

# 面向消化内科辅助诊疗的生成式对话系统<sup>①</sup>



程梦卓, 董兰芳

(中国科学技术大学 计算机科学与技术学院, 合肥 230022)

通讯作者: 董兰芳, E-mail: [lfdong@ustc.edu.cn](mailto:lfdong@ustc.edu.cn)

**摘要:** 社会的高速发展带给群众的压力越来越大, 由于工作压力和自身问题, 越来越多的人饮食不规律且不健康, 导致患有消化系统疾病的人群日益扩大, 而在身体刚出现异样时, 大部分人会首先选择在网上寻找信息帮助, 由于传统搜索引擎的局限性, 过程耗时, 且因为消化内科疾病的多样性, 用户很难准确获取相关信息. 针对这一问题, 考虑到对话系统作为一种较为高级的信息检索系统, 能够根据用户的输入及时返回相关有效信息, 本文探索了一种适用于消化科领域的生成式对话系统, 通过支持向量机分类与主动学习相结合, 在多个医学网站获取消化内科的专业问诊对话语料, 人工与统计相结合构建消化内科疾病、药品和症状专业词典, 改善传统分词工具在消化内科领域分词效果, 在提高分词效果的基础上, 使用 Encoder 与 Decoder 多层结构和门控循环单元 GRU 结合的方式, 加入注意力机制, 提出结合颠倒输入、键值对向量和 Word2Vec 向量的模型加强训练法, 从而获得最终的消化内科生成式问答系统. 实验结果表明, 分词的准确率比传统方式高, 且得出的对话模型能够有效的生成与问句相关的答句, 提高对话系统的回答准确率.

**关键词:** 消化内科; 分词; 主动学习; 词向量; seq2seq; GRU; 注意力机制

引用格式: 程梦卓, 董兰芳. 面向消化内科辅助诊疗的生成式对话系统. 计算机系统应用, 2019, 28(10): 53-60. <http://www.c-s-a.org.cn/1003-3254/7128.html>

## Generating Auxiliary Diagnosis Dialogue System for Gastroenterology

CHENG Meng-Zhuo, DONG Lan-Fang

(School of Computer Science and Technology, University of Science and Technology of China, HeFei 230022, China)

**Abstract:** The rapid development of society brings more and more pressure to people. Due to work pressure and self-problems, more and more people are eating three meals irregularly and unhealthily, which leads to the growing population of people suffering from digestive diseases. When the body just appears abnormal, most people first will choose to find information on the Internet. Due to the limitations of traditional search engines, the process is time consuming, and because of the diversity of diseases, it is difficult for users to accurately obtain relevant information. In view of this problem, considering that the dialogue system is a more advanced information retrieval system. This study explores a generative dialogue system suitable for the field of gastroenterology, using support vector machine and active learning to obtain professional consultation dialogue corpus of gastroenterology on multiple medical websites. Labor and statistics are combined to build a professional dictionary of diseases, drugs, and symptoms of digestive diseases, word segmentation is improved in the medical field. On this basis, multi-Encoder and multi-Decoder structure is combined with gated loop unit GRU, the attention mechanism is also added, then the model strengthening training method is proposed, which combines reverse input and key-value pair vector and Word2Vec vector, to obtain the final model. The experimental results show that the word segmentation result is much higher than the traditional method, and the resulting

<sup>①</sup> 收稿时间: 2019-03-27; 修改时间: 2019-04-26; 采用时间: 2019-04-29; csa 在线出版时间: 2019-10-15

dialogue model can effectively generate the sentence related to the question, which can improve the answer accuracy of the dialogue system.

**Key words:** gastroenterology; participle; active learning; word vector; seq2seq; GRU; attention mechanism

社会的高速发展,带来群众各方面压力的日益增加,工作强度越来越大,饮食不规律,不健康,欠缺睡眠,加上现在食品的安全性依然存在着很大的隐患,导致身患消化内科疾病的人群日益增多.有调查显示接近80%的人在有医药和疾病的疑问时,首先会选择上网寻找帮助,但是传统的搜索引擎在提供相关信息时,大多采用关键词匹配技术,在过去网络中已录入的信息进行匹配,具备诸多限制,类似于“过去几周”的关键词会被忽略,并且由于用户的语言表达差异和网络信息的混乱,网络存在的知识库并不能涵盖用户重要的意图,无法满足实际的应用需求.

