Question Answering in the Biomedical Domain

Vincent Nguyen
Research School of Computer Science, Australian National University
Data61, CSIRO
vincent.nguyen@anu.edu.au

Abstract

Question answering techniques have mainly been investigated in open domains. However, there are particular challenges in extending these open-domain techniques to extend into the biomedical domain. Question answering focusing on patients is less studied. We find that there are some challenges in patient question answering such as limited annotated data, lexical gap and quality of answer spans. We aim to address some of these gaps by extending and developing upon the literature to design a question answering system that can decide on the most appropriate answers for patients attempting to *self-diagnose* while including the ability to abstain from answering when confidence is low.

1 Introduction

Question Answering (QA) is the downstream task of information seeking wherein a user presents a question in natural language, Q, and a system finds an answer or a set of answers from a collection of natural language documents or knowledge bases (Lende and Raghuwanshi, 2016), A, that satisfies the user's question (Molla and Gonzlez, 2007).

Questions fall into one of two categories: factoid and non-factoid. Factoid QA provides brief facts to the users' questions; for example, *Question: What day is it? Answer: Monday*. Non-factoid question answering is a more complex task. It involves answering questions that require specific knowledge, common sense or a procedure due to ambiguity or the scope of the question. An example from the Yahoo non-factoid question answer dataset[1] illustrates this: *Question: Why is it considered unlucky to open an umbrella indoors?*. The answer is not apparent and requires specific knowledge about cultural superstitions.

[1] https://ciir.cs.umass.edu/downloads/nfL6/

Question answering is fundamental in high-level tools such as chatbots (Qiu et al., 2017; Yan et al., 2016; Amato et al., 2017; Ram et al., 2018), search engines (Kadam et al., 2015), and virtual assistants (Yaghoubzadeh and Kopp, 2012; Austerjost et al., 2018; Bradley et al., 2018). However, being a downstream task, question answering suffers from *pipeline error*, as it often relies on the quality of several upstream tasks such as coreference resolution (Vicedo and Ferrández, 2000), anaphora resolution (Ram et al., 2018), named entity recognition (Aliod et al., 2006), information retrieval (Mao et al., 2014), and tokenisation (Devlin et al., 2019).

Thus, there has been a growing demand for these QA systems to deliver precise question-specific answers (Pudaruth et al., 2016) and consequently has sparked much research into improving upon relevant natural language processing approaches (Malik et al., 2013), datasets (Rajpurkar et al., 2016; Kociský et al., 2017) and information retrieval techniques (Weienborn et al., 2013; Mao et al., 2014). These improvements have allowed the domain to evolve from shallow keyword matching to contextual and semantic retrieval systems (Kadam et al., 2015). However, most of these techniques have been focused on the open-domain (Soares and Parreiras, 2018) and the challenges harbouring the biomedical domain have not been well addressed and remain unsolved. Here, we define biomedical QA as either factoid or non-factoid QA on biomedical literature.

One such challenge is due to the creation of complex medical queries which require expert knowledge and up to four hours per query (Russell-Rose and Chamberlain, 2017) to adequately answer. This requirement of expert knowledge leads to a lack of high-quality, publicly available biomedical QA datasets. Furthermore, medical datasets tend to be locked behind ethical, obligatory agreements and are usually small due to

Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: Student Research Workshop, pages 54–63
Florence, Italy, July 28 - August 2, 2019.

cost constraints and lack of domain experts for annotation (Pampari et al., 2018; Shen et al., 2018). Therefore, open-domain techniques which assume data-rich conditions are not suitable for direct application to the biomedical domain.

Another challenge is clinical term ambiguity, which is due to the temporally and spatially varying nature of clinical terminology, and the frequent use of abbreviation and esoteric medical terminology (Lee et al., 2019) (see Table 1 for examples). It is difficult for systems to adequately disambiguate clinical words to be used in downstream QA systems due to the complexity of the ambiguity of medical terminology, such as abbreviations, due to their varying contexts. Though there are existing tools such as MetaMap (Aronson and Lang, 2010) to disambiguate these terms by mapping them to the UMLS (Unified Medical Language System) metathesaurus, coverage of these systems is low and mappings are often inaccurate (Wu et al., 2012).

Furthermore, systems in the open-domain typically retrieve a long answer before extracting a short continuous span of text to present to the user (Soares and Parreiras, 2018; Rajpurkar et al., 2016). However, for biomedical responses, it is not always sufficient to retrieve short answer continuous spans, and *Answer Evidence* spans that are discontinuous that cross the sentence boundary are often required (Pampari et al., 2018; Hunter and Cohen, 2006; Nentidis et al., 2018).

These problems are not yet solved in the biomedical domain and are reflected in the BioASQ challenge (Nentidis et al., 2018), an annual challenge with a biomedical question answering track. Currently, the state-of-the-art systems do not perform much better than random guess with an accuracy of 66.67% for binary question answering (Chandu et al., 2017), 24.24% for factoid (ranked list of named entities as answers) and an F1-score of 0.3312 for list-type (unranked list of named entities) (Peng et al., 2015) suggesting that there is much room for improvement in terms of algorithms and research.

