文章编号：2095-2163(2022)02-0178-04 中图分类号：TP391 文献标志码：A

融合关键字的注意力机制的淋巴水肿病历诊断推理算法

王帅帅，徐 臻
（中国电子科技南湖研究院，浙江 嘉兴 314000）

摘 要： 医疗智能诊断推理模型一直是医疗互联网领域的研究重点，有效的诊断推理模型可以帮助医生提高诊断效率，然而疾病的诊断结论需要考虑各种复杂因素，各种类型的疾病应该有自己的诊断模型。本文基于淋巴水肿电子病历提出了一种融合关键字的注意力机制（key-Attention）疾病诊断推理算法，使用电子病历中主诉、家族史、体格检查等内容推理出初步诊断结果。该算法使用的词频-逆文本频率（TF-IDF）算法提取病历关键词，注意力机制选用改进的指针生成神经网络。实验结果表明，该算法能够有效的解决推理模型缺少关键词问题，可以准确地根据电子病历作出疾病诊断推理。

关键词： 注意力机制；诊断推理；电子病历

Inference algorithm for lymphedema medical record diagnosis with fused keyword attention mechanism

WANG Shuaishuai, XU Zhen
（China Nanhu Academy of Electronics And Information Technology, jiaxing Zhejiang 314000, China）

【Abstract】 Medical intelligent diagnosis reasoning model has always been the research focus in the field of intelligent medical. An effective diagnosis reasoning model can help doctors improve diagnosis efficiency. However, the diagnosis of diseases needs to consider various complex factors, and various diseases should have their own diagnosis. Based on the electronic medical record of lymphedema, this research proposes a keyword fusion attention mechanism disease diagnosis and reasoning algorithm. In this algorithm, the tf-idf algorithm is used to extract keywords in the medical records, and the attention mechanism uses an improved pointer to generate a neural network. Experimental results show that the algorithm can effectively solve the problem of lack of keywords in the previous reasoning model, and can accurately make disease diagnosis reasoning based on electronic medical records.

【Key words】 attention mechanism; diagnostic reasoning; electronic medical record

0 引 言

淋巴水肿是以淋巴管堵塞引起肢体肿胀为代表的一种疾病。根据世界卫生组织统计，淋巴水肿在常见慢性病中列第 11 位，致残类疾病中列第 2 位，全球淋巴水肿患者约达 1.7 亿，中国淋巴水肿患者也高达千万人。淋巴水肿是世界医学难题，目前尚不可治愈，如果早期发现，诊断治疗及时得当，可以不同程度得以缓解[1]。目前国内系统有效诊断治疗淋巴水肿的医疗机构还很少，相关专业医生缺口巨大，淋巴水肿的相关知识尚不普及，大多数患者在发病后得不到有效的诊断和治疗，导致病情不断恶化，因此构建一个淋巴水肿疾病的智能诊断模型具有重要意义。本文利用深度学习技术，使数字赋能病理诊断，通过训练和学习医院收集的淋巴水肿电子病历，快速实现对电子病历内容的的识别与理解，从而大大提升病理诊断的效率和准确率，辅助专业医师，服务更多的患者。

1 诊断推理模型

1.1 电子病历关键词的提取

关键词是文档中能够表达重要内容的词语，关键词提取在信息检索、自动摘要、文本聚类等方面有重要应用[2]。本文认为电子病历中一些关键词语和检查结果对病历诊断结果有重要作用，尤其淋巴水肿相关疾病，不仅要识别出淋巴水肿类型，还需要识别出身体患淋巴水肿的部位。提取病历中关键症状、部位、疾病等关键词，可以更好地帮助模型理解病历的内容。关键词抽取常用的算法有词频-逆文本频率（TF-IDF）、文本排序（TEXTRANK）算法和主题模型算法。本文关键词提取使用 TF-IDF 算法。TF-IDF 的含义是如果某个词或短语在一篇文章中出现的频率（*TF*）高，并且在其他文章中很少出现，则认为此词或者短语具有很好的类别区分能力，适合用来分类[3]。词频（*TF*）是指某一个给定的词语在该文中出现的频率，式(1)。

