Recognizing Question Entailment for Medical Question Answering

Asma Ben Abacha, PhD & Dina Demner-Fushman, MD, PhD
U.S. National Library of Medicine, Bethesda, MD.

Abstract

With the increasing heterogeneity and specialization of medical texts, automated question answering is becoming more and more challenging. In this context, answering a given medical question by retrieving similar questions that are already answered by human experts seems to be a promising solution. In this paper, we propose a new approach for the detection of similar questions based on Recognizing Question Entailment (RQE). In particular, we consider Frequently Asked Question (FAQs) as a valuable and widespread source of information. Our final goal is to automatically provide an existing answer if FAQ similar to a consumer health question exists. We evaluate our approach using consumer health questions received by the National Library of Medicine and FAQs collected from NIH websites. Our first results are promising and suggest the feasibility of our approach as a valuable complement to classic question answering approaches.

1 Introduction

Consumer health queries are increasing at a high rate on the World Wide Web. Recent studies from the Pew Research Center show that 59% of U.S. adults searched for health information online in 2013[1]. In the same study, we find that 35% of U.S. adults attempted to figure out what medical condition they or someone else might have by searching online resources. Consumer health questions cover a wide range of health-related topics asked generally by non-expert persons, patients or professionals. Typical information sought by consumers can be, for instance, information about a disease (e.g. " *My daughter was recently diagnosed with a CTNNB1 mutation and there are only 4 published cases. Can you tell me where I may be able to find more information*"), treatments (e.g. " *Do you have information on the treatment of Parkinson's disease with the use of amino acids*") or more specific questions about drugs (e.g. " *is it safe to take diclofenac when taking lisinopril or aleve or extrastrength Tylenol?* ")[2].

Many websites offer online doctor consultation services in virtual hospitals (e.g. icliniccare.com, www.icliniq.com, www.jeevom.com, www.evaidya.com, www.doctorvista.com) and propose to answer medical questions, give advice or second opinions from thousands of doctors and specialists. These services attract more and more internet users seeking easy and quick answers, as well as free consultations and privacy. Many of these resources publish freely the submitted questions and doctors' answers. New health-dedicated forums and websites continue emerging each year; and major question-answering websites such as Quora[3] or Yahoo! Answers[4] include important and growing health sections.

With this multitude of information, duplicate questions are becoming more frequent, and finding the most appropriate answers becomes problematic. This issue is important for question-answering platforms as it complicates the retrieval of all information relevant to the same topic, particularly when questions similar in essence are expressed differently. Finding similar questions is also important for the users as is implied by the 2013 survey, in which 16% of the users tried to find others with the same health concerns.

Efficient automatic approaches are therefore required to detect similar questions both at search time and at question submission time. In this paper we propose a new approach for the detection of similar questions based on Recognizing Textual Entailment (RTE). RTE is an important component in Natural Language Understanding and in Question Answering. Harabagiu and Hickl argue that RTE can enable question answering systems to identify correct answers with greater precision than keyword or pattern based methods [1].

We particularly tackle the detection of a most similar Frequently Asked Question (FAQ) to a given consumer question. FAQs are both a valuable and widespread source of information. Many trusted sources, such as NIH institutes and

[1] `http://www.pewinternet.org/2013/01/15/health-online-2013/`

[2] Example questions received by the U.S. National Library of Medicine.

[3] `www.quora.com` As of March 1 2016, 7,176,197 questions were posted to Quora.

[4] `answers.yahoo.com`

centers, provide answers to frequently asked questions organized by topics. NIH provides FAQs for many health related problems such as Rare Diseases (Frequently Asked Questions About Rare Diseases[5], or Alzheimer[6]).

To test our approach in a use case, we obtained consumer health questions received by the National Library of Medicine (NLM). NLM receives hundreds of requests per day, e.g., from October 2013 to September 2014, NLM received a total of 102,622 requests, including many consumer health questions. Our final goal is to automatically provide an existing answer if an FAQ similar to a consumer health question received by NLM exists. If we can reliably identify similar questions, retrieving the corresponding answers is relatively straightforward. For example, this approach to answering health related questions is taken in the SimQ system [2].

