Automated Clinical Diagnosis: The Role of Content in Various Sections of a Clinical Document

Vivek Datla, Sadid A. Hasan, Ashequl Qadir, Kathy Lee, Yuan Ling, Joey Liu, and Oladimeji Farri
Artificial Intelligence Laboratory, Philips Research North America, Cambridge, MA
Email: {firstname.lastname,kathy.lee_1,dimeji.farri}@philips.com

***Abstract*—Clinical diagnosis is a critical aspect of patient care that is typically driven by expert medical knowledge and intuition. An automated system for clinical diagnosis could reduce the cognitive burden of clinicians during patient care and medical education. In this paper, we describe a Knowledge Graph (KG)-based clinical diagnosis system that leverages publicly available knowledge sources to infer possible diagnoses from free-text clinical narratives. We experiment with the content in various sections of a clinical document within the electronic health record (EHR) to investigate the contribution of each section to the performance of automated diagnosis systems. Evaluation on MIMIC-III dataset demonstrates that the content of "history of present illness" and "past medical history" sections can play a greater role for clinical diagnosis inference than other sections and all sections combined. Comparison with a state-of-the-art deep learning-based clinical diagnosis system confirms the effectiveness of our system.**

***Keywords*-clinical diagnosis; knowledge graph; electronic health record;**

I. Introduction

Clinical diagnosis is a critical and non-trivial aspect of patient care. Intuition based on past professional experiences and knowledge gained from formal medical training typically drives the clinician's ability to make a diagnosis [1]. Although mimicking the intuition of clinicians can be very challenging, an automated system designed for clinical diagnosis can support expert reasoning based on available knowledge sources, especially when trying to resolve complicated clinical scenarios. Such a system could significantly reduce the cognitive burden of clinicians during patient care so they could be better-informed and adequately engage their patients towards achieving desired health outcomes [2], [3].

Available text-based knowledge sources for medicine include scientific publications and textbooks. However, a significant proportion of these sources are proprietary and require formal and commercial agreements in place for wide-spread use in automated systems for clinical decision support. Instead, we use Wikipedia as our knowledge source given the fact it is publicly available and that medical and colloquial usage of medical terms are represented - a feature that may help build a robust computational model for automated diagnostic inferencing. Furthermore, Wikipedia is used by several researchers in the field of natural language processing (NLP) as a rich multilingual knowledge base useful in various tasks including question answering (QA) and automated reasoning [4], [5], [6].

In this paper, we leverage articles under the "Clinical Medicine" category of Wikipedia to build a knowledge-driven clinical diagnosis system. We use a Knowledge Graph (KG)-based approach to accomplish this goal. Our system takes free-text description of a medical problem (a clinical narrative) as input and provides the most likely diagnoses. We convert the link structure in Wikipedia into a knowledge graph where the nodes represent Wikipedia pages, hyperlinked concepts, and redirect pages, while edges represent the relationships between them. We develop a query system on the knowledge graph that utilizes the content of Wikipedia as well as the link structure in identifying the most probable diagnoses. The structure is used to determine relationships among diseases and symptoms while the content of Wikipedia pages is used to rank their strength of association. Based on the strength of association, the system generates a ranked list of diagnoses for a given medical problem.

Our experiments on MIMIC (Medical Information Mart for Intensive Care)-III [7] discharge summaries demonstrate that identifying relevant sections in these documents can lead to a substantial gain in performance in inferring the most probable diagnoses. We observe that providing content from all sections of the document to the system works poorly compared to using specific sections of the document. We also compare the KG-based system's performance to the state-of-the-art Condensed Memory Networks (C-MemNN)-based clinical diagnosis system [8] also trained on MIMIC-III dataset. Evaluation results reveal that our system performs better in some of the experiments.

Given the increasing interest in artificial intelligence (AI) and clinical decision support systems within the machine learning and health informatics communities, our work helps identify the most appropriate information within electronic clinical documents that would drive automated diagnostic solutions towards optimal accuracy leading to better-informed clinical decisions. Researchers in these communities can utilize findings of our paper to improve quality of training data when developing AI models to address complex reasoning tasks in patient care.

The main contributions of this paper can be summarized

as follows:

- Construction of a KG-based clinical diagnosis inference engine.
- Experiments with different sections of an EHR note to identify sections that contribute the most for an accurate clinical diagnosis inference.
- Comparison with a state-of-the-art system [8] to show the effectiveness of the proposed approach.

In the next sections, we link our work to the literature and describe the proposed KG-based clinical diagnosis inference approach, followed by the detailed experimental setup, results and discussion. Finally, we conclude the paper in Section VI.

II. Related Work

AI systems for clinical decision support have been previously developed using bio-signals from patients [9], [10], [11]. Such structured clinical data contain raw signals without much context for accurate interpretation, whereas unstructured free-text clinical documents contain detailed descriptions of overall clinical scenarios. EHRs typically store both structured clinical data (including physiological signals, vital signs, lab tests, etc.) in addition to unstructured text documents that contain a relatively more complete picture of associated clinical events.

