A Hybrid Approach to Clinical Question Answering

Sadid A. Hasan, Xianshu Zhu, Yao Dong, Joey Liu, and Oladimeji Farri
Philips Research North America, Briarcliff Manor, NY, USA
{sadid.hasan,xianshu.zhu,yao.dong,joey.liu,Dimeji.Farri}@philips.com

Abstract

In this paper, we describe our clinical question answering system developed and submitted for the Text Retrieval Conference (TREC 2014) Clinical Decision Support (CDS) track. The task for this track was to retrieve relevant biomedical articles to answer generic clinical questions about medical case reports. As part of our maiden participation in TREC, we submitted a single run using a hybrid Natural Language Processing (NLP)-driven approach to accomplish the given task. Evaluation results showed that our clinical question answering system achieved the best scores in two of eight dual-judged topics: #5 and 27, and performed relatively better compared to the median scores for topics: #13, 18, 19, 22, and 23.

1 Introduction

The TREC 2014 CDS track[1] aims at investigating techniques to improve patient care through providing pertinent biomedical information related to medical case reports. The primary motivation for such a task relies on the use case where a clinician can seek relevant research-based evidence on how best to care for patients at the point of care. For example, the clinician may require specific information on the patient's most likely diagnosis given a list of signs/symptoms, the most essential tests/procedures in a given scenario, and the most effective treatment plan given a diagnosis. In some cases, these types of information can be obtained from published biomedical literature that can eventually serve as potential clinical evidence to support patient care. However, due to the exponential growth of publications in the biomedical domain over the years, it has become nearly impossible to manually mine such a huge volume of scientific information repositories to find the most relevant and up-to-date details for a particular clinical scenario. Intelligent CDS systems can be useful to overcome this difficulty through automated clinical question answering. Hence, the main goal of the TREC 2014 CDS track is to promote research on systems that can satisfy the information need of the clinicians by retrieving relevant biomedical articles to answer generic clinical questions.

The proposed task for this track was to retrieve a ranked list of the top 1000 biomedical articles that can answer questions related to multiple categories of clinical information needs. In particular, short medical case reports were associated with one of three generic clinical questions: *"What is the patient's diagnosis?"*, *"What tests should the patient receive?"*, and *"How should the patient be treated?"*. The retrieved articles were judged in terms of their relevance to the corresponding clinical question associated with a given case report. Our submission for the CDS track uses a variety of NLP-based techniques to address the clinical questions provided. We present a description of our approach, and discuss our experimental setup, results and evaluation in the subsequent sections.

2 Description of Our Approach

Our hybrid NLP-driven method presents a combination of syntactic, semantic and filtering processes towards extracting relevant biomedical articles corresponding to clinical concepts (diagnoses, treatment and/or test) relevant to each given topic. Our overall

[1] http://www.trec-cds.org/

approach centers on three main processes: (i) Topical Keyword Extraction: extraction of ontology-based topical keywords (e.g. findings, disorders, body structures, procedures, tests, and treatments) along with demographic information from the given medical case reports (i.e., topic descriptions); (ii) Knowledge-based Clinical Inferencing: use of topical keywords as queries to a third-party clinical knowledge base and extraction of a ranked list of inferred diagnoses/tests/treatments corresponding to each given topic; and, (iii) Biomedical Literature Retrieval: retrieval and ranking of pertinent biomedical articles based on the keywords, concepts, and the ranked list of inferred diagnoses/tests/ treatments extracted in the prior steps.

As an initial step, we extract topical keywords from the topic descriptions and map the keywords to categories represented in clinical domain ontologies (e.g. findings, disorders, treatment etc.), in addition to retrieving demographic details from the topic descriptions. The use of clinical domain ontologies is effective in this step as they have been implemented to promote standard clinical vocabulary, and are widely used to semantically categorize clinical concepts, and facilitate information exchange and interoperability (Bodenreider, 2008; Stenzhorn et al., 2008; Garde et al., 2007). We use the following clinical domain ontologies: SNOMED CT[2] (Cornet and de Keizer, 2008) for diagnoses, LOINC[3] for tests, and RxNorm[4] for treatments.

In the next step, we utilize the topical keywords as queries to a clinical knowledge base, which is derived from Wikipedia[5] articles (clinical medicine category) and indexed using Elasticsearch[6] technology. This step aims to find relationships between topical keywords and associated clinical concepts (diagnoses/disorders, treatment and test) within a comprehensive knowledge base for the purpose of biomedical evidence retrieval. Wikipedia has been successfully used as a knowledge source by the information extraction community over the last few years (Wu and Weld, 2010). Clinical concepts found in the Wikipedia articles are filtered using various criteria e.g., location, gender, match with topical keywords, etc., and the resulting list of Wikipedia articles with relevant clinical concepts are mined to retrieve a ranked list of inferred diagnoses/tests/treatments corresponding to each given topic description.

