

基于 LSTM 的食品安全自动问答系统方法研究

陈 瑛 陈昂轩 董玉博 赵筱钰 侯文俊
(中国农业大学信息与电气工程学院, 北京 100083)

摘要: 为高效、准确、全面获取食品安全相关信息,以食品安全文本为研究对象,采用 Lucene 全文检索架构和长期记忆神经网络(Long short-term memory, LSTM)构建了食品安全自动问答系统。依托于从互联网爬取的文本作为非结构化数据集,利用检索架构扩充人工标注的问题答案对规模,并以此训练了可以判断问题和答案候选句匹配程度的 LSTM 模型。基于 Lucene 检索机制进行答案候选集提取和基于 LSTM 模型进行答案提取,构建了一个可根据食品安全相关问题给出答案所在句子的问答系统,并对比了基于 Lucene 直接检索的答案抽取和基于 LSTM 的答案抽取这两种方法。结果表明,当候选文档数量增加时,基于 LSTM 模型的问题答案匹配方法,其平均准确度始终高于基于 Lucene 检索方法的平均准确度;而候选句子数量较小时,基于 LSTM 模型的问题答案匹配方法的平均准确度也高于基于 Lucene 检索方法的平均准确度。

关键词: 食品安全; 问答系统; 答案抽取模型; LSTM 技术; 深度学习

中图分类号: TP182; TS201.6 **文献标识码:** A **文章编号:** 1000-1298(2019)S0-0380-05

Methods of Food Safety Question Answering System Based on LSTM

CHEN Ying CHEN Angxuan DONG Yubo ZHAO Xiaoyu HOU Wenjun
(College of Information and Electrical Engineering, China Agricultural University, Beijing 100083, China)

Abstract: Nowadays, food safety issues have been concerned by both governments and consumers. However, the increasing number of food safety related articles makes it difficult to retrieve useful information from the articles in a short time with high accuracy. In order to improve the efficiency and accuracy of accessing food safety information, a question answering system was proposed, which was based on long short-term memory (LSTM) and information retrieval techniques. The system relied on the food safety unstructured texts obtained by Web crawler technologies, and question answer pairs were selected by using Lucene, and LSTM was used to predict answers according to matching degrees between question and candidate sentences. Based on Lucene's retrieval mechanism and the LSTM model, a question answering system was built which can select sentences that were most likely to contain the answer to given questions. The results showed that the proposed system outperformed the baseline which was only based on retrieval mechanism. Moreover, the performance analysis were made for the two methods with respect to the numbers of candidate articles and candidate sentences.

Key words: food safety; question answering system; answer retrieval model; LSTM; deep learning

0 引言

目前,我国食品安全的首要任务是对食品安全态势有较为客观的把握^[1]。因此,对食品安全事件进行全面、有效的监管对于控制食品安全事件的发生有着重要的意义。

由于信息不对称问题^[2],食品安全类信息的数

据常难以获取。目前,针对食品安全事件语料主要来源于新闻报道^[3]。但是,新闻文本长短不一,答案分布具有不确定性,对于一个食品安全相关问题,在食品安全事件新闻报道中人工寻找答案效率较低。因此,本文构建一个食品安全自动问答系统,该系统可提供相关食品安全事件的查询,给出包含答案的相关语句。

问答系统是一种特殊的信息检索系统。用户输入是以自然语言形式描述的提问,问答系统从大量的数据中查找出准确、简洁的回答,并反馈给用户^[4]。根据处理对象的不同,问答系统可以分为基于答案对、基于自由文本和基于结构化数据库 3 种类型^[5]。本文的食品安全问题自动问答系统基于自由文本这种非结构化数据,即一个由网络爬虫爬取的食品安全新闻报道构成的自由文本库。针对用户给定的问题,食品安全问题自动问答系统以实现在这个文本库中寻找可能包含答案的相关句子为目的。

