50%
22	肾病学	53.84%	96.80%
23	血液内科	61.23%	97.19%
24	烧伤科	60.12%	97.73%
25	外科	57.00%	96.80%
26	传染科	52.43%	95.95%

9 结束语

本文利用 AI 文本挖掘处理技术, 根据 200 万病患的病情自述以及国际标准病理科室知识库, 设计了一种能够理解病患自述语言, 并能提供高质量科室导诊的系统, 为患者准确就医提供帮助。利用了多头注意力与 TextCNN 算法, 有效地提升了文本挖掘和科室推荐的有效性。基于本文提出的算法和平台设计, 开发了面向移动端的“智能与导诊”平台, 已经在“浙江预约挂号”App 上运行, 取得了较好的服务效果, 大大提升了患者自助挂号的准确性, 提升了患者就医的效率, 使医院的医疗资源配置更加有效。

参考文献:

- [1] LOPEZ-NORES M, BLANCO-FERNÁNDEZ Y, PAZOS-ARIAS J J, et al. Property-based collaborative filtering for health-aware recommender systems[C]//2011 IEEE International Conference on Consumer Electron, Jan 9-11, 2011, Las Vegas, USA. Piscataway: IEEE Press, 2011.
- [2] PATTARINTAKORN P, ZAVERUCHA G M, CER CONE N. Web based health recommender system using rough sets, survival analysis and rule-based expert systems [C]//11th International Conference on Rough Sets, Fuzzy Sets, Data Mining and Granular Computing, May 14-16, 2007, Toronto, Canada. Heidelberg: Springer, 2007: 491-499.
- [3] 徐守坤, 吴伟伟. 基于语义的医疗资源均衡推荐算法[J]. 计算机工程, 2015, 41(9): 74-79.

XU S K, WU W W. Balance recommendation algorithm for medical resources based on semantic[J]. Computer Engineering, 2015, 41(9): 74-79.

- [4] 王志宏, 杨震. 人工智能技术研究及未来智能化信息服务体系的思考[J]. 电信科学, 2017, 33(5): 1-11.

WANG Z H, YANG Z. Research on artificial intelligence technology and the future intelligent information service architecture[J]. Telecommunications Science, 2017, 33(5): 1-11.

[作者简介]



汤人杰 (1983-), 男, 中国移动通信集团浙江有限公司高级工程师, 主要研究方向为大数据、人工智能。



江涛 (1977-), 男, 浙江省卫生计生信息中心高级工程师、副主任, 主要研究方向为医疗健康行业信息化。



杨巧节 (1984-), 女, 中国移动通信集团浙江有限公司工程师, 主要研究方向为人工智能。