消化内科对话系统的研究需要大量数据,而现有的医学公开数据集多为影像数据集,没有问诊类的公开数据集.其次,同医学自然语言理解处理研究的大部分挖掘研究类似,第一步都需要对数据进行分词,但是因为消化内科疾病的种类繁多,还有很多新的疾病不断被发现,所涉及疾病名、症状表现名和药名很多,中科院的汉语分词系统NLPIR、中文结巴分词等常用的分词工具在处理该领域分词的结果并不能给生成式对话系统构建提供保障.国内中文医学术语的标准化的研究也比较少,而国外如比较具备代表性如UMLS<sup>[1]</sup>等也缺乏对中文的支持.

本文在谷歌传统seq2seq框架的基础上,运用butterfly爬虫技术和主动学习结合获取相关网站消化内科约48万条问诊语料,解决医疗问诊语料缺少问题.在医学院消化科同学的帮助下,结合医药库的消化科常用药名和常见病症,通过统计分析构建约20万消化科领域词汇表,运用基于词库与最大规则相结合的分词算法,对问诊语料对进行分词处理.结合Word2Vec构建自主的词向量,并提出模型增强训练法获得消化科生成式问诊对话模型.

## 1 处理流程

消化内科生成式问诊系统的实现过程主要涉及3个模块,语料获取与处理、词向量构建、对话模型,整体的框架如图1所示.

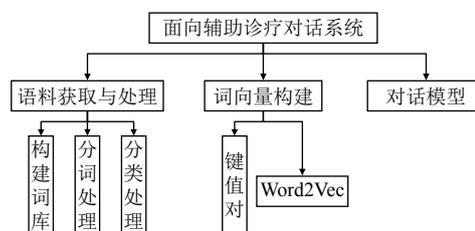


图1 系统模块图

### 1.1 语料获取与预处理

#### 1.1.1 构建词库

因为消化科文本的特殊性,包括方位词、副词以及大量的专业词汇和否定词汇,直接利用结巴分词类通用分词工具进行分词,结果显示,这些词当中大部分无法被识别,分词效果极差.但对话模型的训练过程涉及词之间相关性的建立过程,且键值对向量的处理对分词的准确率有一定的要求,为了提高分词的结果准确率,本文构建自定义消化内科分词词库,其主要由两大部分组成:消化专业词库和停用词词库.

消化内科专业词库主要来源于以下几部分:从现存医学网站爬取获得的消化科常见疾病和症状表现名称,包括有问必答网、寻医问药网;百度文库提供的消化科药品字典;以及医学院同学对字典的补充.词库内部结构如“胰腺炎 nz 260”,其中“胰腺炎”为消化科专业词汇,nz代表其他专有名词,260代表词频.

停用词库包含在问诊语料之中具有高频率但实际意义不大的词,如:“的”、“请”,包含介词、副词、语气词等.

#### 1.1.2 数据预处理

在问答对进行分词操作之前,都会先进行文本的预处理.第一,根据自定义的停用词库,去除文本中“的”、“①”等停用词;第二,由于网络获取的问答对经常会出现多个标点的情况,如“????”,去掉重复;第三,很多药名有简称或者别名,本文为了降低复杂性,将有别名的药统一为同一药名,如“奥美拉唑肠溶片”统一为“兰尼”.

#### 1.1.3 分词

现存的分词大致有3大类:基于理解的分词方

法、基于规则的分词方法和基于统计的分词方法<sup>[2]</sup>。由于本文词典的构建花费时间比较多,相对比较完整,因此在基于最大逆向匹配算法的结巴分词算法的基础上,采用改进规则和统计相结合的方式行进行分词的改进。

文献[3]提出歧义检测,在对句子进行正向和逆向切分得到的两个结果的比较过程中,发现两种方式切分的结果有90%的概率重合且正确,有约9%的概率其中必有一个结果是正确的。因此,本文在此基础上,在进行问答对分词的时候增加比较机制,同时对句子进行正向和逆向最大匹配切分,然后将两个结果进行比较,若前后分词所产生词的个数有差异,则选择其中单字少的分词作为结果,若所得分词数相同,则随机选择一个作为结果。

#### 1.1.4 分类

消化内科一般分为“胃肠病”、“肝病”、“胰胆疾病”、“内镜”和其他疾病五大类,对话模型的训练过程中,为了避免过拟合,需要五大类的语料数据均衡。由于每条问答对可以具备多种疾病的症状表现,且不受其他类影响,并且一个类的疾病具备常见的术语,如“胃肠病”常出现“胃炎”类关键词,因此本文将多标签分类问题转化成单标签分类问题进行求解。

分类模型的训练的过程结合主动学习<sup>[4]</sup>,先准备两万条问答对,人工先对其中6000条数据进行分类处理,完成类别标注,然后将这些数据用Word2Vec向量表示,作为支持向量机的输入获得简单的分类模型,然后每次由分类模型处理2000条新数据,根据结果评测分类器的好坏,并且每次将标记好的数据加入训练集,重新获得新的分类器,直至分类模型达到给定的阈值,过程如图2所示。

