Universitat Oberta de Catalunya (UOC)

Máster Universitario en Ciencia de Datos (*Data Science*)

TRABAJO FINAL DE MÁSTER

Área: 4

Answer generation for retrieval based questions for breastfeeding advice from noisy, user-generated content.

Author: Arcadi Gonzalez Graells

Tutor: Elisenda Bonet Carne

Professor: Nadjet Bouayad-Agha

London, October 23, 2022

Abstract

Question-answering systems for healthcare have been widely implemented, so in the specific area of breastfeeding, where women may not have access to relevant, convenient, on-demand information an automated QA system could prove a valuable resource. With that aim in mind, using expert generated content and user submitted queries, a QA system will be built using state-of-the-art techniques and technologies, leveraging pre-trained models, open source algorithms, and commercially available solutions where relevant.

Els sistemes de resposta a preguntes per a l'assistència sanitària s'han implementat àmpliament, de manera que en l'àrea específica de la lactància materna, on les dones poden no tenir accés a informació rellevant, convenient i immediata, un sistema de pregunta-resposta automatitzat podria crear valor per les usuàries. Amb aquest objectiu en ment, utilitzant contingut generat per experts i consultes enviades per les usuàries, es construirà un sistema de pregunta-resposta utilitzant tècniques i tecnologies d'última generació, aprofitant models pre-entrenats, algorismes de codi font obert i solucions comercials quan sigui necessari.

Keywords: Information Retrieval, Question-Answering, Breastfeeding, Natural Language Processing

Contents

Chapter 1

Introduction

1.1 Proposal Summary

The aim of this project is to build a domain specific question answering engine to provide mothers, and prospective mothers with answers to breastfeeding related queries. The final product should accept natural language questions from users and, where possible, retrieve relevant answers. When those answers are not available in the knowledge base, the user should be not provided an inaccurate answer, but rather given a message detailing that the question was not automatically answerable, or the question should be escalated to a healthcare professional that can address it.

1.2 Rationale

While machine learning has seen significant and sustained increase in usage in healthcare most of that research has mainly focused on Imaging, Public Health and Bioinformatics [15]. Patient facing question answering systems and chatbots have seen widespread usage [1], seeing significant success in the mental health space [10]. Conversational agents have been built for the specific purpose of assisting mothers in regard to breastfeeding, with Public Heath England, through the Start4Life Breastfeeding Friend chatbot, found that new mothers reported they were more likely to attempt breastfeeding than not (59%, n=1000), when in conjunction with other support strategies [8]. Given the success rate of the Start4Life initiative, which provides answers to 42 pre-defined questions [17], a more novel approach using information retrieval techniques should lead to a higher number of possible questions receiving successful answers.

1.3 Personal Motivation

I have been working in Machine Learning for the last five years, with a heavy focus on NLP. I have been able to work on a wide breadth of problems during my career, but while being very interested in the subject, I have never had an opportunity to commercially implement a question and answer system. At the same time, most of my work has had a strictly commercial focus, so being able to work towards a project with a positive social impact is something that drove me to select this specific topic.

1.4 Objectives

The primary objective of the project is to produce accurate answers to breastfeeding related questions. This can be further refined into 3 secondary objectives:

- **Identifying the question type**: Questions must first be divided between factoid questions, and questions containing context.
- **Retrieving the relevant passage for questions**: Extract the passages that may be potential answers for the question, and rank them by relevance.
- **Extracting the specific factoid from the passage**: Find the specific answer to the question within a passage, extracting the relevant keywords or sentence chunks.

1.5 Hypothesis

It is the aim of the project to prove that, given a set of training questions and answers, as well as a tagged of expert articles, user breastfeeding related user generated questions can be answered accurately.

1.6 Methodology

The main data source for the project is the platform LactApp. Two separate sources from the application will be used:

- **Expert content**: The platform contains numerous expert written articles regarding most aspects related to breastfeeding, categorized by topic and other classifications.
- **Conversations**: The live chat functionality has collected conversations between mothers and experts. These range from general factual questions (e.g. "How often do newborns

feed?"), to very specific questions requiring contents (e.g. "I suspect my 3-month-old has bronchitis and is not latching, what should I do?").

As well as these two datasets other breastfeeding FAQ documents and sites will be used to train the restricted domain model.

1.6.1 Research Strategy

The initial stage of the project will consist of evaluating the current state of the art in terms of research. This will be focused on the four main paradigms of question answering [4]: Natural language processing and understanding, information retrieval, knowledge base retrieval, and hybrid methods, with a focus on restricted domain question answering. As part of the literature review, current technology implementations will also be evaluated, including, but not limited to, BERT [6], Haystack [9] as well as other information retrieval systems.

The second stage of the project will focus on evaluating the datasets and separating the training and testing datasets. The data cleaning strategy will also be defined and implemented at this stage. Once the data has been prepared, and a testing dataset defined, the representative sample of questions will be set aside for evaluating the model at the conclusion of the project.