In this paper, RTE is applied to find a frequently asked question similar to a consumer health questions in order to answer consumer health questions with the answers given to similar FAQs. As far as we know, RTE from medical questions has not been studied before for Question Answering. Our contributions can therefore be summarized in three points:

- We address and define the problem of RTE in medical questions for Question Answering.
- We construct automatically an RTE training corpus of medical questions and we study the impact of varying the size and nature of the training examples.
- We use different features for the RTE task in medical questions including classical similarity measures and semantic features related to the medical domain.

The remainder of the paper is organized as follows: Section 2 presents the background. Section 3 presents our approach for RTE in medical questions developed to retrieve existing FAQs and answers. In section 4, we present our methods to construct automatically an RTE training corpus, and to construct semi-automatically an RTE test corpus for medical questions. Our experiments and results are detailed and discussed in sections 5 and 6.

2 Background

Recognizing Textual Entailment (RTE) has been addressed by numerous works in the literature and in the framework of the PASCAL challenge [7]. Dagan et al. [3] present the RTE task and give an overview of the research efforts in this area. A detailed survey of RTE approaches is also presented in [4].

Several machine learning (ML) methods have been explored for RTE using different features, e.g., similarity measures [5, 6, 7]. Some efforts focused on the training corpora as they are an important factor for an efficient supervised learning system. Zanzotto and Pennacchiotti [8] proposed a method to expand the existing textual entailment corpora. They extracted from Wikipedia a large set of textual entailment pairs and used a semi-supervised machine learning method to make the extracted dataset homogeneous with the existing corpora. Other efforts tackled automatic generation of training corpora for a specific language. For instance, Marzelou et al. [9] proposed a method to create a Greek Textual Entailment Corpus that can be exploited for training or evaluating a system for RTE from Greek texts.

While textual entailment in open-domain has been extensively addressed in the literature, RTE has been less addressed for more restricted and specialized fields such as the medical domain. Adler et al. [10] presented a text exploration system, in which search results in the health-care domain can be navigated at propositional level according to textual entailment relation. Ben Abacha et al. [11] proposed a supervised learning method to RTE from medical texts. For the same purpose, they also proposed an automatic method for construction of training corpora from MEDLINE abstracts and compared the results obtained from the open-domain model, derived from the PASCAL training corpus, and the medical-domain model, derived from the automatically-constructed corpus. Their experiments showed the benefits of domain-related models and automatic corpus construction.

[5] `http:/www.genome.gov/27531963` National Human Genome Research Institute, NIH.
[6] `http://nihseniorhealth.gov/alzheimersdisease/faq/faqlist.html`
[7] `http://www.nist.gov/tac/2011/RTE/`

One of the earliest question answering systems based on finding similar questions and re-using the existing answers was FAQ FINDER [12]. Some of the underlying assumptions about FAQs were that the information needed to determine the relevance of a (Question/Answer) QA pair can be found within the QA pair and that the question half of the QA pair is the most relevant for determining the match to a users question. Not surprisingly, much of the subsequent work was dedicated to question matching. Jeon et al.[13] showed that a retrieval model based on translation probabilities learned from a question and answer archive can recognize semantically similar questions. Duan et al. [14] proposed a language modeling using question topic and question focus for question search. An alternative approach, is to return a ranked list of QA pairs in response to a user's question, treating finding an answer as a fielded search task, where the user's question is treated as a query, and the items to be returned are QA pairs [15].

More recent research efforts have focused on retrieving similar consumer health questions. Wang et al. [16] developed a syntactic tree matching method to find similar questions for 0.5 million QA pairs from the Healthcare domain in Yahoo! Answers. SimQ [2] aims to retrieve similar web-based consumer health questions. The system uses syntactic and semantic features and reaches a precision of 72.2% and a recall of 78.0%. SimQ is used to complement the existing Q&A services of Netwellness[8] and allows reducing response delay by instantly providing closely related questions and answers.