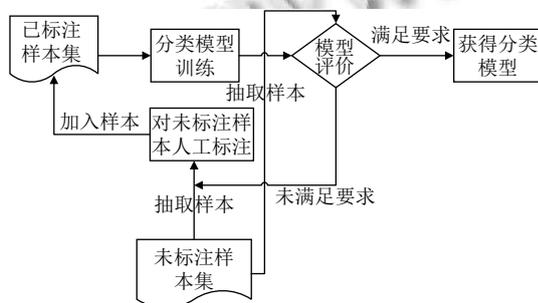


图2 获取语料流程

得到满足的要求模型后,通过该模型本文从有问必答网、寻医问药网共获取约48万条消化科问答对。

## 1.2 词向量构建

词向量又称 Word 嵌入式自然语言理解中的一组语言建模和特征学习技术的统称,是让计算机理解自然语言手段之一,很多研究都表明当将其用作底层输入表示时,可以很大程度的提高 NLP 任务的性能,如语法分析和情感分析。

为了在有限的数据集下能获得较好的词与词之间的相关性,本文使用两种词向量:键值对向量和 Word2Vec 向量。

### 1.2.1 键值对向量

分词之后获得初始语料,对分词之后的语料中词语的出现频次进行统计,从高到底进行排序,形成形同“123-胆囊炎”的键值对,通过实验比较排序字典大小的影响,得到相对于合理的是统计前20000个高频词,也就是在问答对里各抽出前20000个高频词,形成键值对序号,高频词的选择之后会经过比对,如果自定义词典有词未出现在这20000个词内,问答对同时增加未出现的词,因此序号的取值范围会多个“+”。

### 1.2.2 Word2Vec 向量

Word2Vec<sup>[5,6]</sup>包括两种模型:CBOW 和 Skip-Gram。CBOW 模型的训练输入是某一个特征词的上下文相关的词对应的词向量,输出的是这特定的一个词的向量。Skip-Gram 与 CBOW 刚好相反,是在已知当前词的情况下预测上下文。两者均可以和哈夫曼树结合训练得到最终的词向量,但考虑到时间序列 GRU<sup>[7]</sup>,是根据句子前面的词预测后面的词,课题选择基于 Hierarchical Softmax 哈夫曼树的 Skip-Gram 模型获取本文的 Word2Vec 向量。

## 1.3 对话模型

生成对话这个场景,与机器翻译有很多相似之处,都可以简单理解为建立原句和翻译结果两者相同位置的词的相关性,都是根据当前词推算下一个词,现在的研究部分是在不改变主模型的基础上,在传统的 Sutskever 等<sup>[8]</sup>提出 seq2seq 框架上改变编解码结构、神经元等,如加入注意力机制<sup>[9]</sup>,有的已达到比较好的效果。由于对话场景的特殊性,一个极有潜力的改变办法是长短期记忆网络 LSTM<sup>[10]</sup>(包含其各种变体)结合 seq2seq 框架,加入注意力机制,理由在于 LSTM 能够避免长期的依赖问题,适合于解决包含时序先后顺序的序列生成的问题,如阿里小蜜<sup>[11]</sup>和百度自我诊断<sup>[12]</sup>。当本文直接使用基本模型 seq2seq 模型时,所得

结果很不好,生成词的困惑度高,整个句子的结构不全,可读性很差.为了得到更好的结果,本文对模型进行筛选,在公开数据集 WSJ 第 23 部分同等前提下,加州伯克利分校 NLP 实验室开发的 BerkeleyParser 开源句法分析器在测试集上 F1 的分数达到了 90.5, Google 提出的句法成分分析所采取的结构<sup>[13]</sup>F1 分数达到了 95.7. 而 GRU 与 LSTM 相比,只有两个门(更新门和重置门),参数少更容易收敛,本文的实验对比也表明 GRU 比 LSTM 更适用于对话处理场景.因此本文最终选择 multi\_encoder+attention\_decoder+GRU+beamsearch 模型结构,网络结构如图 3 所示.

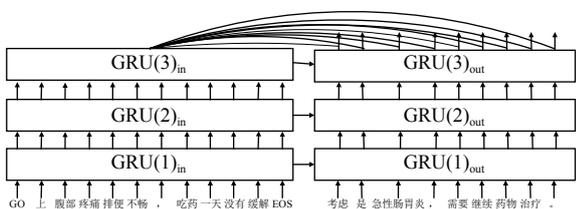


图 3 对话模型网络结构图

给定输入为 $(x_1, \dots, x_T)$ , 文献[14]表明采取倒序输入能够增加输入词之间的相关性, 因此本文也采取倒序原句作为输入.