Furthermore, we found that there is a lack of a biomedical question answering system directed for patients. Biomedical question answering for patients is important as studies from the Pew Research Centre have shown that 35% of U.S. adults have diagnosed themselves using the information they found online[2]. Of these adults, 35% said that they did not get a professional opinion on their self-diagnosis, illustrating that patients may blindly trust the results of search engines without consulting a medical professional. This is cause for concern, as search engines tend to display the most severe ailments first which could lead to a potential waste of hospital resources or deterioration in patient health (Korfage et al., 2006).

Furthermore, although there are negatives to searching symptoms via search engine, for the participants who visited doctors after *self-diagnosis*, research has revealed that doctor-patient relationships and patient compliance with treatment improve as the patients have a clearer understanding of their symptoms and potential disease after *self-diagnosis* (Cocco et al., 2018). These studies motivate the need for a strong biomedical question answering question for patients as it will benefit patients who *self-diagnose* and patients who seek medical advice after looking up their symptoms online.

Finally, we highlight that there is a lexical and semantic gap between clinical and patient language. For example, the expression *"hole in lung"* taken literally is about a punctured lung. However, this colloquialism refers to the condition known as *Pleurisy* (Ben Abacha and Demner-Fushman, 2019; Abacha and Demner-Fushman, 2016), illustrating that patients do not have the level of literacy to formulate complex medical queries nor understand them (Graham and Brookey, 2008).

We aim to address the challenges in applying question answering to biomedical question answering for patients. We highlight that the current gaps of biomedical QA research stem from lack of clinical disambiguation tools, lack of high-quality data, the quality of answer spans, weak algorithms and clinical-patient lexical gaps. Our goal is to present a patient biomedical QA system that can address the gaps in biomedical research and allows a patient to query their symptoms, diseases or available treatment options accurately, but will also abstain from providing answers in cases where there is low confidence in the best answer, question malformation or insufficiency of data to answer the question.

[2]https://www.pewinternet.org/2013/01/15/health-online-2013/

Type	Example	Explanation
Temporally varying	Flu	The Flu evolves every year and the cause is predicated on the year it is contracted
Spatially varying	Cancer	Cancer is a disease that varies with severity based on location (Late stage brain cancer is much worse than early stage skin cancer)
Abbreviation	HR	A common clinical abbreviation that typically means heart rate, but may mean hazard ratio depending on the context
Esoteric terminology	c.248T>C	A gene mutation that does not appear in any open-domain corpus such as Wikipedia and has no layman definition

Table 1: Examples of ambiguity in biomedical text.

2 Literature Review

Here, we detail a review of question answering in the open and biomedical domains.

2.1 Information Retrieval Approaches

Biomedical QA systems up until 2015 relied heavily on Information Retrieval (IR) techniques such as tf-idf ranking (Lee et al., 2006) and entity extraction tools such as MetaMap (Aronson and Lang, 2010) in order to obtain candidate answers (by querying biomedical databases) and feature extraction before using machine learning models such as logistic regression (Weienborn et al., 2013). While other techniques included using cosine similarity between one-hot encoded vectors of answer and question for candidate reranking (Mao et al., 2014). However, these techniques were inherently bag-of-word approaches that ignored the context of words. Furthermore, these techniques relied on complete matches of question terms and answer paragraphs, which is not realistic in practice. Patients use different terminology to that of medical experts and biomedical literature (Graham and Brookey, 2008).

In more recent years, more neural approaches to IR have been used in the biomedical space (Nentidis et al., 2017, 2018) such as *Position-Aware Convolutional Recurrent Relevance Matching* (Hui et al., 2017), *Deep Relevance Matching Model* (Guo et al., 2017) and *Attention Based Convolutional Neural Network* (Yin et al., 2015). However, though these approaches do not rely on complete matching of words and capture semantics, they either ignore local or global contexts which are useful for disambiguation of clinical terminology and comprehension (McDonald et al., 2018).

2.2 Semantic-level Approach

QA requires the retrieval of long answers before summarisation or retrieval of answer spans. Punyakanok et al. (2004) introduced the use of a question's dependency trees and candidate answers' dependency trees and aligning with the Tree Edit Distance metric to augment statistical classifiers such as Logistic Regression and Conditional Random Fields. However, these methods failed to capture complex semantic information due to a reliance on effective part-of-speech tagging and were not attractive end-to-end solutions. Otherwise, WordNet was utilised to extract semantic relationships and estimate semantic distances between answers and questions (Terol et al., 2007). However, WordNet suffered from being open-domain focused and also was not able to capture complex semantic information such as polysemy (Molla and Gonzlez, 2007).

2.3 Neural Approaches

In recent years, approaches that use neural networks have become popular. Word embedding techniques such as Word2vec and GloVe can model the latent semantic distribution of language through unsupervised learning (Chiu et al., 2016). Furthermore, they are quickly adopted into neural networks as these models take fixed-sized vector inputs, where embeddings could be used as encoded inputs into neural networks such as LSTM (Hochreiter and Schmidhuber, 1997) and CNN (LeCun et al., 1999) in the biomedical domain (Nentidis et al., 2017, 2018).