In the final step, topical keywords and the corresponding disorders/diagnoses, tests, and treatments obtained from the clinical knowledge base are used to retrieve candidate biomedical articles by searching through TREC-CDS abstracts of PubMed Central articles. Candidate articles are ranked using multiple weighting algorithms designed to address each category of clinical questions (diagnosis, test, and treatment). The retrieved biomedical articles are further filtered by location, demographic information and other parameters (e.g. species) towards improving the relevance of the results. The final list of top 1000 biomedical articles are ordered by article publication date to support the clinician's synthesis of current research evidence related to the questions for each topic description.

3 Experimental Setup

3.1 Test Data

The test dataset comprises 30 topics divided into three question types as mentioned above. The given topic descriptions (or topics) are essentially medical case narratives that describe scenarios related to patient's medical history, signs/symptoms, diagnoses, tests, and treatments. The topics are provided in two versions depending on the depth of information. Topic "descriptions" include comprehensive descriptions of the patient's situation whereas topic "summaries" contain the most important information. We used descriptions for our experiments in order to utilize the unfiltered and richer context of the available patient information.

3.2 Corpus

The document collection for the track comes from the open access portion of PubMed Central[7] (PMC), a freely available online database of full-text biomedical articles. The provided collection was a snapshot of the open access subset and consisted of over $700,000$ biomedical publications.