基于非结构化数据的问答系统大致可分为 3 个步骤:问题分析、信息检索和答案抽取。问题分析需理解用户问题,并加以分类;信息检索需在文档中利用关键词进行检索,其核心内容是分析问题,并在后台文档库中进行检索匹配^[6],文档检索的精度和速度对系统的整体性能影响很大^[7];答案抽取常使用复杂的自然语言处理技术,考虑候选文档的语义相关性等^[8]。

经过近些年的发展,问答系统已经成为自然语言处理领域和信息检索领域的一个重要分支和新兴的研究热点^[9]。文献[10-11]阐述了早期著名的 LUNAR 系统和 MURAX 系统,实现了对自然语言问答的尝试。文献[12]实现了基于高性能知识库的问答。文献[13-17]对现今美国国家标准与技术研究院组织的 TREC 问答系统评测进行分析。综上所述,英文问答系统研究已经有显著进展,且研究领域也得到了很好的拓展。对于较为复杂的中文问答系统,也有学者进行了探索^[18-22]。

本文结合搜索引擎技术和深度学习技术,针对食品安全事件的自动问答系统进行研究。首先,基于网络爬虫技术和文本分类技术建立一个自动标注的食品安全问题语料库和一个人工标注的食品安全问题答案对语料库;其次,针对每个问题,利用 Lucene 全文检索技术在食品安全事件语料库中提取答案候选集;最后,利用基于 LSTM 的答案抽取模型,从答案候选集中提取出最终答案。其中,基于 LSTM 的答案抽取模型是本文的研究重点,主要包括两部分:利用 Lucene 全文检索技术对人工标注的食品安全问题答案对语料库进行扩充,及利用这个扩充后的问题答案对语料库和 LSTM 技术训练一个答案抽取模型。

1 材料与方法

1.1 语料获取

1.1.1 食品安全事件语料库

利用网络爬虫技术爬取食品伙伴网中的食品安全相关新闻报道(约 17 万篇),再利用文本分类技

术自动抽取出其中的食品安全事件新闻(约 7 000 篇),形成一个自动标注的食品安全问题语料库。

针对这个食品安全事件语料库中的一些文档,本文采用人工标注方式进行食品安全问题提问及答案选取。最后,共标注了 899 条问题答案对,形成一个人工标注的食品安全问题答案对语料库。

1.1.2 基于 Lucene 的问题答案对扩充

针对给定的一个人工标注的问题答案对,本文利用 Lucene 在食品安全事件语料库中搜索和问题最匹配的文档,并提取出其中包含答案的句子,从而扩充匹配的问题答案对(即正样本)。Lucene 全文检索技术是一个优秀的开源全文本搜索技术框架^[23],可以为文本数据建立反向索引,根据索引进行准确快速的查询。然后,本文抽取文档中无法有效回答问题的句子作为不匹配的问题答案对(即负样本)。为保证神经网络的准确性和泛化能力,保证训练数据当中正负样本的比例各占 50%。

在扩充后的问题答案对语料中,每个问题答案对的格式如下:[问题,答案候选句,类别]。其中,类别‘1’表示该问题和答案候选句是匹配的;类别‘0’表示该问题和答案候选句是不匹配的。

例如,匹配的问题答案对:[‘北京 2018 年哪里出现了问题烤鸭?’,’2018 年 6 月 6 日北京某某烤鸭店有顾客反映烤鸭变质’,1];不匹配的问题答案对:[‘北京 2018 年哪里出现了问题烤鸭?’,’记者了解到本次过期的主要原因是冷库损坏’,0]。

1.2 答案抽取方法

给定问题答案对语料,采用 word2vec 进行词的向量表示,采用 LSTM 进行问题答案对的二分类模型训练,即判断该问题和答案候选句是否匹配。

1.2.1 基于 word2vec 方法的词向量生成

将自然语言应用于计算和分析首先需要对其进行数值表示。首先,使用文献[24]的分词方式,将文本通过异构注释的方式进行分词,如:“桶装水有什么问题?”将被分词为“桶装水 NN 有 VE 什么 DT 问题 NN ? PU”。其中字符(如“NN”)表示分词出的词性(如“名词”)。通过异构注释方法的分词能够更加精准地完成分词任务,提高训练模型准确率。