作者简介：王帅帅(1991-)，男，硕士，工程师，主要研究方向：nlp、知识图谱。

收稿日期：2021-10-22

$$TF_{i,j} = \frac{n_{i,j}}{\sum_k n_{i,j}} \tag{1}$$

其中，$n_{i,j}$ 表示该词在文件中的出现次数，分母为文件中所有词的出现次数之和。

TF-IDF 假设高频率词应该具有高权重，除非其在所有的文档中出现的频率都很高。而逆文档频率（*IDF*）的大小与一个词的常见程度成反比，即最常见的词赋予最小的权重，较常见的词赋予较小的权重，而较小频率的词赋予较大的权重，式(2)。

$$IDF_{i,j} = \log\left(\frac{|D|}{1 + |\{j: t_j \in d_j\}|}\right) \tag{2}$$

其中，|D| 表示语料库中的文档总数，$|\{j: t_j \in d_j\}|$ 表示包含词语 t_i 的文档数目。

在分别计算得到 *TF* 和 *IDF* 后，将其相乘就能得到 TF-IDF 的值，如式(3)所示。

$$\text{TF} - \text{IDF}_{i,j} = TF_{i,j} * IDF_{i,j} \tag{3}$$

计算得到文本的关键词特征向量和 TF-IDF 特征向量后，将其拼接为一维的长向量，该向量就是最终机器学习算法要学习的特征向量。本文使用的淋巴水肿电子病历数据，如图 1 所示。对其中一个电子病历主诉、现病史、个人史、家族史和体格检查内容使用 TI-IDF 提取关键词，按重要性排序如图 2 所示。

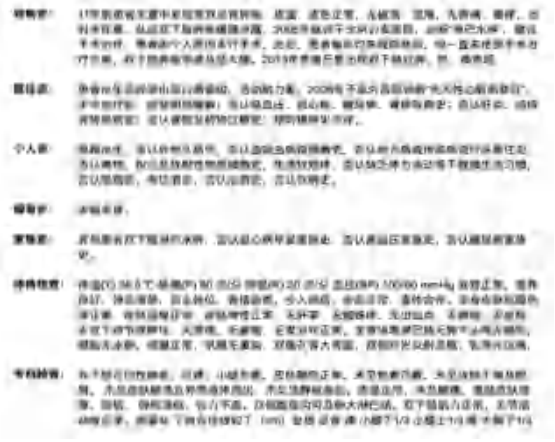

图 1　淋巴水肿电子病历

Fig. 1　Lymphedema electronic medical record

图 2　关键词排序

Fig. 2　Keyword ranking

1.2　融合关键词的注意力诊断推理模型

在自然语言领域序列到序列（seq2seq）模型和注意力机制（attention）为生成式推理提供了一种可行方法[4]。但这些模型存在两个问题：

（1）不能准确把握文章细节，无法处理未登录单词问题；

（2）倾向于重复自己的内容，使生成的句子不连贯。

指针生成网络（PGN）通过指向从源文本中复制单词，有助于准确地复制信息，同时保留通过生成器产生新单词的能力，使用覆盖（coverage）机制来跟踪已总结的内容，防止重复[5]。PGN 是在 seq2seq 模型的基础上构建，PGN 模型架构如图 3 所示。

该模型在 seq2seq+attention 模型的基础上增加了 P_{gen}，在每个解码器过程中，计算一个生成概率 $P_{gen}[0,1]$，该值决定有多大的概率从单词表中生成单词，模型中的最终分布根据词汇分布和注意分布加权求和得到，根据最终分布进行预测。PGN 既允许通过指针复制单词，也允许根据词汇生成单词。在指针生成器模型中，时间步长 t 的生成概率 P_{gen} 是根据上下文向量 $\boldsymbol{h}_t$、解码器状态 $\boldsymbol{S}_t$ 和解码器输入 $\boldsymbol{x}_t$ 计算[6]，如式(4) ~ 式(6) 所示。本文的编码端和解码端使用的是长短时记忆网络（LSTM）。编码端的输入为 $\boldsymbol{x} = (\boldsymbol{x}_1, \cdots, \boldsymbol{x}_n)$，解码端的输入为 $y = (y_1, \cdots, y_n)$。