用 $x_t, h_t, \tilde{h}_t$ 分别代表 $t$ 时刻的输入、输出状态、隐藏状态, GRU 结构如图 4 所示.

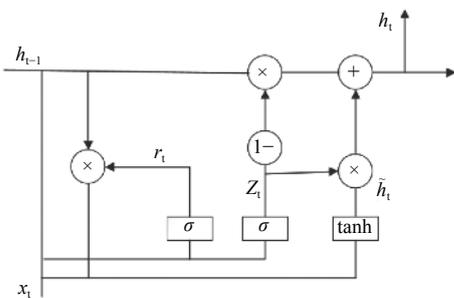


图 4 GRU 结构图

第一层得到中间状态 $(h_1, h_2, \dots, h_t)$ 和 $(\tilde{h}_1, \tilde{h}_2, \dots, \tilde{h}_t)$ 的计算如下:

$$z_t = \sigma(W_z \cdot [h_{t-1}, x_t]) \quad (1)$$

$$r_t = \sigma(W_r \cdot [h_{t-1}, x_t]) \quad (2)$$

$$\tilde{h}_t = \tanh(W_h \cdot [r_t * h_{t-1}, x_t]) \quad (3)$$

$$h_t = (1 - z_t) * h_{t-1} + z_t * \tilde{h}_t \quad (4)$$

其中,  $[\ ]$ 表示两个向量相连接,  $*$ 表示矩阵的相乘,  $\cdot$ 表示

矩阵的点乘,  $W_z, W_r, W_h$ 是模型需要学习的模型参数.  $r_t$ 重置门决定是否将之前的状态忘记, 当其值趋近 0 的时候, 前一个时候的状态信息 $h_{t-1}$ 会被忘掉, 隐藏状态 $\tilde{h}_t$ 会被重置成当前输入信息, 更新门 $z_t$ 决定是否将隐藏状态更新为新的状态 $h_t$ .

在一个深层的 GRU, 每层将上一层的得到的 $h_t$ 作为该层的输入序列  $X$ , 定义输出的分布为:

$$P(B|A) \equiv \prod_{t=1}^{T_B} P(B_t|A_1, \dots, A_{T_A}, B_1, \dots, B_{t-1}) \quad (5)$$

$$\equiv \prod_{t=1}^{T_B} \text{softmax}(W_o \cdot h_{T_A+t})^T \delta_{B_t}$$

输入序列 $X = (A_1, \dots, A_{T_A}, B_1, \dots, B_{T_B}), (A_1, \dots, A_{T_A})$ 代表上一层的输出 $h_t, (B_1, \dots, B_{t-1})$ 代表 $t-1$ 时刻前面 $t-1$ 神经元的输出,  $W_o$ 为权重参数,  $\delta_{B_t}$ 为克罗内克函数.

直接使用该方法来处理医疗问答会出现一些问题, 其一在于问句对疾病的描述语句可能会很长, 解码阶段 GRU 无法很好的针对序列前面部分进行解码, 其二在于没有关键点, 解码阶段应该更关注于疾病症状词, 而直接单纯的使用同一个中间向量进行解码, 显然是不合理. 因此, 本文加入注意力机制, 来源于文献[10], 也就是图 3 中最上方的黑线曲线, 使得每一步解码都有不同的中间向量  $c$ . 中间状态表示为 $(h_1, \dots, h_{T_A})$ , Decoder 中间状态用 $(d_1, \dots, d_{T_B}) := (h_{T_A+1}, \dots, h_{T_A+T_B})$ 定义, 计算过程:

$$u_i^t = v^T \tanh(W'_1 h_i + W'_2 d_t) \quad (6)$$

$$a_i^t = \text{softmax}(u_i^t) \quad (7)$$

$$d_t^t = \sum_{i=1}^{T_A} a_i^t h_i \quad (8)$$

其中,  $t$ 指的是 $t$ 时刻,  $i \in [1, T_A]$ ,  $v$ 和矩阵 $W'_1, W'_2$ 是模型需要学习的参数, 由于前面说过编解码都使用相同规格 GRU, 因此 $W'_1, W'_2$ 的维度一样.  $u_i^t$ 的长度与 Encoder 产生的 $T_A$ 具备相同长度, 其中 $i$ 的值代表关注 Encoder 中 $h_i$ 的程度, 使用 Softmax 进行规范化, 最后通过将 $d_t^t, d_t$ 拼接得到新的中间状态, 作为解码的中间向量  $c$ .