The third stage will consist of building the question answering model iteratively, testing the improvement in quality with each modification applied to the pipeline, including features generated, tweaking machine learning models, or adding new stages on the pipeline.

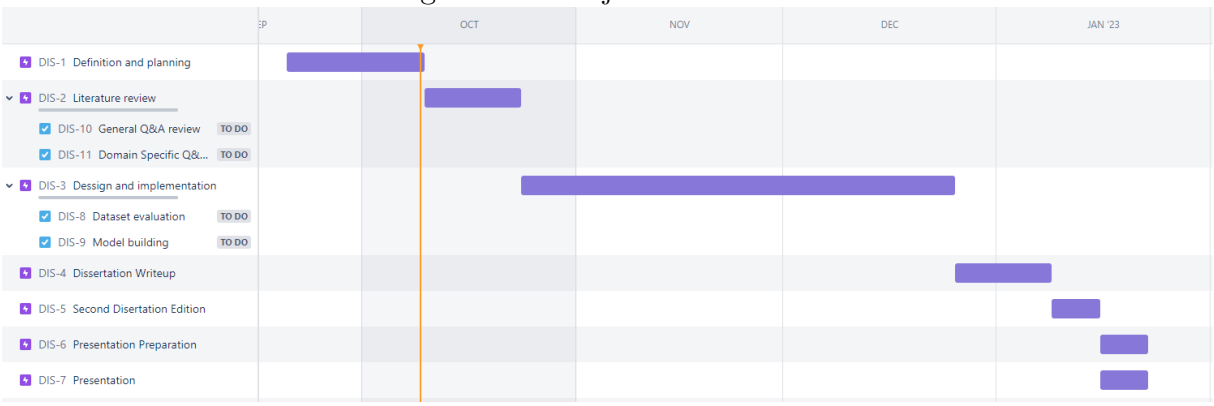
1.6.2 Project Methodology

The desired outcome of the project is to create a new product, by adapting existing open source tools, training custom models for the target data, and using commercially available tools to build an end-to-end solution that can assist mothers with their breastfeeding queries, served via an API.

1.6.3 Project Plan

As detailed in figure 1.1, the project plan has been scheduled finish the implementation work by the 25th of December, and the finished paper by the 8th of January.

Figure 1.1: Project Gantt Chart



1.7 Global ethical commitment and Sustainable Development Objectives

If the question answering accuracy is high enough, the project should reduce the turn-around time for mothers' queries, leading to better advice access for end users.

1.7.1 Sustainability

The main carbon footprint impact of the project will likely be in compute, given the size of the dataset, the likelihood is that the impact on the environment will be minor, given that the bulk of the work will be performed on a single, consumer grade hardware, workstation.

1.7.2 Diversity and Inclusivity

Active care will be taken to ensure that any biases present in the training data do not carry to the machine learning model. A more specific strategy will be devised after evaluating the training dataset.

1.7.3 Ethical behaviour and social responsibility

The data used will be anonymized and kept secure, limiting both the likelihood of data leaks, as well as mitigating the potential impact on the platform users.

Chapter 2

Literature Review

2.1 Introduction

There are several methods to build question answering systems, this literature review focuses on document processing and answer processing, detailing the most common and latest techniques used when building the systems.

2.2 Term similarity based QA

Traditional search engines use terms in the search query to match the similarity of the question to the available documents. One of the first systems to implement term based search was Smart (System for the Mechanical Analysis and Retrieval of Text) [16], developed in 1971, it used TF-IDF to rank the relevant documents for a query. Similar implementations of Smart have lead to successful open-source and commercial implementations like Lucene [3] or Elasticsearch [7], which are widely used as single-website search engines [2].

Term based searches can be further refined to extract passages within the documents, these can be defined to be individual sentences, series of sentences, or other pre-defined or dynamic lengths of text. There are several techniques to extract and rank the passages [18], which include:

- **MITRE**: Each individual sentence becomes a passage, and word overlap is used to rank them [12].

- **MultiText**: The passage is determined by a variable sized window that starts and ends with a query term, creating passages of arbitrary length. The passages are ranked by inverse document frequency while penalizing longer passages [5].

- **Alicante**: The passage length is parametrized at runtime, by specifying the number of sentences to be included in each overlapping window passage. Each window is then scored using cosine similarity between each sentence and the query's TF-IDF vector, as well as the similarity amongst the sentences within the passage [19].