3 Recognizing Question Entailment (RQE)

A. Problem Definition

Textual Entailment (TE) is a directional relation between two text snippets called *text* (T) and *hypothesis* (H), expressing the fact that the meaning of T is contained in the meaning of H [3]. In a similar definition, the first PASCAL Recognizing Textual Entailment Challenge[9] (2004-2005) defined the task of Recognizing Textual Entailment (RTE) as deciding, given two text fragments, whether or not the meaning of one text (H) can be inferred (entailed) from the other one (T) [17].

In our case, the (T, H) pairs refer to pairs of questions (Q1, Q2). Our goal in to retrieve answers to Q1 by retrieving entailed questions Q2 that have associated answers. Groenendijk and Stokhof (1984) define an entailment relation between two questions Q1, Q2 if every proposition giving an answer to Q1 is also giving an answer to Q2 [18]. Roberts (1996) called the first question Q1 as a superquestion and Q2 a subquestion (if we answer enough subquestions, we have the answer to the superquestion) [19]. In accordance with these definitions, we define question entailment as follows:

Question Entailment. Question A entails Question B if every answer to B is also a correct answer to A exactly (cf. Example 1) or partially (cf. Example 2).

Consumer health questions often contain heterogeneous information, e.g. information about the patient's history. Our rule for ignoring additional information in the question is based on the assumption that it is not required to retrieve a correct answer (cf. Example 1, in which the fact that the patient has only central vision will not help retrieving information related to treatment of the disease). Here are two examples from our datasets:

- **Example 1:**
- A1 (consumer health question): Hi I have retinitis pigmentosa for 3years. Im suffering from this disease. Please intoduce me any way to treat mg eyes such as stem cell ....I am 25 years old and I have only central vision. Please help me. Thank you
- B1 (FAQ): Are there treatments for RP?
- A1 $\rightarrow$ B1
- **Example 2:**

[8] http://netwellness.org/

[9] http://pascallin.ecs.soton.ac.uk/Challenges/RTE/

- A2 (consumer health question): Can sepsis be prevented? Can someone get this from a hospital?
- B2 (FAQ): Who gets sepsis?
- A2 → B2

A2 includes 2 questions about prevention and susceptibility. An answer to B2 (about the susceptibility to Sepsis) is considered as a partially correct answer to A2. In this case, we consider that A2 implies B2.

B. Learning Method for RQE

We propose a supervised machine learning approach to determine whether or not a question Q2 can be inferred from a question Q1. In order to extract relevant features, we first remove stop words and perform word stemming using the Porter algorithm [20] for all (Q1,Q2) training pairs.

Lexical Features:

We compute different similarity measures between the pre-processed questions and use their values as features:

- Word Overlap: we compute the word overlap as the proportion of words that appear in both Q1 and Q1 and normalize by the length of Q1.
- Bigram: we compute the bigram similarity between Q1 and Q2 as the total number of matched bigrams in (Q1,Q2) pair normalized by the number of Q1 bigrams.
- Best similarity value: the maximum similarity between five similarity measures: Levenshtein, Bigram, Jaccard, Cosine and Word Overlap.

Semantic Features:

Negation: we use NegEx [21] for identifying negation scope in Q1 and Q2.

Medical entities: we annotate all (Q1,Q2) pairs with medical entities using two supervised systems. The first system uses a CRF classifier trained on the i2b2 corpus [22] to recognize medical entities of 3 types: Problem, Treatment and Test [23]. These three medical categories are the most frequent and important categories in consumer questions. The second system is a meta-classifier [24] trained on four corpora: two clinical texts corpora i2b2 and SemEval [25] and two scientific abstracts corpora NCBI [26] and Berkeley [27] to recognize medical problems. In many cases, the focus of the consumer health question and the FAQ is a medical problem (e.g. questions about treatment, symptoms or medical exam). Starting from the obtained annotations, we generate the following semantic features for each question pair (Q1,Q2):

- Number of medical entities in Q1. Number of medical entities in Q2.
- Number of medical problems in Q1. Number of medical problems in Q2
- Number of medical entities in common between Q1 and Q2.
- Number of medical problems in common between Q1 and Q2.
- Common medical problem: binary feature indicating whether or not at least one medical problem from Q1 is mentioned in Q2.