其次,当前广泛使用的两种词语数值表示方式分别是:独热表示法(One-hot representation)和词嵌入表示法(Word embedding)。独热表示法就是把每个词转为离散的单独的符号,直接面向词语建立标记矩阵,任意两个词之间都是孤立的,这对于语义的关联分析十分不利。而词嵌入表示法是把自然语言

的词语在特定的向量空间中生成连续实向量的词语表示方法。其中,各个维度的值代表在语义和语法上某些特征^[25],含义相近的词语会被对应映射到向量空间中的相邻位置,很好地保留了原有语义。

采用词嵌入表示法,将食品伙伴网食品安全新闻报道(约 17 万篇)作为训练语料,采用 word2vec 中的 Skip - Gram 模型进行训练,得到每个词的向量表示。其中,词向量维度设置为 20,窗口值设置为 5,alpha 值设置为 0.01。

1.2.2 基于 LSTM 的答案抽取模型

循环神经网络^[26] (Recurrent neural networks, RNNs)是前向传播神经网络的一种。模型的隐层加入了自连接和互连接,并且通过相应的矩阵运算传播延迟信号,从而使得 RNN 获得短期记忆属性^[27],但是传统的 RNN 神经网络随着层数的增加会引起梯度消失或梯度爆炸的问题,难以处理长序列问题^[28]。

长短期记忆神经网络(Long short-term memory, LSTM)对 RNN 进行了改进。LSTM 引入了“细胞状态”和“门”进行状态传递和控制,能够较好地解决序列中的较长期依赖问题。在自然语言处理中也有助于保留序列整体的语义信息,清除无用的状态信息^[29]。

实验当中,本文基于 LSTM 模型构造了判断答案候选句和问题是否匹配的模型,如图 1 所示。

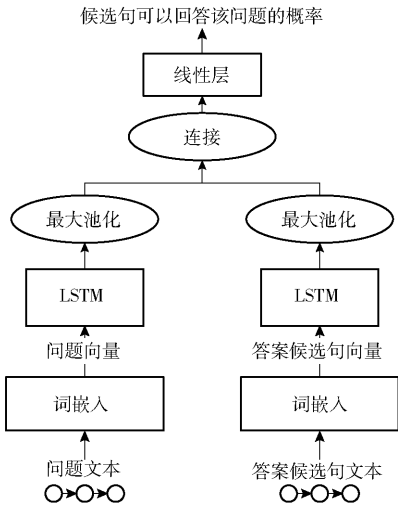


图 1 基于 LSTM 的问题答案匹配方法

Fig. 1 Question - answer matching method based on LSTM

对于输入的分好词的问题句 S ,有

$$S = \{w_1, w_2, \dots, w_i, \dots, w_n\} \tag{1}$$

式中 w_i ——句子中的第 i 个词

首先通过 word2vec 生成词 w_i 对应的词向量 E_i 。将 E_i 通过 LSTM 编码得到对应的向量 h_i 。

对于问题句 S 通过最大池化得到代表问题文本的向量

$$V_Q = \text{MaxPooling}(h_1, h_2, \dots, h_n) \tag{2}$$

同样地,对于答案候选句 S'

$$V_A = \text{MaxPooling}(h'_1, h'_2, \dots, h'_n) \tag{3}$$

式中 V_A ——答案候选句文本的向量

$\text{MaxPooling}(\cdot)$ ——最大池化方法

连接 V_Q, V_A ,得到特征向量

$$V = [V_Q \quad V_A] \tag{4}$$
$$P = \sigma(W_R^T V + b) \tag{5}$$

其中

$$\sigma(x) = \frac{1}{1 + e^{-x}} \tag{6}$$

式中 P ——处理该特征向量得到问题答案相匹配的概率

W_R ——线性层权重矩阵

b ——偏置

2 实验结果与分析

2.1 实验设置

2.1.1 数据划分及实验过程

随机从人工标注的食品安全事件问题答案对语料库中抽取其中的 80 条作为测试集,其余作为训练集。对于训练集,利用基于 Lucene 的问题答案对扩充方法得到一个问答对(约 19 160 条)。在实验中,问答对对语料按照 9:1 的比例分为训练集和交叉验证集,供 LSTM 模型使用。

