Using Neural Embeddings for Diagnostic Inferencing in Clinical Question Answering

Sadid A. Hasan, Yuan Ling, Joey Liu, and Oladimeji Farri

Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA, USA {sadid.hasan, yuan.ling, joey.liu, Dimeji.Farri}@philips.com

Abstract

In this paper, we describe our clinical question answering system implemented for the Text Retrieval Conference (TREC 2015) Clinical Decision Support (CDS) track. We submitted six runs for two related tasks using a multi-step approach that leverages Natural Language Processing (NLP) and neural embeddings to retrieve relevant biomedical articles for answering generic clinical questions. Evaluation results demonstrated that our system achieved higher scores for most clinical questions requiring answers that pertain to treatment and test/investigations, topics in which the ground-truth diagnoses were provided by the track organizers. However, our system was less accurate with questions requiring answers pertaining to diagnosis. We conclude that diagnostic inferencing may be the most important determinant of the accuracy of the clinical question answering systems, and that the use of neural embeddings and advanced deep learning techniques could help improve the quality of such systems in order to effectively support decision-making in patient care.

1 Introduction

Similar to 2014, the main objective of the 2015 CDS track¹ was to retrieve a ranked list of the top 1000 biomedical articles that can answer questions related to multiple categories (diagnosis, test, and treatment) of clinical information needs. There were two tasks in the 2015 edition of the CDS track: *Task A* was identical to the 2014 task. In addition, based on the feedback from our participation in the

2014 edition where we demonstrated that clinical knowledge-based inferencing is a critical factor for more focused retrieval of biomedical articles (Hasan et al., 2014), the organizers defined a new task this year, *Task B*, where the ground truth diagnoses were provided for the test and treatment topics in order to verify if such important information can actually improve the accuracy of the relevant biomedical article retrieval. We submitted six runs for the two tasks using a variety of NLP-based techniques to address the clinical questions provided.

Motivated by the recent surge of Deep Learning (DL) techniques in demonstrating superior performance over the traditional supervised and unsupervised Machine Learning (ML) techniques for various NLP tasks, we exploited the strength of neural word and phrase embeddings in extending the context of the underlying topics (clinical case narratives) towards improved diagnostic inferencing for our clinical question answering system. Current applications of deep learning in NLP tasks largely rely on learning high-dimensional vector representations of words and their relationships (called word embeddings) using neural network architectures. Once a language model is trained in this manner, semantically related words would be transformed into similar vector representations (Mikolov et al., 2013). Following the success of word embeddings in modeling high-level abstractions in textual data, researchers have progressed beyond the wordlevel to propose deep learning algorithms for learning phrase-level, sentence-level, and document-level representations (Le and Mikolov, 2014). The main advantage of this architecture over the traditional bag-of-words model is its ability to capture the embedded ordering and semantics of the words

¹http://www.trec-cds.org/

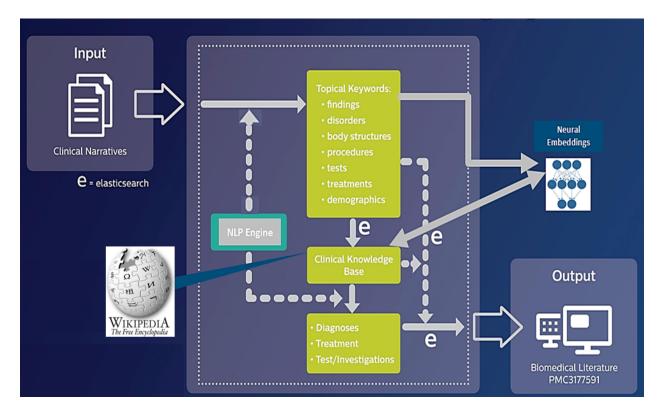


Figure 1: Clinical question answering system architecture

while learning a fixed-length vector representation for variable-length text structures. Our experiments showed that deep neural embeddings can provide performance improvements for a number of topics and the evaluation results further emphasized the significance of accurate diagnostic inferencing for effective retrieval of relevant biomedical articles for the automated clinical question answering task. In the subsequent sections, we describe the overall architecture of our system, and present the evaluation results with analyses.