4 Data

A. Automatic Construction of Training Data

We used the NLM collection of 4,655 clinical questions asked by family doctors [28] to construct our training corpus for RQE. An extract from this collection is presented below (Clinical question with NLM ID: NQ003094):

- <original_question> 6 and 1/2-year-old girl. What's causing her ear pain? She has a normal ear on exam and normal tympanogram. Is it just eustachian tube dysfunction? </original_question>
- <short_question> What is causing the ear pain in this child with a normal ear exam? </short_question>
- <general_question> What would cause ear pain in a child with a normal exam? </general_question>
- <content>History</content>
- <keyword>Earache</keyword>
- <keyword>Diagnosis</keyword>

The content field includes Device, Diagnosis, Epidemiology, Etiology, History, Management, Pharmacological, Physical Finding, Procedure, Prognosis, Test, Treatment & Prevention.

To obtain positive QE examples, we use the original form and the short form of the question as expressed in the collection.

- original question → short form (4,655 positive examples)

To obtain the final corpus, we studied five different construction methods by varying the number and type of negative pairs, and we analyzed the impact of each construction method on the results. Each of these methods led to different training sets, described below:

- Training set 1: 8,588 training pairs, containing 54.2% positive pairs. The remaining pairs (3,933) are negative examples collected by associating a random short form having at least one common keyword and at least one different keyword for each original question.
- Training set 2: 17,898 training pairs, containing 26 % positive pairs. The remaining pairs (13,243) are negative examples constructed as follows:
  - original question → short form of random question having at least 1 different keyword and at least 1 common keyword
  - original question → short form of random question having at least 1 different keyword and at least 1 common content
  - original question → short form of random question having at least 1 different keyword
- Training set 3: 13,918 training pairs containing 33.44 % positive pairs. Two types of negative examples are considered in this set:
  - original question → short form of random question having (i) at least one different keyword and at least one common keyword or (ii) at least one different <content> tag and one equal <content> tag (cf. example question above)
  - original question → short form of random question having (i) at least one different keyword and at least one equal <content> tag or (ii) at least one different <content> tag and at least one equal keyword.
- Training set 4: 18,573 training pairs with 25 % positive pairs. Three types of negative examples (13,918 negative pairs):
  - Same 2 types of negative pairs as in training set 3.
  - original question → short form of random question having (at least one different CONTENT) OR (at least one different keyword)
- Training set 5: 9,310 training pairs with 50 % positive pairs. One type of negative examples regrouping the 2 types of Training set 3 (4,655 negative pairs):

- original question → short form of random question having (i) at least one different keyword and at least one common keyword or (ii) at least one different <content> tag and at least one equal <content> tag, or (iii) at least one different keyword and at least one equal <content> tag, or (iv) at least one different <content> tag and at least one equal keyword. Duplicated questions were removed.

The four examples below are extracted from our training corpora to better show the intuition behind these different construction procedures.

- <pair id="29" type="originalQ-shortRandQ" value="**negative**">
  <t> What is the cause and treatment of this old man's stomatitis? </t>
  <h> What would cause a patient to have a low thyroxine and a high thyroid stimulating hormone when the patient is on Synthroid chronically? </h> </pair>
- <pair id="37" type="originalQ-shortQ" value="**positive**">
  <t> Is melatonin good for anything? I don't know anything about melatonin. I need to know the dose. </t>
  <h> Is melatonin good for anything? What is the dose? </h> </pair>
- <pair id="4358" type="originalQ-shortRandQ" value="**negative**">
  <t> How should you work up someone who tried to give blood and has positive Hepatitis B core antibody? </t>
  <h> Why did the blood pressure go up with addition of Vasotec? (Was already on Procardia.) </h> </pair>
- <pair id="10366" type="originalQ-shortQ" value="**positive**">
  <t> Young woman with acute Fifth disease (son also has it). She is not pregnant. She has increased swelling of her hands and her arms go numb and she has trouble sleeping. Wants to know what to do. </t>
  <h> What is the treatment for Fifth disease? </h> </pair>

We evaluate the constructed training corpora in terms of Precision, Recall and F-measure in section 5.