## 2 实验分析

### 2.1 分词实验结果

从获取的 48 万对问答对中随机抽取 2000 条问答对, 进行人工分词并统计作为标准. 结果评估采用第二届国际汉语分词评测发布的国际中文分词标准进行评

测,计算方法:

$$precision = \frac{CN}{CS} * 100\% \quad (9)$$

$$recall = \frac{CN}{TS} * 100\% \quad (10)$$

$$F = \frac{2 * precision * recall}{precision + recall} * 100\% \quad (11)$$

其中, *precision* 表示准确率, *recall* 表示召回率, *F* 值为正确率和召回率的调和平均值, *CN* 表示正确切分词

数, *CS* 表示切分的总词数, *TS* 表示答案中的词语总数.

对这 2000 条问答对, 使用 3 种方法进行分词: 直接使用结巴分词、结巴结合清华医学词库、结巴分词+清华医学库+自定义词典, 以及结巴分词+清华医学库+自定义词典+歧义消除. 分词结果如表 1 所示, 分词样例如表 2 所示.

表 1 分词评估结果 (%)

| 评价指标      | 召回率  | 准确率  | <i>F</i> 值 |
|-----------|------|------|------------|
| 结巴分词      | 81.2 | 70.5 | 75.1       |
| 结巴+清华医学词库 | 83.4 | 73.8 | 78.2       |
| 结巴三加      | 98.6 | 97.3 | 97.9       |

表 2 分词样例对比

| 分词结果                     |  |
|--------------------------|--|
| 原句                       | 由于胃粘膜炎症引起的症状, 通常会有上腹部隐痛, 暖气等症状. 建议日常注意清淡饮食, 避免吃生冷辛辣刺激油腻的食物, 禁忌咖啡浓茶等, 同时口服胃炎胶囊, 999胃泰颗粒, 丽珠得乐等药物, 适当口服吗丁啉, 再热敷一下腹部, 症状应该会有所改善, 避免肠胃疾病恶化.  |
| 结巴                       | 由于/胃粘膜/炎症/引起/的/症状/, /通常/会/有/上/腹部/隐痛/, /暖气/等/症状/. /建议/日常/注意/清淡/饮食/, /避免/吃/生冷/辛辣/刺激/油腻/的/食物/, /禁忌/咖啡/浓茶/等/, /同时/口服/胃炎/胶囊/, /999/胃泰/颗粒/, /丽珠/得乐/等/药物/, /适当/口服/吗丁啉/, /再/热敷/一下/腹部/, /症状/应该/会/有所改善/, /避免/肠胃/病/恶化/. |
| 结巴+清华医学词库                | 由于/胃粘膜炎症/引起/的/症状/, /通常/会/有/上/腹部/隐痛/, /暖气/等/症状/. /建议/日常/注意/清淡/饮食/, /避免/吃/生冷/辛辣/刺激/油腻/的/食物/, /禁忌/咖啡/浓茶/等/, /同时/口服/胃炎/胶囊/, /999/胃泰/颗粒/, /丽珠/得乐/等/药物/, /适当/口服/吗丁啉/, /再/热敷/一下/腹部/, /症状/应该/会/有所改善/, /避免/肠胃/疾病/恶化/. |
| 结巴+清华医学词库+自定义词典          | 由于/胃粘膜炎症/引起/的/症状/, /通常/会/有/上腹部/隐痛/, /暖气/等/症状/. /建议/日常/注意/清淡/饮食/, /避免/吃/生冷/辛辣/刺激/油腻/的/食物/, /禁忌/咖啡/浓茶/等/, /同时/口服/胃炎胶囊/, /999胃泰颗粒/, /丽珠得乐/等/药物/, /适当/口服/吗丁啉/, /再/热敷/一下/腹部/, /症状/应该/会/有所改善/, /避免/肠胃/病/恶化/.       |
| 结巴+清华医学词库+自定义词典+前后选择歧义消除 | 由于/胃粘膜炎症/引起/的/症状/, /通常/会/有/上腹部/隐痛/, /暖气/等/症状/. /建议/日常/注意/清淡/饮食/, /避免/吃/生冷/辛辣/刺激/油腻/的/食物/, /禁忌/咖啡/浓茶/等/, /同时/口服/胃炎胶囊/, /999胃泰颗粒/, /丽珠得乐/等/药物/, /适当/口服/吗丁啉/, /再/热敷/一下/腹部/, /症状/应该/会/有所改善/, /避免/肠胃病恶化/.         |
| 人工                       | 由于/胃粘膜炎症/引起/的/症状/, /通常/会/有/上腹部/隐痛/, /暖气/等/症状/. /建议/日常/注意/清淡/饮食/, /避免/吃/生冷/辛辣/刺激/油腻/的/食物/, /禁忌/咖啡/浓茶/等/, /同时/口服/胃炎胶囊/, /999胃泰颗粒/, /丽珠得乐/等/药物/, /适当/口服/吗丁啉/, /再/热敷/一下/腹部/, /症状/应该/会/有所改善/, /避免/肠胃病恶化/.         |