2.1.2 评价标准

在实验过程中,若系统抽取的句子中包含人工标注的答案,则认定系统输出结果正确。若不含有人工标注的答案则认为结果错误。

基于 LSTM 的食品安全自动问答系统性能的评价主要采用平均准确度 $\bar{P}$,计算式为

$$\bar{P} = \frac{1}{n} \sum_{k=1}^n \lambda(q_k) \tag{7}$$

其中

$$\lambda(q_k) = \frac{A}{A+B} \times 100\% \tag{8}$$

式中 A ——系统给出包含正确答案的结果个数

B ——系统给出不包含正确答案的结果个数

n ——测试集问题数目

$\lambda(q_k)$ ——第 k 条测试问题对应的准确度

2.1.3 测试方法

如图 2 所示,在测试阶段,对于每一个测试问题,首先利用 Lucene 抽取近似度最高的文档作为候选文档,框定候选文档范围,提取出文档中的所有句子作为针对于当前测试问题的答案候选集。然后,采用基于 LSTM 的答案抽取模型对答案候选集中每一个句子和当前问题的匹配度进行评分,该得分反映了该句子可以回答该问题的概率。按照得分由高

到低的顺序抽取一些句子作为该测试问题的答案。

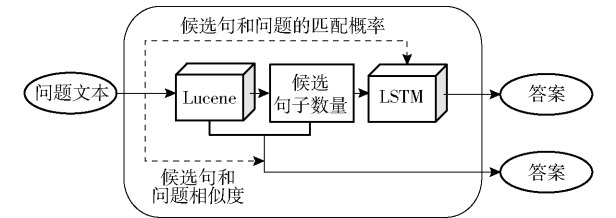


图 2 测试阶段的方法
Fig. 2 Methods for test

2.1.4 基准实验

采用基于 Lucene 的答案抽取方法作为基准模型,其将答案候选集重新生成索引, Lucene 会按照与测试问题的匹配度将所有的候选句子进行排序,最后选取得分较高的句子作为答案。

2.2 结果分析

从候选文档数量和候选句子数量两个方面分别对基于 LSTM 的答案抽取方法和基于 Lucene 的答案抽取方法进行比较。

在候选文档数量方面,从图 3 发现,当候选文档数量增加时,基于 LSTM 的问题答案匹配方法的平均准确度至少比基于 Lucene 的检索方法得到的平均准确度高出 7~8 个百分点,体现出 LSTM 模型在文档数量较多时给定问题寻找正确答案的能力相对较强。

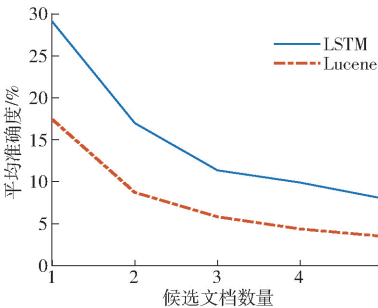


图 3 平均准确度和候选文档数量的关系
Fig. 3 Average precision with regard to number of candidate articles

在候选句子数量方面,随着候选句子数量的增加,两种方法的平均准确度都有不同程度的下降。从表 1 可得出,基于 LSTM 的问题答案匹配方法在抽取答案句个数不多于 2 个时,其平均准确度优于基于 Lucene 的检索方法,但随着候选句子数量的增加,基于 LSTM 的问题答案匹配方法的平均准确度

降低速度较快。

表 1 平均准确度与候选句子数量的关系

Tab. 1 Average precision with regard to number of extracted sentences

候选句子数量	平均准确度		%
	LSTM	Lucene	
1	29.11	17.44	
2	14.79	13.95	
3	10.43	12.40	
4	8.16	11.33	
5	6.84	9.30	

通过分析,基于 LSTM 的问题答案匹配方法的平均准确度降低速度较快的原因可能是其形式概率的相对值相差较大,导致随着候选句子集的增加, LSTM 模型给出的答案候选集个数基本不变,使其匹配的答案准确率下降较快。