2 System Description

Figure 1 shows the generic architecture of our clinical question answering system. Our hybrid NLPdriven 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. Similar to last year (Hasan et al., 2014), our overall approach centers on three steps: (i) Topical Analysis: identifying the most relevant topical keywords from the given topic descriptions or summaries; (ii) Clinical Inferencing: reasoning through the topical keywords to determine the appropriate diagnoses, tests, and treatments based on underlying clinical context by using neural embeddings and/or an external clinical knowledge base; and, (iii) Relevant Article Retrieval: retrieving and ranking pertinent biomedical articles based on the topical keywords and clinical inferences from steps (i) and (ii). Our submitted runs are varied based on the use of topical descriptions or summaries in association with different clinical inferencing methodologies.

In the first step, we extracted term frequencyinverse document frequency (TF-IDF) weighted topical keywords from the given descriptions/summaries and mapped them to categories represented in controlled clinical vocabularies/ontologies. We also identified relevant demographic information, interpreted vital patient parameters based on standard normal range values, and filtered out negated clinical concepts in order to give more weight to positive clinical manifestations in a given patient scenario. 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). In our system, we used the following clinical domain ontologies: SNOMED CT² (Cornet and de Keizer, 2008) for diagnoses, LOINC³ for tests, and RxNorm⁴ for treatments.

In the next step, we implemented a word- and phrase-to-vector neural embedding model trained on over 4 million clinically relevant sentences garnered from multiple clinical data sources: PubMed Central⁵ articles, Wikipedia (Clinical Medicine⁶) articles, and MIMICII discharge summaries (Saeed et al., 2011). The generated model is used to capture the overall context of a given topic description or summary towards inferring the differential diagnoses based on the commonest clinical diagnoses represented in the clusters of topical keywords from the first step. We used the skip-gram model architecture to learn the vector representations of words and phrases as it has been reported to provide better results in comparison to the continuous bag-of-words (CBOW) model (Mikolov et al., 2013). The list of possible diagnoses was further validated, filtered, and ranked by referencing a clinical knowledge base (derived from Wikipedia articles and indexed using Elasticsearch⁷) to extract a list of candidate articles with relevant diagnoses corresponding to each topical keyword. Our main goal was 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. Candidate Wikipedia articles were filtered using various criteria e.g., location, gender, match with topical keywords, etc., and the resulting list of Wikipedia articles with relevant clinical concepts were mined to retrieve specific diagnoses (from the title of the Wikipedia article), and tests and treatments (from

sections and subsections of the Wikipedia article). It is important to note that for *Task B*, we skipped inferring the differential diagnoses for the topic descriptions/summaries under the "treatment" and "test" categories, given that the organizers included the diagnoses (ground truth) in the test data.

In the final step, topical keywords and the corresponding disorders/diagnoses, tests, and treatments obtained from the clinical inferencing step were used to retrieve candidate biomedical articles by searching through the TREC-CDS database of PubMed Central articles. Candidate articles were ranked using multiple weighting algorithms specific to the three types of clinical questions (diagnosis, test, and treatment). The retrieved biomedical articles were further filtered by location (e.g. USA/Canada), demographic information and other contextual information from the topic description/summary towards improving the relevance of the results. The final list of top 1000 biomedical articles was ordered by article publication date to provide chronological biomedical evidence for the answers to each topic.

3 Experimental Setup

3.1 Test Data

The test dataset comprises 30 topics divided into three question types: topic 1-10 (diagnosis), topic 11-20 (test), and topic 21-30 (treatment). 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 an abridged version of the most important information. In addition, for *Task B*, ground truth diagnoses were provided for the test and treatment topics.