Though these embedding techniques were useful in capturing latent semantics, they did not distinguish between multiple meanings of clinical text (Molla and Gonzlez, 2007; Vine et al., 2015).

There have been several solutions to this prob-

lem (Peters et al., 2018; Howard and Ruder, 2018; Devlin et al., 2019) proposed but they are not relevant specifically to the biomedical domain. Instead, we highlight *BioBERT* (Lee et al., 2019), a biomedical version of *BERT* (Devlin et al., 2019) which is a deeply bidirectional transformer (Vaswani et al., 2017) that is able to incorporate rich context into the encoding or embedding process that has pre-trained on the Wikipedia and PubMed corpora. However, this model fails to account for the spatial and temporal aspects of diseases in biomedical literature as temporality is not encoded into its input. Furthermore, Biobert uses a WordPiece tokeniser (Wu et al., 2016) which keeps a fixed-size vocabulary dictionary for learning new words. However, the vocabulary within the model is derived from Wikipedia, a general domain corpus, and thus Biobert is unable to learn distinct morphological semantics of medical terms like *-phobia*, where '-' denotes suffixation, meaning *fear* as it only has the internal representation for *-bia*.

3 Research Plan

We list the research questions to address some of the research gaps in biomedical QA and the system we aim to design, alongside baseline approaches and methodology as starting points. We will also mention future directions to address these research questions.

RQ1: What are the limitations of current biomedical QA? The limitations in current biomedical QA include the lack of: sufficient ambiguity resolution tools (Wu et al., 2012), robust techniques to using semantic neural approaches (Lee et al., 2019; Nentidis et al., 2018). The lack of strong comprehension from systems to produce sufficient answer spans that cross the sentence boundary as reflected by poor results in *ideal answer production in BioASQ* (Nentidis et al., 2018, 2017) and addressing issues using real-world patient queries rather than artificially curated queries (Pampari et al., 2018; Guo et al., 2006) which contain colloquial ambiguous non-medical terminology such as *hole in lung*.

In our research, we aim to address each of these gaps by researching into: higher coverage clinical ambiguity tools that use contexts in the spatial and temporal domains, summarisation techniques that can translate from biomedical terminology to patient language (Mishra et al., 2014; Shi et al., 2018) and tuning biomedical models to solve complex answer span tasks that cross sentence boundaries (Kociský et al., 2017) or require common sense (Talmor et al., 2018).

RQ2: Data-driven approaches require high-quality datasets. How can we construct or leverage existing datasets to mimic real-world biomedical question answering? By leveraging existing techniques such as variational auto-encoder (Shen et al., 2018) and Snorkel (Bach et al., 2018), we will be able to generate, label and process additional data that can meet stringent data requirements of neural approaches.

However, synthetic datasets generally perform weaker than handcrafted datasets (Bach et al., 2018). In order to bridge this gap in the research, we propose augmenting these data generation methods via crowd-sourcing methods with textual entailment (Abacha and Demner-Fushman, 2016) and natural language inference (Johnson et al., 2016) to improve the quality of the generated labels and data. For instance, we can use forums like Quora or medical specific forums such as Health24[3] and utilise techniques such as question entailment to find questions that are related to ones seen in the dataset in order to generate higher-quality annotated labels.

We will then develop techniques that can combine synthetic and higher-quality labelled datasets that can be utilized downstream in a QA system. We will compare this against baselines such as majority voting and Snorkel to evaluate our approaches.

Allowing the model to abstain from a decision, through comprehension, has been the focus of many datasets as of late (Rajpurkar et al., 2016; Kociský et al., 2017). We can use these datasets as a starting problem to solve before applying these techniques to the biomedical domain. However, we will also develop and research further techniques in order to allow for improved confidence and low uncertainty from the model.

RQ3: How do we indicate the confidence of the answer that the model has provided? Often researchers interpret softmax or confidence scores from the classifier models as direct correlations to probability but often forget about uncertainties in this measurement (Kendall and Gal, 2017). Due to the real-world application and sensitivity of pre-

[3]https://www.health24.com/Experts

dictions in a health-based QA system, there needs to be guarantees that predictions are of both high accuracy and low uncertainty.

In order to account for uncertainty, techniques such as *Inductive Conformal Prediction* (Papadopoulos, 2008) and *Deep Bayesian Learning* (Siddhant and Lipton, 2018) can be used to model *epistemic uncertainty*, which is not inherently captured by the model during training, in order to make the loss function more robust to noise and uncertainty and thereby strengthen the predictions of the model. This would then allow softmax scores to be used as confidence scores within a reasonable level of uncertainty.