$$\boldsymbol{h}_t = LSTM_{enc}(\boldsymbol{E}(\boldsymbol{x}_t), \boldsymbol{h}_{t-1}) \tag{4}$$

$$s_t = LSTM_{dec}(\boldsymbol{E}(\boldsymbol{y}_{t-1}), \boldsymbol{s}_{t-1}) \tag{5}$$

$$p_{gen} = \sigma(\boldsymbol{w}_h^{\mathrm{T}}\boldsymbol{h}_t + \boldsymbol{w}_s^{\mathrm{T}}s_t + \boldsymbol{w}_x^{\mathrm{T}}x_t + \boldsymbol{b}_{ptr}) \tag{6}$$

其中，$\boldsymbol{E}(x_t)$ 为单词的词向量；向量 $\boldsymbol{w}_h$、$\boldsymbol{w}_s$、$\boldsymbol{w}_x$、$\boldsymbol{b}_{ptr}$ 是学习参数；σ 是激活函数 *sigmoid*。

P_{gen} 用来决定从词汇表生成单词，还是从源文本复制单词的概率，用来对词汇分布和注意力分布进行加权平均，得到扩展词汇表上的概率分布，如式(7)所示。

$$p(\boldsymbol{w}) = p_{gen}p_{vocab}(\boldsymbol{w}) + (1 - p_{gen})\sum_{i:\boldsymbol{w}_i = \boldsymbol{w}} a_i^t \tag{7}$$

其中，p_{vocab} 为单词表分布，a_i^t 表示原文档中的词。当一个词不出现在常规的单词表中时，$p_{vocab}(\boldsymbol{w})$ 为 0。

使用 PGN 进行病历诊断结果生成的结果往往会忽略一些重要的词，比如部位等，而且现有的深度学习生成方法只关注结果与原始文本的总体关系，有时可能会对文本中的主要细节内容把握不准确，导致生成的结果不全面，很容易丢失病历中部位、症状等关键信息。将病例中关键词作为 PGN 模型生

成结果的提示,病历的关键词可以是部位、疾病之类名词,也可以是症状类的描述性短语,让模型在解码时更加关注这些关键词汇,从而使生成的结果更加准确。如图 4 所示,将病历关键词通过注意力机制融入到 PGN 模型中,模型就可以通过关键词所包含的语义和病历中其他信息生成概括性的诊断结论。