B. Semi-automatic Construction of Test Data

For test pairs, we collected two types of test data: (i) pairs of manually validated questions from the NLM collections and (ii) pairs of questions including FAQs retrieved online with a manual search of NIH websites. We constructed the two parts of our test corpus using the following methods:

We constructed the first part using two collections of (i) 300 consumer health questions annotated with the focus of the question and (ii) 349 FAQs from NIH websites, in two steps:

1. Extracting the question pairs that have the same focus (medical problem).
2. Manual annotation and selection of 70 positive and negative pairs.

Here are two examples from the first part of our test corpus:

- <pair id="6" type="Part1" value="**TRUE**">
  <t> No. hi my name is _NAME_ I'm currently working with Friends Community Center in Hollywood California and I was wondering I came across some of you healthy tip fliers for HIV/Aids treatment .at the moment we have a study going on that helps HIV positive transgender women into HIV quality care .so it would be great to have some more information on HIV/Aids treatment </t>
  <h> How is HIV/AIDS treated? </h> </pair>
- <pair id="66" type="Part1" value="**FALSE**">
  <t>recovery after stroke?. what is the pattern of recovery after stroke? </t>
  <h> Is there any treatment for Stroke and Atrial Fibrillation? </h> </pair>

For the second part, we searched online for FAQs for 116 from the remaining consumer health questions from NIH websites to construct our positive pairs. For negative pairs, we selected for each consumer health question, a random FAQ from the list of 116 found online. Here are two examples from the second part of our test corpus:

- <pair id="79" type="Part2" value="**TRUE**">
  <t> Help for my diagnose. I have been diagnosed with SCA3. I was wondering if MedlinePlus is able to help me with resources that I may need on my journey through this disease? If not, can you help me find an organization or association that can help me. </t>
  <h> Where can I find additional information about SCA3? </h> </pair>

- <pair id=”302” type=”Part2” value=”**FALSE**”>
  <t>I will love to try diet I need. To lose weight </t>
  <h> What Are Asbestos-Related Lung Diseases? How Are Asbestos-Related Lung Diseases Diagnosed? </h> </pair>

Our final test corpus contains 302 pairs of questions consisting of 173 negative pairs and 129 positive pairs.

5 Evaluation

In this section, we evaluate the corpus construction methods as well as several algorithms to recognize textual entailment between questions.

Table 1 presents the results in terms of Precision (P), Recall (R) and F-measure (F), obtained using different training corpora. We tested different configurations (i) by adding new types of negative pairs (e.g. training set 3 vs. training set 4) to evaluate the impact of the new example types and (ii) by regrouping types of pairs to evaluate the impact of reducing the number of negative examples (e.g. training set 3 vs. training set 5).

Training Data	P	R	F
Set 1	**75.0**	**75.2**	**75.0**
Set 2	72.5	71.9	70.7
Set 3	**73.4**	73.5	73.2
Set 4	71.6	71.2	70.1
Set 5	72.7	72.8	72.4

Table 1: Results of a SVM classifier trained on five datasets having different sizes and types.

Table 2 presents the obtained results using four statistical learning algorithms that are usually used for RTE (SVM, Logistic Regression, Naive Bayes and J48). The best results are obtained using the SVM classifier (75% F-measure). The Logistic Regression classifier gives a slightly lower F-measure on our dataset (74.7%).

Algorithm	P	R	F
SVM	**75.0**	**75.2**	**75.0**
Logistic Regression	74.7	74.8	74.7
Naive Bayes	73.1	72.5	71.5
J48	70.9	70.2	70.3

Table 2: Results of different classifiers trained on Set1 of training pairs.

6 Discussion

By testing the impact of different automatic corpus construction methods to train RQE classifiers we found that increasing the number of negative examples does not always improve the results. The nature and size of the training corpus obtained by our best automatic construction method (8,588 question pairs with 54.2 % of positive pairs) provided a good start for building a system for question entailment recognition that can be used as an important support to find additional answers that complement classical question-answering results.