2.3 Natural Language Processing

Layering Natural Language Processing techniques over traditional information retrieval methods adds nuance to the extraction. Multiple techniques can be used to improve results, encoding both the query and the potential passages:

- **TF-IDF**: A non-contextual purely statistical methods use the relative frequency of a term when compared to the frequency of the same term over the whole corpus [13]. While efficient and widely used, it does not use any external corpora to improve results. Cosine similarity is generally used between the encoded question vectors and the encoded passages.
- **Word2Vec**: The encoder is built by training a shallow neural network to learn word associations based on their context [14]. The vectors can be trained on a large external corpus, on the target documents containing the potential answer, or on a combination of both. The vectors of all the words in a given passage can be aggregated with Doc2Vec [11], by averaging or concatenating them, and appending a paragraph ID.
- **BERT**: This open source framework, Bidirectional Encoder Representations from Transformers [6], uses a combination of two techniques to encode passages: Masked Language Modelling, where the model is trained to predict the missing word on a given sentence; and Next Sentence Prediction, where the model is trained to predict the sentence after the current one. BERT can be further fine-tuned to the specific corpora to improve question answering quality on restricted domains.

These techniques can be used individually or in conjunction to build a question answering pipeline.

2.4 Knowledge Base QA

Knowledge Base Question Answering works with the potential answer in structured data instead of unstructured documents. These systems rely on translating unstructured questions to database queries. Questions have to be first categorized by their goal [20], e.g. is it a "who"

question or a "what" question, named entities are then extracted from the question, and the combination of those is used to map a predicate to be extracted from the knowledge base.

Bibliography

[1] Alaa Abd-Alrazaq, Zeineb Safi, Mohannad Alajlani, Jim Warren, Mowafa Househ, and Kerstin Denecke. Technical metrics used to evaluate health care chatbots: Scoping review. *J Med Internet Res*, 22(6):e18301, Jun 2020.

[2] Whit Andrews and Rita E Knox. Magic quadrant for information access technology. *Technical report, Gartner Research (G00131678)*, 2005.

[3] Andrzej Białecki, Robert Muir, Grant Ingersoll, and Lucid Imagination. Apache lucene 4. In *SIGIR 2012 workshop on open source information retrieval*, page 17, 2012.

[4] Marco Antonio Calijorne Soares and Fernando Silva Parreiras. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*, 32(6):635–646, 2020.

[5] Charles LA Clarke, Gordon V Cormack, Derek IE Kisman, and Thomas R Lynam. Question answering by passage selection (multitext experiments for trec-9). In *TREC*, 2000.

[6] Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*, 2018.

[7] BV Elasticsearch. Elasticsearch. *Internet: https://www. elastic. co/pt/,[Sep. 12, 2019]*, 2018.

[8] Public Health England. Latest technology supports new mums to breast-feed. https://www.gov.uk/government/news/new-technology-supports-new-mums-to-breastfeed, Mar 2018.

[9] Haystack. What is haystack? Website, https://haystack.deepset.ai/overview/intro, October 2022.

[10] Simon Hoermann, Kathryn L McCabe, David N Milne, and Rafael A Calvo. Application of synchronous text-based dialogue systems in mental health interventions: Systematic review. *J Med Internet Res*, 19(8):e267, Jul 2017.

[11] Quoc Le and Tomas Mikolov. Distributed representations of sentences and documents. In *International conference on machine learning*, pages 1188–1196. PMLR, 2014.

[12] Marc Light, Gideon S Mann, Ellen Riloff, and Eric Breck. Analyses for elucidating current question answering technology. *Natural Language Engineering*, 7(4):325–342, 2001.

[13] H. P. Luhn. A statistical approach to mechanized encoding and searching of literary information. *IBM Journal of Research and Development*, 1(4):309–317, 1957.

[14] Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*, 2013.

[15] Daniele Ravì, Charence Wong, Fani Deligianni, Melissa Berthelot, Javier Andreu-Perez, Benny Lo, and Guang-Zhong Yang. Deep learning for health informatics. *IEEE Journal of Biomedical and Health Informatics*, 21(1):4–21, 2017.

[16] G. Salton. *The SMART Retrieval System—Experiments in Automatic Document Processing*. Prentice-Hall, Inc., USA, 1971.

[17] Start4Life. Breastfeeding friend from start for life, what questions can i ask the breastfeeding friend? Website, https://www.nhs.uk/start4life/baby/feeding-your-baby/breastfeeding/breastfeeding-friend-from-start4life/breastfeeding-friend-on-google-home.

[18] Stefanie Tellex, Boris Katz, Jimmy Lin, Aaron Fernandes, and Gregory Marton. Quantitative evaluation of passage retrieval algorithms for question answering. In *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, pages 41–47, 2003.

[19] Jose Luis Vicedo, Antonio Ferrández, and Fernando Llopis. University of alicante at trec-10. In *AUTHOR Voorhees, Ellen M., Ed.; Harman, Donna K., Ed. TITLE The Text REtrieval Conference (TREC-2001)(10th, Gaithersburg, Maryland, November 13-16, 2001). NIST Special*, volume 500, page 419. ERIC, 2002.

[20] Min-Chul Yang, Do-Gil Lee, So-Young Park, and Hae-Chang Rim. Knowledge-based question answering using the semantic embedding space. *Expert Systems with Applications*, 42(23):9086–9104, 2015.