通过表 1 结果不难看出, 当直接使用结巴分词进行消化内科语料的分词操作时, 准确率最低, 为 70.5%, 可以从表 2 中发现此时分词产生的结果中, “胃粘膜炎症”、“胃炎胶囊”、“999 胃泰颗粒”、“丽珠得乐”和“吗丁啉”药名和症状名出现错分, 且“肠胃病恶化”通过查询中间结果发现, 出现歧义, 有多种切分结果: “肠胃/病/恶化”、“肠胃病/恶化”、“肠/胃/病/恶化”. 增加清华医学词库后, 3 个指标都得到提升, 准确率达到 73.8%, 可以从样例中发现, “胃黏膜炎症”得到正确且切分, 同时也发现, 药名仍然处于错分状态, 且歧义没有得到解决. 进一步增加自定义词典, 由于本文的自定义词典花费的时间和人力较大, 包含绝大部分的消化内科常见疾病名称和药名, 因此通过表 1 可以发现, 实验结果得到

了很大提升, 准确率达到 96.6%, 召回率达到 96.4%, *F* 值达到 95.3%, 从样例中可以发现, “胃炎胶囊”、“999 胃泰颗粒”、“丽珠得乐”和“吗丁啉”都得到正确切分, 但歧义仍然存在.

最后一步, 歧义的消除, 可以发现准确率有了进一步的提升, 针对表 2 例子中“肠胃病恶化”, 根据本文采取取最大操作, 直接被切分成一个词, 符合人工切分要求.

## 2.2 对话模型实验

### 2.2.1 数据处理

获取的 48 万条问答对, 长短不一, 短的问句只有十几个词, 长的超过 60 个词, 本文采取桶装方式对数据进行分组, 共分 (10, 15)、(20, 30)、(30, 45)、(40,

60)、(40+, 60+) 5 个桶, (20, 30) 代表问句词数在 10~20 之间, 答句词数在 15~30 之间, 如果问句和答句的词数分布在两个桶内, 以答句词数为准, 如“12, 40”放至 (30, 45) 桶内, (40+, 60+) 代表多问答句词数超过 40 和 60 的放至该桶. 将五组分开进行模型的训练.

### 2.2.2 模型训练

本文提出一种加强训练法, 过程如下:

1) 使用键值对向量进行预训练, 得到初始模型 Model\_1;

2) 将键值对问句前后颠倒, 加载 Model\_1 进行训练, 得到 Model\_2;

3) 在 Model\_2 的基础上, 使用 Word2vec 向量进行训练, 得到 Model\_3;

4) 将 Word2vec 向量表示的问句前后颠倒, 加载 Model\_3, 得到最终模型 Model\_4.

### 2.2.3 评估方法

本文在 48 万问答对随机选择 1 万条问答对作为测试集, 通过两个方面进行对话模型质量评估: 困惑度和词向量匹配评价.

(1) 困惑度是衡量一个语言模型好坏的指标, 用来估算一句话是否通顺, 主要根据每个词来估计一句话出现的概率, 计算方法:

$$PP(S) = 2^{-\frac{1}{N} \sum \log(P(\omega_i))} \quad (12)$$

其中,  $S$  代表句子,  $N$  代表句子的长度,  $P(\omega_i)$  代表第  $i$  个词的概率, 困惑度越小, 期望句子出现的概率也就越大.

(2) 词向量的评价方法是通过计算目标句子与生成句子之间句向量之间的余弦距离的大小来估算两者的相似度, 本文采取 BLEU 值进行评估, 也就是比较两者中的  $n$ -gram (实验取值为 3, 也就是每次三词一组进行比对) 词组在整个训练语料中出现的总次数, 出现的次数越高, 则模型的效果越好. 计算方法:

$$P_n(r, \hat{r}) = \frac{\sum_k \min(h(k, r), h(k, r_i))}{\sum_k h(k, r_i)} \quad (13)$$

$$BP = \begin{cases} 1, & \text{if } c > r \\ e^{1-r/c}, & \text{else} \end{cases} \quad (14)$$

$$BLEU = BP \cdot \exp\left(\sum_{n=1}^N \omega_n \log P_n\right) \quad (15)$$

其中,  $P_n(r, \hat{r})$  计算  $n$ -gram 的短语词组在整个数据集中的准确度,  $h(k, r)$  表示每个  $n$ -gram 词组在参考答案语

句中出现的次数 (因为对于每个  $n$  而言, 都会存在多个  $n$ -gram 词组, 因此有一个求和);  $\omega$  表示各个  $n$ -gram 的权重; BP 是长度过短惩罚因子, 取值范围 (0, 1], 候选句子越短, 越接近 0, 加入该因子的目的在于改善生成答案过短的效果.