RQ4: How do we include temporality or locality of diseases into answers? Diseases are non-static, they evolve such as the flu or are seasonal such as the summer cold. Current models utilise only static vector inputs, such as word embeddings, that do not account for this temporal aspect of the input. Furthermore, though diseases are non-static, they may be more likely in different countries as there is a spatiotemporal relationship where countries will experience different seasons and thus different diseases. In order to accommodate for these relationships, we can draw on prior research as starting points such as space-time local embeddings (Sun et al., 2015), dynamic word embeddings (Bamler and Mandt, 2017) or time-embeddings (Barbieri et al., 2018) as baselines and extend them into the biomedical setting.

RQ5: How do we bridge the semantic gap between clinical text and terminology that a patient can understand? Most patients lack the expertise in utilising resources such as biomedical literature in order to self-diagnose. Therefore, knowledge or answers should be presented in a form that they can understand (Graham and Brookey, 2008). Biomedical language and patient language can be construed as two separate languages as biomedical language changes and evolves over time (Yan and Zhu, 2018) and also pose the same problems (Hunter and Cohen, 2006). Therefore, we can model this problem as a language translation problem and thus can use techniques in neural machine translation (Qi et al., 2018; Chousa et al., 2018) based on word embeddings.

However, as biomedical language and patient English are primarily borne of the same language, this poses unique problems. For instance, a token in plain English may translate to several tokens in the biomedical space or vice versa. This is known as the alignment problem (Qi et al., 2018). We can potentially remedy this by borrowing ideas from n-gram embedding (Zhao et al., 2017) as a starting point or using Biobert (Lee et al., 2019) projected to a dual-language embedding space and use attention to produce the alignment. Furthermore, there are biomedical abbreviations that need to be disambiguated before translation (Festag and Spreckelsen, 2017), for which we would use direct, rule-based approaches using thesauri or tools such as Metamap (Aronson and Lang, 2010) as our baseline approaches and extend upon using data-driven approaches (Wu et al., 2017).

4 Experimental Framework

4.1 Datasets

High-quality data is required to address the challenges we outlined. We therefore consider the following datasets: (1) MEDNLI (Johnson et al., 2016; Goldberger et al., 2000) for medical language inference; (2) i2b2 in the form of emrQA (Pampari et al., 2018) for synthetic question-answer pairs; (3) SQuAD (Rajpurkar et al., 2016) for open-domain transfer learning; (4) the question-answering datasets provided on MediQA 2019[4]; (5) the question entailment dataset and MedQuAD (Ben Abacha and Demner-Fushman, 2019); (5) CLEF eHealth (Suominen et al., 2018) to utilize and evaluate IR methods; and (6) we will supplement our datasets by generating labels for unlabelled data by leveraging the signals from the labelled datasets through the use of tools such as Snorkel (Bach et al., 2018) and CVAE (Shen et al., 2018).

4.2 Evaluation Metrics

In our experiments, we will evaluate our summarisation strategies with metrics such as ROGUE (Lin, 2004), in particular, rogue-2 (Owczarzak and Dang, 2009) and BLEU (Papineni et al., 2002). For question-answering, we use standard ranking metrics such as *Mean Average Precision* and *Mean Reciprocal Rank* for evaluating candidate ranking and standard metrics such as *f1-score*, *Precision*, *Accuracy* and more medical targeted metrics such as *sensitivity* and *specificity* (Parikh et al., 2008).

[4]https://sites.google.com/view/mediqa2019

4.3 Proposed Framework

From the research questions mentioned, we propose a framework to unify their solutions.

Embeddings To begin, we need to construct our date/seasonal embeddings (Barbieri et al., 2018), to do this, we will need datasets that have mentions of the seasonality and locality of disease entities. Also, we will require embeddings that are representative of the text, we will consider state-of-the-art word-level context sensitive embeddings (Lee et al., 2019; Peters et al., 2018) and word-level context insensitive embeddings (Chiu et al., 2016) and ensure they properly represent the biomedical datasets. For instance, *BERT* will need to pretrained with a biomedical vocabulary rather than a general purpose open-domain one, and, in doing so, we will be able to resolve ambiguity in polysemy or abbreviations.

Furthermore, we will also be researching methodologies to handle out-of-vocabulary words as the current *WordPiece* tokenization (Devlin et al., 2019) or character-level embeddings (Barbieri et al., 2018) would not be sufficient to address esoteric terminology (Lee et al., 2019). The time embeddings and the word-level embeddings will be concatenated and used as input to the model.

Model Architecture Given the success of multi-task learning (Zhao et al., 2018; Liu et al., 2019), and having been proposed as the *blocking task* in NLP (McCann et al., 2018) that needs to be solved. We therefore apply multi-task learning to this problem. From the state of the art multi-task learning models, we borrow the fundamental building blocks such as multi-headed self-attention (Liu et al., 2019) and multi-pointer generation (McCann et al., 2018) to be used as decisions in a Neural Architecture Search (NAS) (Zoph and Le, 2016). NAS will use reinforcement learning techniques to find a suitable architecture for multi-task learning. We elect to find the architecture to represent our problem this way due to one main reason. The reason is that the field of deep learning in NLP is quickly changing, and thus the state-of-the-art techniques will always change. Therefore, by having a tool that builds architectures from the building blocks of state-of-the-art models is vital. However, crucially, we must add *Heteroscedastic Aleatoric Uncertainty* and *Epistemic Uncertainty* minimisation to the model by adjusting the loss function and weight distribution which will allow the model to be more certain about decisions (Kendall and Gal, 2017). One such decision must be the ability to abstain from answering.