3.2 Corpus

Similar to 2014, a snapshot of the open access portion of PubMed Central (PMC), a freely available online database of full-text biomedical articles comprising 733, 138 biomedical publications was made available by the TREC CDS track organizers.

²http://www.ihtsdo.org/snomed-ct/

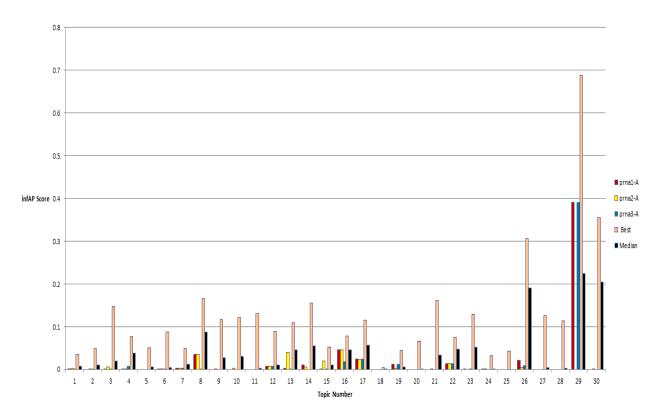
³http://loinc.org/

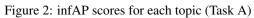
⁴http://www.nlm.nih.gov/research/umls/rxnorm/

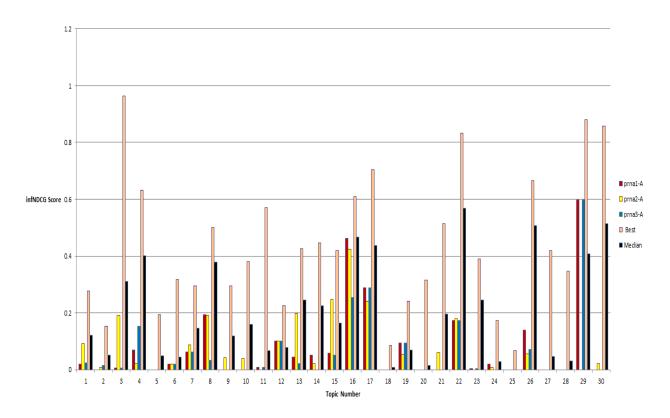
⁵http://www.ncbi.nlm.nih.gov/pmc/

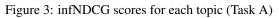
⁶https://en.wikipedia.org/wiki/Category:Clinical_medicine