### 2.2.4 测试阶段

在测试解码阶段, 传统的广度优先策略虽然能够找到最优的路径, 但是由于本文数据有限, 困惑度会很大, 因此搜索的空间会非常大, 如果使用广度优先策略会导致内存占用指数级增长, 内存会溢出, 因此采用 Beamsearch 算法<sup>[15]</sup> 进行寻找最优解, 该算法是一种启发式图搜索算法, 一般当解空间特别大时会被使用到, 该算法的优点在于能够减少空间和时间的使用, 在每一步路径的选择中, 不同于传统的广度优先策略, 会根据定义的策略对一些质量差的点进行去除操作, 保留质量高点, 最终找到最优的路径. 算法的流程如下:

1) 初始化节点, 将节点插入到列表中;

2) 从列表中提取节点出堆, 如果为目标节点, 则结束, 否则扩展该节点, 并取集束的宽度的节点入堆, 既取前  $k$  个概率最大的节点入堆, 不断循环, 直至找到最优解或堆为空. 每个  $t$  时刻有几个大概率的词候选, 选择可能性最大的前几个候选, 降低复杂度.

### 2.2.5 对话模型实验结果

表 3 表明, 在 Encoder 和 Decoder 层数都为 3 层, 单元数都为 256 的条件下, GRU 比 LSTM 所产生的困惑度降低约 18, 得出 GRU 更适用于对话系统模型; 表格也显示在使用 GRU 作为神经元的前提下, 神经元数目超过 256 个, 模型会出现不收敛和困惑度持续增加的情况; 而在神经元为 256 个相同数目下, 3 层因为计算量少, 比 5 层的效果好, 困惑度降低 21, 因此最终本文选择 3 层 256 个神经元进行最终的模型的实验对比, 结果如表 4 所示.

表 3 模型层数选择对比

| Cell model | layers | units | First Last    |               | trend | steps     | preplexity |
|------------|--------|-------|---------------|---------------|-------|-----------|------------|
|            |        |       | learning rate | learning rate |       |           |            |
| GRU        | 3      | 256   | 0.5           | 0             | down  | 2000 000  | 3.55       |
| GRU        | 3      | 512   | 0.5           | 0.0528        | down  | 140 000   | 21         |
| GRU        | 3      | 512   | 1             | /             | up    | /         | inf        |
| GRU        | 2      | 256   | 0.5           | 0.0216        | down  | 140 000   | 16         |
| GRU        | 5      | 256   | 0.5           | 0.0654        | down  | 140 000   | 26.17      |
| LSTM       | 3      | 256   | 0.5           | 0             | down  | 2 000 000 | 22         |

表4 对话模型对比

| 评价指标       | Seq2seq | Multi_en+<br>atten_de | Multi_en+atten<br>_de+GRU | Multi_en+atten_de+<br>GRU+加强训练 |
|------------|---------|-----------------------|---------------------------|--------------------------------|
| preplexity | 26.01   | 15.47                 | 10.21                     | 3.55                           |
| BLEU       | 0.0573  | 0.0939                | 0.1856                    | 0.2675                         |

表4表明层数都为3层,神经元都为256个,训练次数都为2000 000的情况下,本文所使用的模型结构

和训练方法获得最好的结果,困惑度只有3.55,且bleu值能提高到0.2675.

对于表5中的问句,模型最后获得两个句子:“你好,根据你的描述,你这种情况应该是有便秘的情况.建议口服,可以口服润肠通便的药物,如可以口服润肠通便的药物。”和“你好,根据你的描述,我们的需求,最终模型取第一个作为最终结果.