Concretely, we use NAS to discover models for NMT from clinical text to the patient language by conditioning to an encoder-decoder structure. From here, using this model a starting point, NAS will add task-specific layers that will minimise the joint loss over the biomedical tasks such as question answering (Nentidis et al., 2018), question entailment (Abacha and Demner-Fushman, 2016) and natural language inference (Johnson et al., 2016). In doing so, multi-task learning will allow for stronger generalisability and end-to-end training (McCann et al., 2018; Liu et al., 2019).

5 Summary

We highlight gaps within the literature in question answering in the biomedical domain. We outline challenges associated with implementing these systems due to the limitations of current work: lack of annotated data, ambiguity in clinical text and lack of comprehension of question/answer text by models.

We motivate this research in the area of patient QA due to the high volume of medical queries in search engines that are trusted by patients. Our research aims to build upon the strengths of the current state-of-the-art and research new strategies in solving technical challenges to support a patient in retrieving the answers they require with low uncertainty and high confidence.

Acknowledgements

I thank for my supervisors, Dr Sarvnaz Karimi and Dr Zhenchang Xing for providing invaluable insight into the writing of this proposal. This research is supported by the Australian Research Training Program and the CSIRO Postgraduate Scholarship.

References

Ben Abacha and Demner-Fushman. 2016. Recognizing Question Entailment for Medical Question Answering. *American Medical Informatics Association Annual Symposium Proceedings*, 2016:310–318.

Diego Aliod, Menno Zaanen, and Daniel Smith. 2006. Named entity recognition for question answering. In *The Australasian Language Technology Association*, Sydney, Australia.

Flora Amato, Stefano Marrone, Vincenzo Moscato, Gabriele Piantadosi, Antonio Picariello, and Carlo Sansone. 2017. Chatbots meet ehealth: Automatizing healthcare. In *Proceedings of the Workshop on Artificial Intelligence with Application in Health co-located with the 16th International Conference of the Italian Association for Artificial Intelligence*, Bari, Italy.

Alan Aronson and François-Michel Lang. 2010. An overview of Metamap: Historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

Jonas Austerjost, Marc Porr, Noah Riedel, Dominik Geier, Thomas Becker, Thomas Scheper, Daniel Marquard, Patrick Lindner, and Sascha Beutel. 2018. Introducing a virtual assistant to the lab: A voice user interface for the intuitive control of laboratory instruments. *SLAS Technology: Translating Life Sciences Innovation*, 23:476–482.

Stephen Bach, Daniel Rodriguez, Yintao Liu, Chong Luo, Haidong Shao, Cassandra Xia, Souvik Sen, Alexander Ratner, Braden Hancock, Houman Alborzi, Rahul Kuchhal, Christopher Ré, and Rob Malkin. 2018. Snorkel drybell: A case study in deploying weak supervision at industrial scale. *Computing Research Repository*, abs/1812.00417.

Robert Bamler and Stephan Mandt. 2017. Dynamic Word Embeddings. *arXiv e-prints*, page arXiv:1702.08359.

Francesco Barbieri, Luís Marujo, Pradeep Karuturi, William Brendel, and Horacio Saggion. 2018. Exploring emoji usage and prediction through a temporal variation lens. *Computing Research Repository*, abs/1805.00731.

Asma Ben Abacha and Dina Demner-Fushman. 2019. A question-entailment approach to question answering. *Computing Research Repository*, abs/1901.08079.

Nick Bradley, Thomas Fritz, and Reid Holmes. 2018. Context-aware conversational developer assistants. In *Proceedings of the 40th International Conference on Software Engineering*, pages 993–1003, New York, NY, US.

Khyathi Chandu, Aakanksha Naik, Aditya Chandrasekar, Zi Yang, Niloy Gupta, and Eric Nyberg. 2017. Tackling biomedical text summarization: OAQA at BioASQ 5B. In *BioNLP 2017*, pages 58–66, Vancouver, Canada,.

Billy Chiu, Gamal Crichton, Anna Korhonen, and Sampo Pyysalo. 2016. How to train good word embeddings for biomedical NLP. In *Proceedings of the 15th Workshop on Biomedical Natural Language Processing*, pages 166–174, Berlin, Germany.

Katsuki Chousa, Katsuhito Sudoh, and Satoshi Nakamura. 2018. Training neural machine translation using word embedding-based loss. *Computing Research Repository*, abs/1807.11219.

Anthony Cocco, Rachel Zordan, David Taylor, Tracey Weiland, Stuart Dilley, Joyce Kant, Mahesha Dombagolla, Andreas Hendarto, Fiona Lai, and Jennie Hutton. 2018. Dr Google in the ED: searching for online health information by adult emergency department patients. *The Medical Journal of Australia*, 209:342–347.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, Minneapolis, MN.