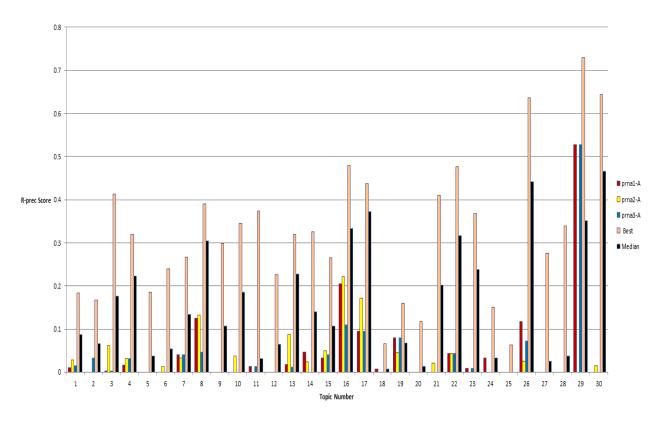
⁷http://www.elasticsearch.org/

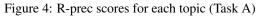












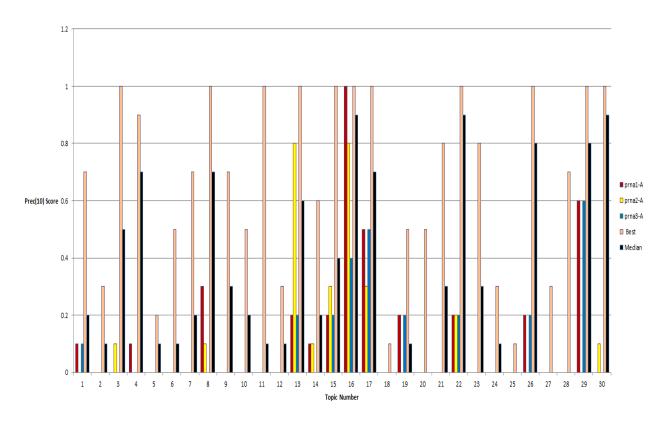
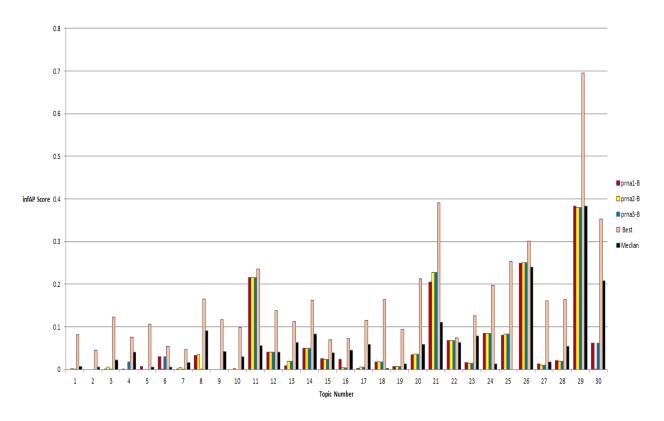
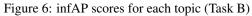


Figure 5: Prec(10) scores for each topic (Task A)





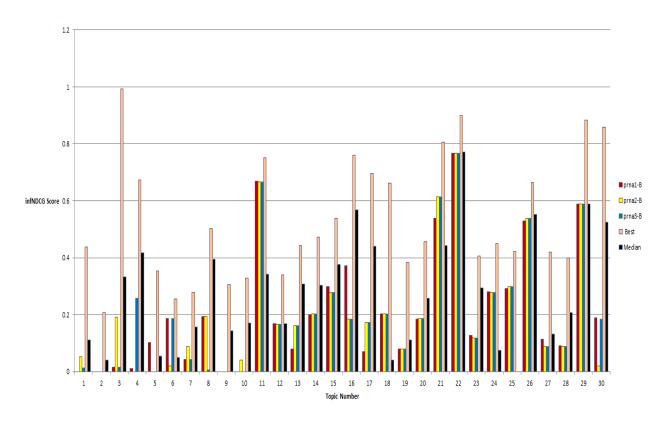
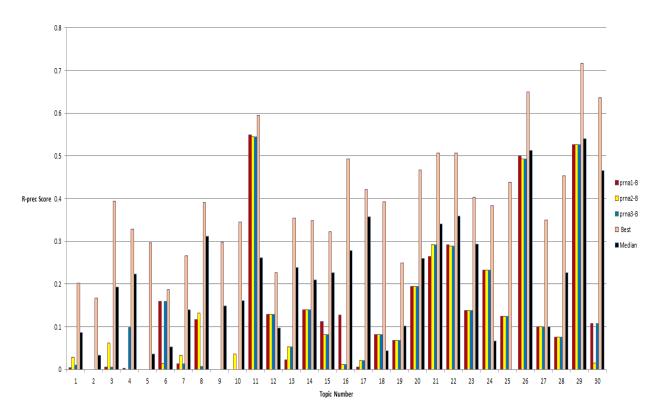


Figure 7: infNDCG scores for each topic (Task B)