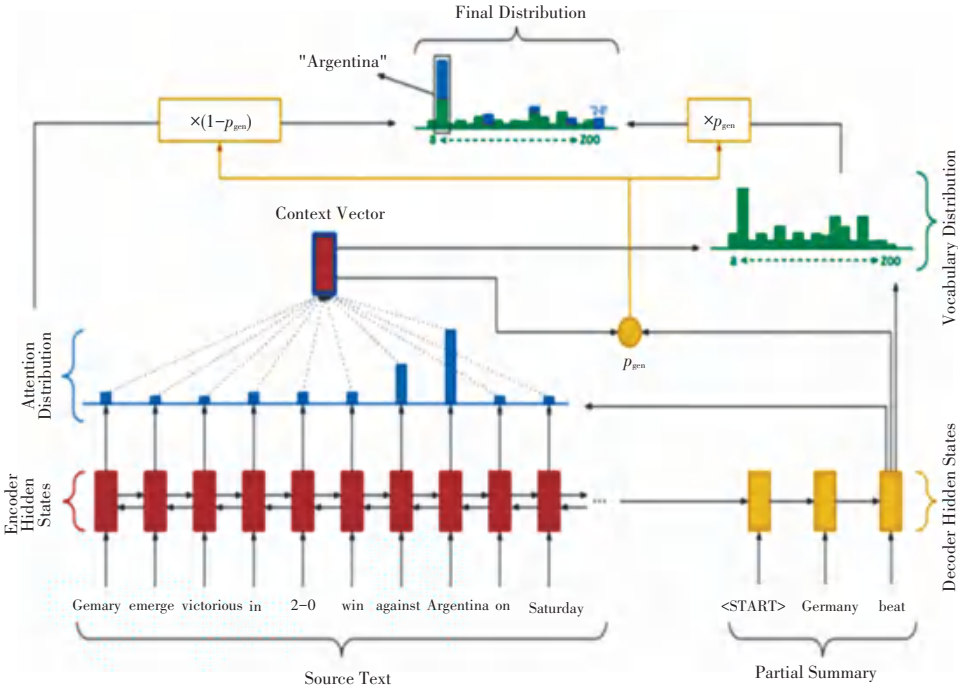

图 3 PGN 模型

Fig. 3 PGN model

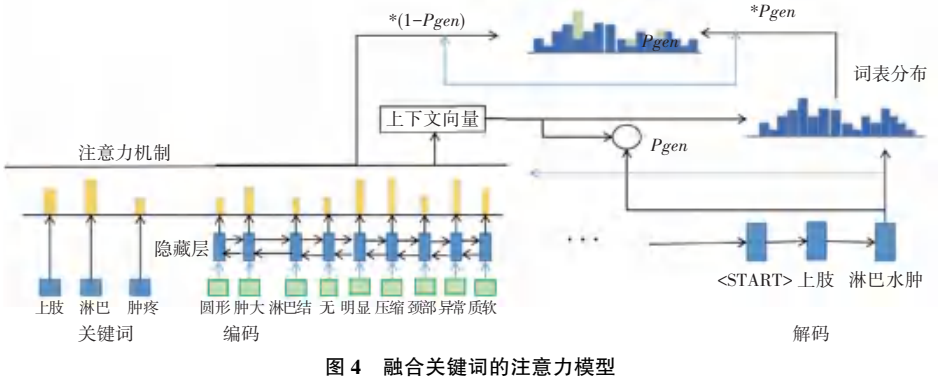

图 4 融合关键词的注意力模型

Fig. 4 Attention model fused with keywords

病历输入时使用 word2vec 训练生成词向量,对于本文选取到的关键词,使用 word2vec 生成词向量 $\boldsymbol{k}=\{\boldsymbol{k}_1,\boldsymbol{k}_2,\cdots,\boldsymbol{k}_d\}$,将所有的 $\boldsymbol{k}$ 相加作为输入融合注意力机制。计算方法如式(8) 和式(9) 所示。

$$q=\sum_{i=1}^{s}\boldsymbol{k}_i \tag{8}$$

$$p_{gen}=\sigma(\boldsymbol{w}_h^{\mathrm{T}}\boldsymbol{h}_t+\boldsymbol{w}_s^{\mathrm{T}}\boldsymbol{s}_t+\boldsymbol{w}_x^{\mathrm{T}}\boldsymbol{x}_t+\boldsymbol{w}_q^{\mathrm{T}}q_t+\boldsymbol{b}_{ptr}) \tag{9}$$

2 数据与验证

2.1 实验设置与数据

本文使用 8 000 份电子病例作为实验的训练集,使用 500 份电子病历作为测试集,300 份淋巴水肿病历作为验证数据集。实验在 GPU 服务器上进行,采用 pytorch 深度学习框架。本文使用的词汇表大小为 8 000 词,单词向量的维度是 128,编码器和解码器的输入维度是 256,batch_size 大小为 64。