Sven Festag and Cord Spreckelsen. 2017. *Word Sense Disambiguation of Medical Terms via Recurrent Convolutional Neural Networks*, volume 236. Health Informatics Meets eHealth.

Ary Goldberger, Luis Amaral, Leon Glass, Jeffrey Hausdorff, Plamen Ivanov, Roger Mark, Joseph Mietus, George Moody, Chung-Kang Peng, and Eugene Stanley. 2000. PhysioBank, PhysioToolkit, and PhysioNet: components of a new research resource for complex physiologic signals. *Circulation*, 101(23):E215–220.

Suzanne Graham and John Brookey. 2008. Do patients understand? *The Permanente journal*, 12(3):67–69.

Jiafeng Guo, Yixing Fan, Qingyao Ai, and Bruce Croft. 2017. A deep relevance matching model for ad-hoc retrieval. *Computing Research Repository*, abs/1711.08611.

Yikun Guo, Robert Gaizauskas, Ian Roberts, and George Demetriou. 2006. Identifying personal health information using support vector machines. In *i2b2 Workshop on Challenges in Natural Language Processing for Clinical Data*, Washington, DC, US.

Sepp Hochreiter and Jrgen Schmidhuber. 1997. Long short-term memory. *Neural Computing*, 9(8):1735–1780.

Jeremy Howard and Sebastian Ruder. 2018. Universal language model fine-tuning for text classification. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 328–339, Melbourne, Australia.

Kai Hui, Andrew Yates, Klaus Berberich, and Gerard de Melo. 2017. A position-aware deep model for relevance matching in information retrieval. *Computing Research Repository*, abs/1704.03940.

Lawrence Hunter and Bretonnel Cohen. 2006. Biomedical language processing: what's beyond pubmed? *Molecular Cell*, 21(5):589–594.

Alistair Johnson, Tom Pollard, Lu Shen, Li-wei Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Anthony Celi, and Roger Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3:160035.

Aniket Kadam, Shashank Joshi, Sachin Shinde, and Sampat Medhane. 2015. Notice of retraction question answering search engine short review and roadmap to future qa search engine. In *International Conference on Electrical, Electronics, Signals, Communication and Optimization*, pages 1–8, Visakhapatnam, India.

Alex Kendall and Yarin Gal. 2017. What uncertainties do we need in bayesian deep learning for computer vision? *Computing Research Repository*, abs/1703.04977.

Tomás Kociský, Jonathan Schwarz, Phil Blunsom, Chris Dyer, Karl Hermann, Gábor Melis, and Edward Grefenstette. 2017. The narrativeqa reading comprehension challenge. *Computing Research Repository*, abs/1712.07040.

Ida Korfage, Harry Koning, Monique Roobol, Fritz Schrder, and Marie-Louise Essink-Bot. 2006. Prostate cancer diagnosis: The impact on patients mental health. *European Journal of Cancer*, 42(2):165 – 170.

Yann LeCun, Patrick Haffner, Léon Bottou, and Yoshua Bengio. 1999. Object recognition with gradient-based learning. In *Shape, Contour and Grouping in Computer Vision*, page 319, London, UK.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. BioBERT: A pre-trained biomedical language representation model for biomedical text mining. *arXiv e-prints*, page arXiv:1901.08746.

Minsuk Lee, James Cimino, Hai Zhu, Carl Sable, Vijay Shanker, John Ely, and Hong Yu. 2006. Beyond information retrieval–medical question answering. *American Medical Informatics Association Annual Symposium Proceedings*, 2006:469–473.

Sweta Lende and Mukesh Raghuwanshi. 2016. Question answering system on education acts using nlp techniques. In *World Conference on Futuristic Trends in Research and Innovation for Social Welfare (Startup Conclave)*, pages 1–6, Coimbatore, India.

Chin-Yew Lin. 2004. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the Association for Computational Linguistics Workshop*, pages 74–81, Barcelona, Spain.

Xiaodong Liu, Pengcheng He, Weizhu Chen, and Jianfeng Gao. 2019. Improving multi-task deep neural networks via knowledge distillation for natural language understanding. *Computing Research Repository*, arXiv:1904.0948.

Nidhi Malik, Aditi Sharan, and Payal Biswas. 2013. Domain knowledge enriched framework for restricted domain question answering system. In *IEEE International Conference on Computational Intelligence and Computing Research*, pages 1–7, Madurai, Tamilnadu, India.

Yuqing Mao, Chih-Hsuan Wei, and Zhiyong Lu. 2014. NCBI at the 2014 BioASQ challenge task: Large-scale biomedical semantic indexing and question answering. In *Conference and Labs of the Evaluation Forum*, Sheffield, UK.

Bryan McCann, Nitish Shirish Keskar, Caiming Xiong, and Richard Socher. 2018. The natural language decathlon: Multitask learning as question answering. *Computing Research Repository*, abs/1806.08730.