对于 LSTM 模型未能正确打标签的情况,本文对其例子做出如下讨论:

预测错误例子:问题:“福州 哪儿 的 食堂 发生 食物 中毒?”;答案:“她 的 班主任 张老师 已 要求 全班 学生 不要 去 食堂 吃饭。”;系统标签:1;真实标签:0。

可以看出,此句问题答案对中的问题和答案均含有“食堂”关键词,而且问题询问的是和“食堂”同类型的地点问题,容易造成模型的混淆,从而使 LSTM 模型得到的标签与真实值不符合。

3 结论

(1) 针对食品安全问答系统的特点,基于 Lucene 文本搜索技术框架实现答案候选集的抽取,提出了基于 LSTM 技术的答案抽取方法,有利于提高问答系统的准确度。

(2) 对比了基于 Lucene 检索机制的答案抽取方法和基于 LSTM 的答案提取方法,发现基于 LSTM 的问题答案对匹配方法的平均准确度相对较高。

(3) 基于 LSTM 的问题答案对匹配方法抽取答案句个数不多于 2 个时,其准确度高于 Lucene 检索方法,同时发现,在大候选文档集下,基于 LSTM 的问题答案对匹配方法的准确度仍然优于 Lucene 检索方法。食品安全问答系统问题答案对抽取的效果不仅取决于抽取方法,与抽取候选集的大小也有一定关系。

参 考 文 献

[1] 王常伟, 顾海英. 我国食品安全态势与政策启示——基于事件统计、监测与消费者认知的对比分析[J]. 社会科学, 2013 (7):24-38.
WANG Changwei, GU Haiying. The food safety situation of our country and policy implications: comparative analysis based on incidents statistic, monitoring and consumers' perception[J]. Journal of Social Sciences, 2013(7):24-38. (in Chinese)