While we obtained relatively good results in our evaluation, several challenges have still to be tackled to enhance further the performance of similarity-based approaches. If we consider the example of questions Q1 and Q2 below, they are not linked by an an entailment relation even though they share three different keywords (cause, dry, mouth).

- *Q1 (Consumer health question)*: treatment for *dry mouth caused* by necessary medicine. My provider can't help (I asked.) I am intolerant of all the sugar alcohols such as maltilol, sorbitol, xylitol, etc. and need something for dry mouth caused by med which I have to take. Biotene products help for only about two minutes.

- *Q2 (FAQ)*: What *causes dry mouth*?

- *Entailment*: false.

More generally, relying on number of shared keywords and medical entities provided a high recall upper bound for the recognition of question entailment, however, it did not reach very high precision due to the lack of advanced features such as the identification of the answer type (e.g. treatment in Q1) and the main semantic relations (e.g. *causes* in Q2).

7 Conclusion

In this paper, we described our approach for recognizing question entailment (RQE) in order to answer new questions using existing Question-Answer pairs. We presented an automatic method for the construction of training corpora for RQE and a semi-automatic method for the construction of a test corpus for medical questions. Our experiments confirm the feasibility of medical question entailment even with a small set of features. In future work, we plan to extend our test corpus and to include more semantic features. We also plan to study the adaptation of our training corpus consisting of doctors' questions by including a small number of consumer health questions (e.g. questions asked by non experts, patients) and manually collected FAQs.

References

[1] Harabagiu S, Hickl A. Methods for Using Textual Entailment in Open-domain Question Answering. In: Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics. Association for Computational Linguistics; 2006. p. 905–912.

[2] Luo J, Zhang GQ, Wentz S, Cui L, Xu R. SimQ: Real-Time Retrieval of Similar Consumer Health Questions;.

[3] Dagan I, Roth D, Sammons M, Zanzotto FM. Recognizing Textual Entailment: Models and Applications. Synthesis Lectures on Human Language Technologies. Morgan & Claypool Publishers; 2013.

[4] Androutsopoulos I, Malakasiotis P. A Survey of Paraphrasing and Textual Entailment Methods. J Artif Int Res. 2010 May;38(1):135–187.

[5] Kozareva Z, Montoyo A. MLEnt: The Machine Learning Entailment System of the University of Alicante. In: In Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment, Pages 17-20; 2006. .

[6] Malakasiotis P, Androutsopoulos I. Learning Textual Entailment Using SVMs and String Similarity Measures. In: Proceedings of the ACL-PASCAL Workshop on Textual Entailment and Paraphrasing. Association for Computational Linguistics; 2007. p. 42–47.

[7] Zanzotto FM, Pennacchiotti M, Moschitti A. A machine learning approach to textual entailment recognition. Natural Language Engineering. 2009;15(4):551–582.

[8] Zanzotto FM, Pennacchiotti M. Expanding textual entailment corpora from Wikipedia using co-training. In: In Proc of the 2nd Coling Workshop on The Peoples Web Meets NLP: Collaboratively Constructed Semantic Resources; 2010. p. 28–36.

[9] Marzelou E, Zourari M, Giouli V, Piperidis S. Building a Greek corpus of Textual Entailment. In: Proceedings of the 6th Language Resources and Evaluation Conference. Marrakech, Morocco; 2008. p. 1680–1686.

[10] Adler M, Berant J, Dagan I. Entailment-based Text Exploration with Application to the Health-care Domain. In: Proceedings of the ACL 2012 System Demonstrations. Jeju Island, Korea: Association for Computational Linguistics; 2012. p. 79–84. Available from: http://www.aclweb.org/anthology/P12-3014.

[11] Ben Abacha A, Dinh D, Mrabet Y. Semantic Analysis and Automatic Corpus Construction for Entailment Recognition in Medical Texts. In: Artificial Intelligence in Medicine - 15th Conference on Artificial Intelligence in Medicine, AIME 2015, Pavia, Italy, June 17-20, 2015. Proceedings; 2015. p. 238–242. Available from: http://dx.doi.org/10.1007/978-3-319-19551-3_31.