Ryan McDonald, Georgios-Ioannis Brokos, and Ion Androutsopoulos. 2018. Deep relevance ranking using enhanced document-query interactions. *Computing Research Repository*, abs/1809.01682.

Rashmi Mishra, Jiantao Bian, Marcelo Fiszman, Charlene Weir, Siddhartha Jonnalagadda, Javed Mostafa, and Guilherme Del Fiol. 2014. Text summarization in the biomedical domain: a systematic review of recent research. *Journal of Biomedical Informatics*, 52:457–467.

Diego Molla and Jos Gonzlez. 2007. Question answering in restricted domains: An overview. *Computational Linguistics*, 33:41–61.

Anastasios Nentidis, Konstantinos Bougiatiotis, Anastasia Krithara, Georgios Paliouras, and Ioannis Kakadiaris. 2017. Results of the fifth edition of the bioasq challenge. In *Biomedical Natural Language Processing*, pages 48–57, Vancouver, Canada.

Anastasios Nentidis, Anastasia Krithara, Konstantinos Bougiatiotis, Georgios Paliouras, and Ioannis Kakadiaris. 2018. Results of the sixth edition of the BioASQ challenge. In *Proceedings of the 6th BioASQ Workshop A challenge on large-scale biomedical semantic indexing and question answering*, pages 1–10, Brussels, Belgium.

Karolina Owczarzak and Hoa Dang. 2009. Evaluation of automatic summaries: Metrics under varying data conditions. In *Proceedings of the Workshop on Language Generation and Summarisation*, pages 23–30, Stroudsburg, PA, US.

Anusri Pampari, Preethi Raghavan, Jennifer Liang, and Jian Peng. 2018. emrqa: A large corpus for question answering on electronic medical records. *Computing Research Repository*, abs/1809.00732.

Harris Papadopoulos. 2008. *Inductive Conformal Prediction: Theory and Application to Neural Networks*.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of 40th Annual Meeting of the Association for Computational Linguistics*, pages 311–318, Philadelphia, Pennsylvania, US.

Rajul Parikh, Annie Mathai, Shefali Parikh, Chandra Sekhar, and Ravi Thomas. 2008. Understanding and using sensitivity, specificity and predictive values. *Indian Journal of Ophthalmology*, 56(1):45–50.

Shengwen Peng, Ronghui You, Zhikai Xie, Beichen Wang, Yanchun Zhang, and Shanfeng Zhu. 2015. The fudan participation in the 2015 bioasq challenge: Large-scale biomedical semantic indexing and question answering. In *Conference and Labs of the Evaluation Forum 2015: Conference and Labs of the Evaluation Forum Experimental IR meets Multilinguality, Multimodality and Interaction*, volume 1391, Toulouse, France.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *Computing Research Repository*, abs/1802.05365.

Sameerchand Pudaruth, Kajal Boodhoo, and Lushika Goolbudun. 2016. An intelligent question answering system for ICT. In *2016 International Conference on Electrical, Electronics, and Optimization Techniques*, pages 2895–2899, Paralakhemundi, Odisha, India.

Vasin Punyakanok, Dan Roth, and Wen tau Yih. 2004. Mapping dependencies trees: An application to question answering. In *In Proceedings of the 8th International Symposium on Artificial Intelligence and Mathematics, Fort*, Fort Lauderdale, Flordia.

Ye Qi, Devendra Sachan, Matthieu Felix, Sarguna Padmanabhan, and Graham Neubig. 2018. When and why are pre-trained word embeddings useful for neural machine translation? *Computing Research Repository*, abs/1804.06323.

Minghui Qiu, Feng-Lin Li, Siyu Wang, Xing Gao, Yan Chen, Weipeng Zhao, Haiqing Chen, Jun Huang, and Wei deep-multi-task-learning Chu. 2017. AliMe chat: A sequence to sequence and rerank based chatbot engine. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 498–503, Vancouver, Canada.

Pranav Rajpurkar, Jian Zhang, Konstantin Lopyrev, and Percy Liang. 2016. Squad: 100, 000+ questions for machine comprehension of text. *Computing Research Repository*, abs/1606.05250.

Ashwin Ram, Rohit Prasad, Chandra Khatri, Anu Venkatesh, Raefer Gabriel, Qing Liu, Jeff Nunn, Behnam Hedayatnia, Ming Cheng, Ashish Nagar, Eric King, Kate Bland, Amanda Wartick, Yi Pan, Han Song, Sk Jayadevan, Gene Hwang, and Art Pettigrue. 2018. Conversational AI: the science behind the alexa prize. *Computing Research Repository*, abs/1801.03604.

Tony Russell-Rose and Jon Chamberlain. 2017. Expert search strategies: The information retrieval practices of healthcare information professionals. *JMIR Medical Informatics*, 5(4):e33.

Sheng Shen, Yaliang Li, Nan Du, Xian Wu, Yusheng Xie, Shen Ge, Tao Yang, Kai Wang, Xingzheng Liang, and Wei Fan. 2018. On the generation of medical question-answer pairs. *Computing Research Repository*, abs/1811.00681.