2.2 评价方法

本实验使用的评价指标为 rouge(recall-oriented understudy forgisting evaluation),是文章摘要提取和机器翻译常用的评价指标。ROUGE 主要有 ROUGE-N、ROUGE-L 和 ROUGE-W 3 种方法,本文使用的是 ROUGE-L 方法, l 指最长公共子序列,使用了机器译文 C 和参考译文 S 的最长公共子序列,式(10) ~式(12)。

$$R_{LCS} = \frac{LCS(C,S)}{len(S)} \tag{10}$$

$$P_{LCS} = \frac{LCS(C,S)}{len(C)} \tag{11}$$

$$F_{LCS} = \frac{(1+\beta^2)R_{LCS}P_{LCS}}{R_{LCS}+\beta^2 P_{LCS}} \tag{12}$$

其中，LCS 表示文本公共长度；R_{LCS} 表示召回率；P_{LCS} 表示精确率；F_{LCS} 就是 ROUGE-L。

3　结果与分析

本文使用每个病历数据集经处理后长度为 900 字左右，一个病历诊断结果生成示例如图 5 所示。

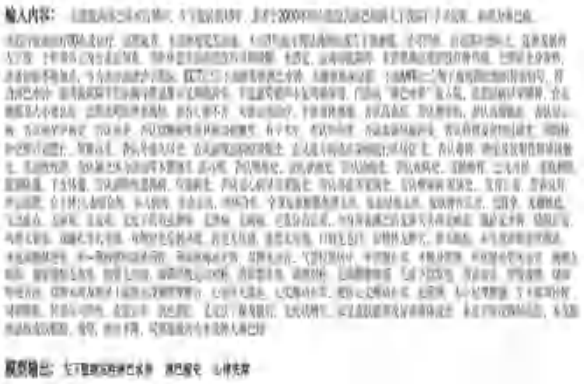

图 5　诊断结果生成

Fig. 5　Diagnosis result generation

为了验证算法的有效性，本文对比了序列到序列，注意力机制和指针生成网络模型，评价指标使用 ROUGE-1、ROUGE-2，ROUGE-L，实验结果见表 1。实验表明，本文提出的融合关键字的注意力机制疾病诊断推理模型识别准确率优于其他的算法。

表 1　模型结果对比

Tab. 1　Comparison of model results

模型	ROUGE-1	ROUGE-2	ROUGE-L
序列到序列	38.45	31.46	34.85
注意力机制	39.51	35.22	36.4
指针生成网络	39.89	36.38	36.78
融合关键字注意力机制	41.23	40.76	38.9

由表 1 可以看出，在将病历中关键信息加入到模型后，本文的融合关键字的注意力机制明显优于指针生成网络，在生成的诊断结论中，融合关键字的注意力机制可以有效的提取到病历中关键信息，ROUGE-2 指标提升最多，能够得到性能更好地淋巴水肿诊断推理模型，为淋巴水肿相关疾病诊断提供了可靠的辅助支持。

4　结束语

本文提出的融合关键词的注意力机制模型即保持了模型的文本生成能力，又可以让模型可以向医生一样依据病历中的核心症状和核心部位等信息进行疾病的推理，生成的诊断结果更加连贯，更能覆盖病历信息。利用深度学习技术构建水肿诊断推理模型可以帮助医生进行疾病的快速诊断，让患者可以及时得到治疗，在一定程度上缓解医疗资源不足问题。

参考文献

[1] 周文红，张玄，井月秋，等. 乳腺癌术后上肢淋巴水肿患者生活质量的调查[J]. 中华护理杂志，2008，43(7)：661-664.

[2] 胡少虎，张颖怡，章成志. 关键词提取研究综述[J]. 数据分析与知识发现，2021，5(3)：45-59.

[3] 张鑫明，白冬立. 一种基于优化 TFIDF 的特征提取方法及系统[P]. 北京市：CN111062212B，2020-06-30.

[4] 董谱. 改进的 Seq2Seq 文本摘要生成方法[D]. 广州：广东工业大学，2021.

[5] 巩轶凡，刘红岩，何军，等. 带有覆盖率机制的文本摘要模型研究[J]. 计算机科学与探索，2019，13(2)：29-37.

[6] SEE A，LIU P J，MANNING C D. Get To The Point：Summarization with Pointer-Generator Networks [C]// Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1：Long Papers). 2017.