[12] Burke RD, Hammond KJ, Kulyukin V, Lytinen SL, Tomuro N, Schoenberg S. Question answering from frequently asked question files: Experiences with the faq finder system. AI magazine. 1997;18(2):57.

[13] Jeon J, Croft WB, Lee JH. Finding Similar Questions in Large Question and Answer Archives. In: Proceedings of the 14th ACM International Conference on Information and Knowledge Management. CIKM '05. New York, NY, USA: ACM; 2005. p. 84–90. Available from: http://doi.acm.org/10.1145/1099554.1099572.

[14] Duan H, Cao Y, Lin CY, Yu Y. Searching Questions by Identifying Question Topic and Question Focus.; 2008. .

[15] Jijkoun V, de Rijke M. Retrieving answers from frequently asked questions pages on the web. In: Proceedings of the 14th ACM international conference on Information and knowledge management. ACM; 2005. p. 76–83.

[16] Wang K, Ming Z, Chua TS. A syntactic tree matching approach to finding similar questions in community-based qa services. In: Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval. ACM; 2009. p. 187–194.

[17] Dagan I, Glickman O, Magnini B. The PASCAL Recognising Textual Entailment Challenge. In: Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment; 2005. Available from: http://www.cs.biu.ac.il/~glikmao/rte05/.

[18] Groenendijk J, Stokhof M. Studies on the Semantics of Questions and the Pragmatics of Answers. University of Amsterdam, Amsterdam; 1984. Available from: http://dare.uva.nl/document/2/27444.

[19] Roberts C. Information Structure in Discourse: Towards an Integrated Formal Theory of Pragmatics. Ohio State University Working Papers in Linguistics. 1996;49.

[20] Porter M. An Algorithm for Suffix Stripping. Program. 1980;14(3):130–137.

[21] Chapman WW, Bridewell W, Hanbury P, Cooper GF, Buchanan BG. A Simple Algorithm for Identifying Negated Findings and Diseases in Discharge Summaries. Journal of Biomedical Informatics. 2001;34(5):301–310.

[22] Uzuner Ö, South BR, Shen S, DuVall SL. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. JAMIA. 2011;18(5):552–556. Available from: http://dx.doi.org/10.1136/amiajnl-2011-000203.

[23] Ben Abacha A, Zweigenbaum P. Medical Entity Recognition: A Comparison of Semantic and Statistical Methods. In: BioNLP 2011 Workshop. Portland, Oregon, USA: Association for Computational Linguistics; 2011. p. 56–64. Available from: http://www.aclweb.org/anthology/W11-0207.

[24] Ben Abacha A, Demner-Fushman D. Meta-Learning with Selective Data Augmentation for Medical Entity Recognition. In: 17th International Conference on Intelligent Text Processing and Computational Linguistics, CICLing 2016. Konya, Turkey; 2016. .

[25] Pradhan S, Elhadad N, South B, Martinez D, Christensen L, Vogel A, et al. Evaluating the State of the Art in Disorder Recognition and Normalization of Clinical Narrative. Journal of the American Medical Informatics Association (JAMIA). 2015;22(1):143–154.

[26] Dogan RI, Leaman R, Lu Z. NCBI disease corpus: A resource for disease name recognition and concept normalization. Journal of Biomedical Informatics. 2014;47:1–10. Available from: http://dx.doi.org/10.1016/j.jbi.2013.12.006.

[27] Rosario B, Hearst MA. Classifying Semantic Relations in Bioscience Texts. In: Proceedings of the 42nd Annual Meeting of the Association for Computational Linguistics, 21-26 July, 2004, Barcelona, Spain.; 2004. p. 430–437. Available from: http://acl.ldc.upenn.edu/acl2004/main/pdf/309_pdf_2-col.pdf.

[28] Ely JW, Osheroff JA, Gorman PN, Ebell MH, Chambliss ML, Pifer EA, et al. A taxonomy of generic clinical questions: classification study. British Medical Journal. 2000;321:429–432.