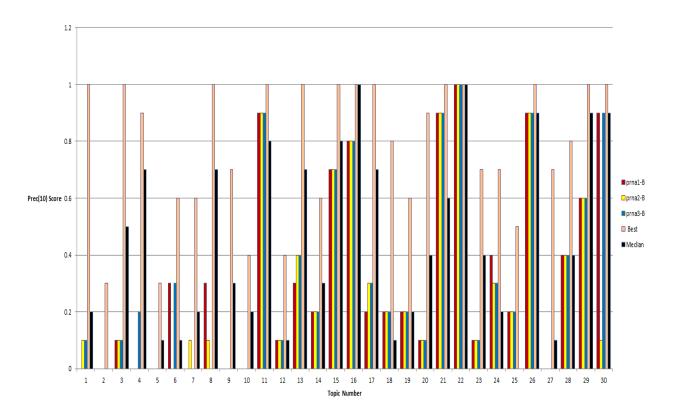


Figure 9: Prec(10) scores for each topic (Task B)

3.3 Run Description

For *Task A*, we submitted three runs as follows: 1) prna1-A: considers topic descriptions with clinical knowledge base inferencing, 2) prna2-A: considers topic summaries with clinical knowledge base inferencing, and 3) prna3-A: considers topic descriptions with clinical inferencing using neural word/phrase embeddings and the clinical knowledge base. Our three runs for *Task B* were designed in the same fashion with the exception that we skipped inferring the differential diagnoses for the topic descriptions/summaries under the "treatment" and "test" categories as we simply used the ground truth diagnoses provided in the test data.

3.4 Evaluation and Analysis

The evaluation of the CDS track was conducted using the standard TREC evaluation procedures for adhoc 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 2 to Figure 5 and Figure 6 to Figure 9 show the overall scores of our runs for *Task A* and *Task B* respectively across all the topics as compared to the *median* and *best* scores for the submitted automatic runs for the following evaluation measures: inferred average precision⁸ (infAP), inferred normalized discounted cumulative gain⁹ (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 comprehensive assessment of the quality of a system.

The reported results show that our clinical question answering system performs close to or better than the median scores for 43% of the topics in Task A and 50% of the topics in Task B across all evaluation measures. Analysis of these results also demonstrates that our clinical question answering system can achieve better results for 53% of the topics when topic summaries are used whereas neural word/phrase embeddings could improve upon the scores for 13% of the topics. Moreover, we find that the results for Task B improves drastically over those for Task A for most of the test and treatment topics across all runs by often outperforming the median scores - thus, emphasizing that accurate differential diagnoses can have a significant impact on the accuracy of the relevant biomedical article retrieval.

4 Conclusion and Future Work

In this paper, we described our participation in the TREC 2015 CDS Track. Evaluation results showed additional gains with the implementation of neural word and phrase embeddings in extending relevant context for our clinical question answering system. Results also demonstrated the importance of accurate inferencing of diagnosis in retrieving relevant biomedical articles corresponding to underlying clinical narratives. In future, we plan to improve our clinical inferencing algorithms towards extracting the most accurate differential diagnoses by employing advanced deep learning architectures (e.g. Recurrent Neural Networks (RNNs)) for training models from large collections of clinical knowledge sources.

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

⁸Average 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.

⁹Discounted 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.

Records. *Methods of Information in Medicine*, 46(3):332–343.

- S. A. Hasan, X. Zhu, Y. Dong, J. Liu, and O. Farri. 2014. A Hybrid Approach to Clinical Question Answering. In Proceedings of the Twenty-Third Text REtrieval Conference, TREC 2014, Gaithersburg, Maryland, USA, November 19-21, 2014.
- Q. V. Le and T. Mikolov. 2014. Distributed representations of sentences and documents. In *Proceedings of* the 31th International Conference on Machine Learning, ICML 2014, Beijing, China, 21-26 June 2014, pages 1188–1196.
- T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 27th Annual Conference on Neural Information Processing Systems NIPS 2013*, pages 3111– 3119.
- M. Saeed, M. Villarroel, A. T. Reisner, G. Clifford, L. Lehman, G. Moody, T. Heldt, T. H. Kyaw, B. Moody, and R. G. Mark. 2011. Multiparameter intelligent monitoring in intensive care ii MIMIC-II: A public-access intensive care unit database. *Critical Care Medicine*, 39:952–960.
- 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.
- 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.