表5 对话模型问答样例

|             | 内容  |
|-------------|---|
| 问句          | 这/几天/饿/不饿/肚子/都/是/鼓鼓的/,/并且/偶尔/有/腹部/隐痛/,/痛感/不/明显/. /今天/中午/吃/了/三碗/饭多/,/并/没有/饱腹感/,/过/一会/感觉/比较/饿/. /趴在/桌子/上/睡/一会/之后/起来/感觉/肚子/里/一股/气/,/接着/打/几个/嗝/就/舒服/了./ |
| 模型生成的<br>问句 | 你好/,/根据/你/的/描述/,/你/这种/情况/应该/是/有/便秘/的/情况/. /建议/口服/,/可以/口服/润肠/通便/的/药物/,/如/可以/口服/润肠/通便/的/药物/./   |

我们咨询了消化内科领域的专家,给出的回答考虑是胃肠胀气的症状,建议您首先注意饮食,避免吃生冷辛辣刺激油腻的食物,可以适当口服吗丁啉,再热敷一下腹部,症状应该会有所改善”. 概率分别为0.87和0.81,其实可以发现两种答案都满足为:“考虑胃肠功能紊乱的可能性较大,可以吃点促进胃肠蠕动药物,比如多潘立酮片或者是枸橼酸莫沙必利胶囊,适当运动,不要长时间久坐”. 在语义方面和模型生成的是接近的,都是肠胃蠕功能出现异常的回答,证明本文得出的模型能够有效的回答.

### 3 结论与展望

通过实验发现增加消化科比较充足的自定义词典和增加前后选择机制,分词的准确率能够达到97.3%,支持向量机和主动学习的结合能够有效获取均衡数据. 本文使用的模型结构,与传统主流的模型相比,困惑度和BLEU值都要好,且在键值对向量和Word2Vec向量两者组合增加训练法,对话系统的性能得到了进一步的提升. 为了进一步提高实用性,接下来的研究工作,会研究消化科语料中的语义信息和数据特征,然后研究增强词之间的相关性,以及寻找更好的模型结构和训练方法.

#### 参考文献

- 1 Bodenreider O. The unified medical language system (UMLS): Integrating biomedical terminology. *Nucleic Acids Research*, 2004, 32(S1): D267–D270.
- 2 奉国和, 郑伟. 国内中文自动分词技术研究综述. *图书情报*

工作, 2011, 54(2): 41–45.

- 3 Sun MS, T'sou BK. Ambiguity resolution in Chinese word segmentation. *Proceedings of the 10th Pacific Asia Conference on Language, Information and Computation*. Hong Kong, China. 1995. 121–126.
- 4 白龙飞, 王文剑, 郭虎升. 一种新的支持向量机主动学习策略. *南京大学学报(自然科学)*, 2012, 48(2): 182–189.
- 5 Mikolov T, Sutskever I, Chen K, *et al.* Distributed representations of words and phrases and their compositionality. *Proceedings of the 26th International Conference on Neural Information Processing Systems*. Lake Tahoe, NV, USA. 2013. 3111–3119.
- 6 Mikolov T, Chen K, Corrado G, *et al.* Efficient estimation of word representations in vector space. *arXiv preprint arXiv: 1301.3781*, 2013.
- 7 Chung J, Gulcehre C, Cho KH, *et al.* Empirical evaluation of gated recurrent neural networks on sequence modeling. *arXiv preprint arXiv: 1412.3555*, 2014.
- 8 Sutskever I, Vinyals O, Le QV. Sequence to sequence learning with neural networks. *Proceedings of the 27th International Conference on Neural Information Processing Systems*. Montreal, Quebec, Canada. 2014. 3104–3112.
- 9 Luong MT, Pham H, Manning CD. Effective approaches to attention-based neural machine translation. *arXiv preprint arXiv: 1508.04025*, 2015.
- 10 Sundermeyer M, Schlüter R, Ney H. LSTM neural networks for language modeling. *Proceedings of the International Speech Communication Association*. Portland, OR, USA. 2012. 601–608.
- 11 Qiu MH, Li FL, Wang SY, *et al.* *Alime chat*: A sequence to sequence and rerank based chatbot engine. *Proceedings of*

- the 55th Annual Meeting of the Association for Computational Linguistics. Vancouver, Canada. 2017. 498–503.
- 12 Liu CC, Sun H, Du N, *et al.* Augmented LSTM framework to construct medical self-diagnosis android. Proceedings of the IEEE 16th International Conference on Data Mining. Barcelona, Spain. 2016. 251–260.
  - 13 Vinyals O, Kaiser Ł, Koo T, *et al.* Grammar as a foreign language. Proceedings of the 28th International Conference on Neural Information Processing Systems. Montreal, Quebec, Canada. 2015. 2773–2781.
  - 14 Bahdanau D, Cho K, Bengio Y. Neural machine translation by jointly learning to align and translate. arXiv preprint arXiv: 1409.0473, 2014.
  - 15 Freitag M, Al-Onaizan Y. Beam search strategies for neural machine translation. arXiv preprint arXiv: 1702.01806, 2017.

WWW.C-S-A.ORG.CN

WWW.C-S-A.ORG.CN