[2] http://www.ihtsdo.org/snomed-ct/

[3] http://loinc.org/

[4] http://www.nlm.nih.gov/research/umls/rxnorm/

[5] https://www.wikipedia.org/

[6] http://www.elasticsearch.org/

[7] http://www.ncbi.nlm.nih.gov/pmc/

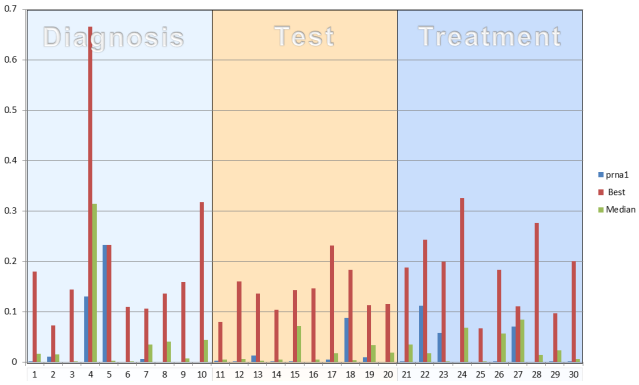

Figure 1: infAP scores for each topic

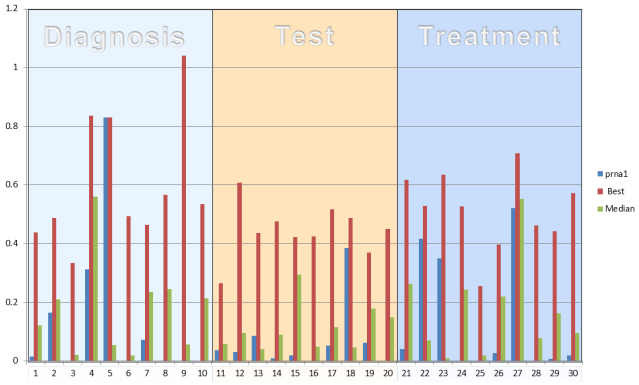

Figure 2: infNDCG scores for each topic

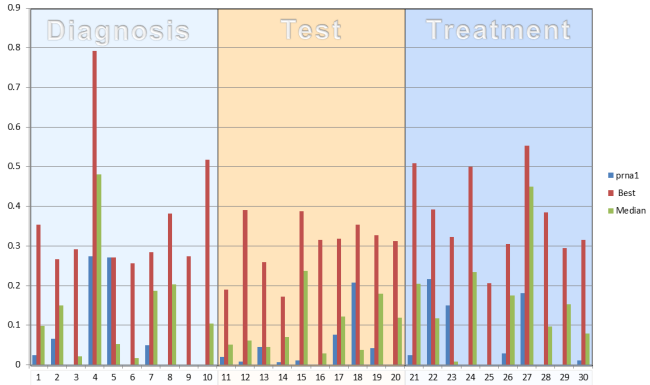

Figure 3: R-prec scores for each topic

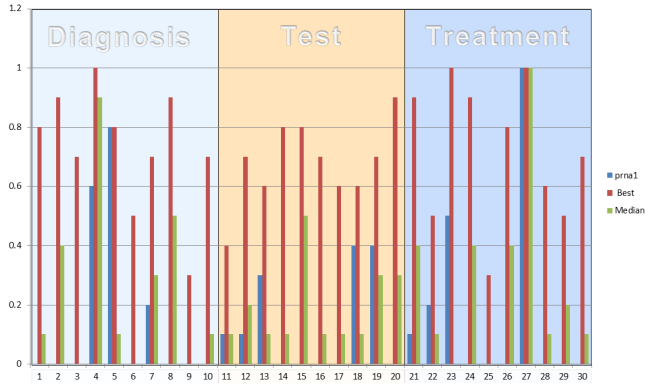

Figure 4: Prec(10) scores for each topic

3.3 Evaluation and Analysis

The evaluation of the CDS track was conducted using the standard TREC evaluation procedures for ad-hoc information retrieval tasks (Yilmaz et al., 2008; Voorhees, 2014). The highest ranked biomedical articles were sampled and judged by medical domain experts on a three-point scale of 0: not relevant, 1: possibly relevant, and 2: definitely relevant depending on the relevance of the answer to the associated question type about a given case report.

Figure 1 to Figure 4 show the overall scores of our system (*prna1*) across all the topics (categorized into three groups: diagnosis, test, and treatment) as compared to the *median* and *best* scores across all the submitted automatic runs for the following evaluation measures: inferred average precision[8] (infAP), inferred normalized discounted cumulative gain[9] (infNDCG), precision at R where R is the number of known relevant documents (R-prec), and precision at 10 documents (Prec (10)). The two inferred measures are used to provide more accurate estimates of a system's performance when relevance judgments are incomplete due to dynamic and/or larger document collections (Yilmaz and Aslam, 2006; Yilmaz et al., 2008). All the evaluation measures used for the CDS track contribute towards providing a sound view about the quality of a system. The reported results show that our clinical question answering system mostly performs close to the *median* scores for all evaluation measures.

[8] *Average Precision (AP)* is a measure that combines precision and recall for evaluating systems that retrieve a ranked list of articles. In particular, AP is the mean of the precision scores after each relevant article is retrieved.

[9] *Discounted Cumulative Gain (DCG)* measures the quality of ranking for a system when it retrieves a ranked list of results and the results are graded with relevance judgment. In particular, DCG computes the usefulness of an article based on its rank in the retrieved list. *Normalized DCG (NDCG)* is computed by using the maximum possible DCG (calculated by sorting the result list by relevance) as the normalization factor.

Analysis of these results also demonstrates that our clinical question answering system has achieved the best scores in two of eight dual-judged topics: #5 and 27, and performed relatively better compared to the median scores for topics: #13, 18, 19, 22, and 23. These results further emphasize the overall performance of our system in terms of answering the various question types represented in the topic descriptions.

4 Conclusion

In this paper, we described our participation in the inaugural TREC 2014 Clinical Decision Support Track. Evaluation results showed the effectiveness of our clinical question answering system. Next steps include improving the system's performance with more domain-specific clinical knowledge bases along with more NLP algorithms (e.g., paraphrasing and textual entailment) for better clinical reasoning and question answering.

References

O. Bodenreider. 2008. Biomedical Ontologies in Action: Role in Knowledge Management, Data Integration and Decision Support. *IMIA Yearbook of Medical Informatics*, 47(1):67–79.

R. Cornet and N. de Keizer. 2008. Forty Years of SNOMED: A Literature Review. *BMC Medical Informatics and Decision Making*, 8:1–7.

S. Garde, P. Knaup, E. Hovenga, and S. Heard. 2007. Towards Semantic Interoperability for Electronic Health Records. *Methods of Information in Medicine*, 46(3):332–343.

H. Stenzhorn, S. Schulz, M. Boeker, and B. Smith. 2008. Adapting Clinical Ontologies in Real-World Environments. *Journal of Universal Computer Science*, 14(22):3767–3780.

E. M. Voorhees. 2014. The Effect of Sampling Strategy on Inferred Measures. In *Proceedings of the 37th International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '14, pages 1119–1122.

F. Wu and D. S. Weld. 2010. Open Information Extraction Using Wikipedia. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ACL '10, pages 118–127.

E. Yilmaz and J. A. Aslam. 2006. Estimating Average Precision with Incomplete and Imperfect Judgments. In *Proceedings of the 15th ACM International Conference on Information and Knowledge Management*, CIKM '06, pages 102–111.

E. Yilmaz, E. Kanoulas, and J. A. Aslam. 2008. A Simple and Efficient Sampling Method for Estimating AP and NDCG. In *Proceedings of the 31st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, SIGIR '08, pages 603–610.