Tian Shi, Yaser Keneshloo, Naren Ramakrishnan, and Chandan Reddy. 2018. Neural abstractive text summarization with sequence-to-sequence models. *Computing Research Repository*, abs/1812.02303.

Aditya Siddhant and Zachary Lipton. 2018. Deep bayesian active learning for natural language processing: Results of a large-scale empirical study. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2904–2909, Brussels, Belgium.

Marco Soares and Fernando Parreiras. 2018. A literature review on question answering techniques, paradigms and systems. *Journal of King Saud University - Computer and Information Sciences*.

Ke Sun, Jun Wang, Alexandros Kalousis, and Stephane Marchand-Maillet. 2015. Space-time local embeddings. In *Advances in Neural Information Processing Systems 28*, pages 100–108.

Hanna Suominen, Liadh Kelly, Lorraine Goeuriot, Aurélie Névéol, Lionel Ramadier, Aude Robert, Evangelos Kanoulas, Rene Spijker, Leif Azzopardi, Dan Li, Jimmy, João Palotti, and Guido Zuccon. 2018. Overview of the conference and labs of the evaluation forum ehealth evaluation lab 2018. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 286–301, Avignon, France.

Alon Talmor, Jonathan Herzig, Nicholas Lourie, and Jonathan Berant. 2018. Commonsenseqa: A question answering challenge targeting commonsense knowledge. *Computing Research Repository*, abs/1811.00937.

Rafael Terol, Patricio Martinez-Barco, and Manuel Palomar. 2007. A knowledge based method for the medical question answering problem. *Computers in Biology and Medicine*, 37(10):1511–1521.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. *Computing Research Repository*, abs/1706.03762.

José Vicedo and Antonio Ferrández. 2000. Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, pages 555–562, Hong Kong.

Lance Vine, Mahnoosh Kholghi, Guido Zuccon, Laurianne Sitbon, and Anthony Nguyen. 2015. Analysis of word embeddings and sequence features for clinical information extraction. In *Proceedings of the Australasian Language Technology Association Workshop 2015*, pages 21–30, Parramatta, Australia.

Dirk Weienborn, George Tsatsaronis, and Michael Schroeder. 2013. Answering factoid questions in the biomedical domain. In *CEUR Workshop Proceedings*, volume 1094, Valencia, Spain.

Yonghui Wu, Joshua Denny, Trent Rosenbloom, Randolph Miller, Dario Giuse, and Hua Xu. 2012. A comparative study of current clinical natural language processing systems on handling abbreviations in discharge summaries. *American Medical Informatics Association Annual Symposium Proceedings*, 2012:997–1003.

Yonghui Wu, Joshua Denny, Rosenbloom Trent, Randolph Miller, Dario Giuse, Lulu Wang, Carmelo Blanquicett, Ergin Soysal, Jun Xu, and Hua Xu. 2017. A long journey to short abbreviations: developing an open-source framework for clinical abbreviation recognition and disambiguation (CARD). *Journal of the American Medical Informatics Association*, 24(e1):e79–e86.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, Jeff Klingner, Apurva Shah, Melvin Johnson, Xiaobing Liu, Lukasz Kaiser, Stephan Gouws, Yoshikiyo Kato, Taku Kudo, Hideto Kazawa, Keith Stevens, George Kurian, Nishant Patil, Wei Wang, Cliff Young, Jason Smith, Jason Riesa, Alex Rudnick, Oriol Vinyals, Greg Corrado, Macduff Hughes, and Jeffrey Dean. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *Computing Research Repository*, abs/1609.08144.

Ramin Yaghoubzadeh and Stefan Kopp. 2012. Toward a virtual assistant for vulnerable users: Designing careful interaction. In *Proceedings of the 1st Workshop on Speech and Multimodal Interaction in Assistive Environments*, pages 13–17, Jeju, Republic of Korea. Association for Computational Linguistics.

Erjia Yan and Yongjun Zhu. 2018. Tracking word semantic change in biomedical literature. *International Journal of Medical Informatics*, 109:76 – 86.

Zhao Yan, Nan Duan, Junwei Bao, Peng Chen, Ming Zhou, Zhoujun Li, and Jianshe Zhou. 2016. DocChat: An information retrieval approach for chatbot engines using unstructured documents. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 516–525, Berlin, Germany. Association for Computational Linguistics.

Wenpeng Yin, Hinrich Schütze, Bing Xiang, and Bowen Zhou. 2015. ABCNN: attention-based convolutional neural network for modeling sentence pairs. *Computing Research Repository*, abs/1512.05193.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2018. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. *Computing Research Repository*, abs/1812.06081.

Zhe Zhao, Tao Liu, Shen Li, Bofang Li, and Xiaoyong Du. 2017. Ngram2vec: Learning improved word representations from ngram co-occurrence statistics. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 244–253, Copenhagen, Denmark.

Barret Zoph and Quoc Le. 2016. Neural architecture search with reinforcement learning. *Computing Research Repository*, abs/1611.01578.