- [2] 全世文, 曾寅初. 消费者对食品安全信息的搜寻行为研究: 基于北京市消费者的调查[J]. 农业技术经济, 2013(4): 43–52.
- [3] 沈思, 胡业勋. 基于深度学习的食品安全事件实体自动抽取模型研究[J]. 信息与电脑(理论版), 2018(23): 11–13, 16. SHEN Si, HU Yexun. Research of automatic extraction of entities from of food safety event based on deep learning model[J]. China Computer & Communication, 2018(23): 11–13, 16. (in Chinese)
- [4] 李沛晏, 朱露, 吴多胜. 问答系统综述[J]. 数字技术与应用, 2015(4): 69–71.
- [5] 毛先领, 李晓明. 问答系统研究综述[J]. 计算机科学与探索, 2012, 6(3): 193–207. MAO Xianling, LI Xiaoming. A survey on question and answering systems[J]. Journal of Frontiers of Computer Science and Technology, 2012, 6(3): 193–207. (in Chinese)
- [6] 吴灵慧. 问答系统研究综述[J]. 科技传播, 2019, 11(5): 147–148.
- [7] COLLINS-THOMPSON K, CALLAN J, TERRA E, et al. The effect of document retrieval quality on factoid question answering performance[C] // Proceedings of the 27th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval (SIGIR 04). New York, NY, USA: ACM, 2004: 574–575.
- [8] 王慧慧. 基于自然语言处理的问答系统研究[D]. 成都: 电子科技大学, 2006.
- [9] 吴友政, 赵军, 段湘煜, 等. 问答式检索技术及评测研究综述[J]. 中文信息学报, 2005, 19(3): 1–13. WU Youzheng, ZHAO Jun, DUAN Xiangyu, et al. Research on question answering & evaluation: a survey[J]. Journal of Chinese Information Processing, 2005, 19(3): 1–13. (in Chinese)
- [10] WOODS W A. Lunar rocks in natural English: explorations in natural language question answering[C] // Linguistic Structures Processing, Volume 5 of Fundamental Studies in Computer Science, North Holland, 1977: 521–569.
- [11] JULIAN K. MURAX: a robust linguistic approach for question answering using an on-line encyclopedia [C] // Proceedings of the Sixteenth Annual International ACM SIGIR Conference on Research and Development in Information Retrieval, 1993: 181–190.
- [12] PAUL C, ROBERT S, ERIC J, et al. DARPA high-performance knowledge bases project [C] // AI Magazine, 1998: 25–49.
- [13] ELLEN M V, DAWN M T. The TREC–8 question answering track evaluation[C] // The Eighth Text REtrieval Conference (TREC–8). Washington DC: NIST, 1999: 77–82.
- [14] ELLEN M V. Overview of the TREC2003 question answering track[C] // Proceedings of the Twelfth Text REtrieval Conference (TREC 2003), 2003.
- [15] ELLEN M V. Overview of the TREC–9 question answering track[C] // The Ninth Text REtrieval Conference (TREC–9). Washington DC: NIST, 2000: 77–82.
- [16] ELLEN M V. Overview of the TREC2001 question answering track[C] // The Tenth Text REtrieval Conference (TREC–01). Washington DC: NIST, 2001: 42–51.
- [17] ELLEN M V. Overview of the TREC2002 question answering track[C] // The Eleventh Text REtrieval Conference (TREC–02). Washington DC: NIST, 2002.
- [18] LI Xiaoyan, CROFT W B. Evaluating question-answering techniques in Chinese [C] // Proceedings of the First International Conference on Human Language Technology Research, 2001: 1–6.
- [19] 郑实福, 刘挺, 秦兵, 等. 自动问答综述[J]. 中文信息学报, 2002, 16(6): 46–52. ZHENG Shifu, LIU Ting, QIN Bing, et al. Overview of question-answering[J]. Journal of Chinese Information Processing, 2002, 16(6): 46–52. (in Chinese)
- [20] 张刚, 刘挺, 郑实福, 等. 开放域中文问答系统的研究与实现[C] // 中国中文信息学会二十周年学术会议, 2001.
- [21] WU Lide, HUANG Xuanjing, NIU Junyu, et al. FDU at TREC 2002: filtering, Q & A, web and video tasks[C] // 11th Text REtrieval Conference. Gaithersburg, 2002.
- [22] XU Hongbo, ZHANG Hao, BAI Shuo. ICT experiments in TREC–11QA main task[C] // The Eleventh Text REtrieval Conference (TREC–11), 2002.
- [23] 管建和, 甘剑峰. 基于 Lucene 全文检索引擎的应用研究与实现[J]. 计算机工程与设计, 2007, 28(2): 489–491. GUAN Jianhe, GAN Jianfeng. Design and implementation of web search engine based on Lucene[J]. Computer Engineering and Design, 2007, 28(2): 489–491. (in Chinese)
- [24] SUN Weiwei, WAN Xiaojun. Reducing approximation and estimation errors for Chinese lexical processing with heterogeneous annotations[C] // Proceedings Association for Computational Linguistics, 2012.
- [25] 周练. Word2vec 的工作原理及应用探究[J]. 科技情报开发与经济, 2015, 25(2): 145–148. ZHOU Lian. Exploration of the working principle and application of Word2vec[J]. Sci-Tech Information Development & Economy, 2015, 25(2): 145–148. (in Chinese)
- [26] SCHMIDHUBER J. Deep learning in neural networks: an overview[J]. Neural Networks, 2015, 61: 85–117.
- [27] TOMÁŠ M, KOMBRINK S, LUKÁŠ B, et al. Extensions of recurrent neural network language model[C] // Acoustics, Speech and Signal Processing (ICASSP). 2011 IEEE International Conference on. IEEE, 2011.
- [28] KOLEN J F, KREMER S C. Gradient flow in recurrent nets: the difficulty of learning long term dependencies [M]. Pensacola: Univ. of West Florida, 2001.
- [29] 闫琰. 基于深度学习的文本表示与分类方法研究[D]. 北京: 北京科技大学, 2016. YAN Yan. Text representation and classification with deep learning[D]. Beijing: University of Science & Technology Beijing, 2016. (in Chinese)