Diagnostic inferencing from medical narratives has gained much attention in recent times [2], [3], [12], [13], [14], [15], [16], [17], [18]. Researchers have formulated the problem of diagnostic inferencing from free-text as a document retrieval task (medical literature retrieval) [19], [20] or as a multiclass-multilabel classification task [8], [14]. In a document retrieval task, the goal is to obtain documents from a given database that mention and/or describe possible diagnoses for the underlying clinical scenario. In the multiclass-multilabel classification approach, the classes are predefined to represent the most frequent diagnoses in the training set, and classification models are developed to analyze clinical scenarios to generate a list of differential diagnoses.

To infer clinical diagnoses, a few research works have explored graph-based reasoning methods where the graphs incorporate relevant medical concepts and their associations. Shi et al. [21] organized textual medical knowledge into conceptual graphs and proposed a contextual information pruning algorithm to conduct semantic reasoning over the graph. Geng et al. [22] constructed a causal graph for medical knowledge representation and inferred diagnosis over the graph, but focused on selected diseases.

Overcoming the current limitations of improving the accuracy levels of Clinical QA, especially scenario-based analysis, [23], [24] may require leveraging domain expertise from a variety of sources (e.g. domain-specific knowledge bases). However, knowledge representation is one of the fundamental problems in Artificial Intelligence. Over the decades of research, there have been several solutions and ontologies proposed to address the problem. Examples of successful open domain knowledge graph (KG) representations include Freebase [25], YAGO [26], and DBpedia [27]. These KG representations have been successfully used to answer factoid-based questions such as *'Who won the 2017 super bowl?"*, but it is not clear how to address non-factoid questions such as *"Why is there a drop in blood pressure while a person reaches higher altitudes?"*. Many methods [28] convert the natural language question into a structured query and then search efficiently over large collection of resource description format (RDF) triples. These methods have the underlying assumption that the answer is a node or a path in the knowledge graph. Also, they are shown to work well in open domains, but it is not clear how they can be adapted for specific domains such as clinical medicine.

To facilitate the development of artificially intelligent systems assisting the clinical diagnosis inferencing process, the Text Retrieval Conference (TREC) has recently initiated the clinical decision support (CDS) track[3] that requires retrieval of relevant biomedical articles for a given clinical case narrative. The challenges of the task include answering three types of generic clinical questions: 1) Diagnosis ("*what is the patient's diagnosis?*"), 2) Tests ("*what tests should the patient receive?*"), and 3) Treatment ("*how should the patient be treated?*"). The organizers provided a collection of 30 topics for the task. One of the major challenges of building effective models for such intelligent clinical decision support applications arises from the unavailability of a large volume of annotated corpus for training and testing the models [29].

More recently, recurrent neural networks (RNNs) have been implemented in systems for clinical inferencing by utilizing structured clinical data [9], [10]. Prakash et al. [8] created a novel C-MemNN model based on RNNs with a memory component to enhance the possibility of arriving at the potential diagnosis for a given medical problem. We benchmark our KG-based diagnosis inference system against this state-of-the-art deep learning system to determine its effectiveness and recommend future directions for performance improvements.

Our KG-based system differs from existing reasoning models with respect to the construction of the knowledge graph, which we build incorporating medical concepts in Wikipedia represented by their corresponding pages. We also employ a hybrid activation-based querying mechanism, which is both knowledge-driven and data-driven. Furthermore, although there are few systems [30], [31] that extract section information directly from free-text in clinical narratives, to the best of our knowledge this is the first study that explores the impact of different sections in free text medical notes for predicting diagnoses from clinical narratives.

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

III. Inferring Clinical Diagnosis with Knowledge Graph

In this work, we introduce a novel hybrid approach to address the clinical diagnostic inferencing problem. Figure 1 shows the overall architecture of our system. We first build a structured knowledge graph (KG) using contents from Wikipedia that are relevant for this problem. Given a clinical narrative, we then identify the patient's symptoms in the narrative using an information extraction engine. The extracted symptoms are used to query the knowledge graph for predicting a set of diagnoses for the given narrative. The following sections discuss the details of the knowledge graph construction and our method for predicting diagnoses from a clinical narrative.

A. Knowledge Graph Construction

For constructing our knowledge graph, we used Wikipedia as our knowledge source. We collected all documents under the clinical medicine category in Wikipedia. This category served as the root node of our knowledge graph. The subcategories and any page in Wikipedia under this root category became the initial children nodes in our graph. The nodes representing the sub-categories might not have had any content text, whereas the nodes representing the pages had their content text. We further expanded the nodes recursively up to a depth of 10 using breadth-first search, extracted all subcategories and pages, mining a total of 188,139 Wikipedia pages from 17,121 categories. These pages and the categories were then added to the graph.

Some of the categories were verified by Domain experts as unrelated to clinical medicine, so we pruned our graph at these categories. Furthermore, we created an edge for any hyperlink associated with a term in any of the retrieved pages. These edges connected the page node that contained the term with the page node that was the hyperlink destination. The resulting knowledge graph (KG) contained a total of 381,964 nodes and 1,906,302 edges. The constructed knowledge graph is represented in the 4^{th} component in Figure 1.

B. Inferencing Diagnosis

1) Symptom Extraction from Clinical Narratives: The diagnosis inferencing process begins with a clinical narrative that describes the symptoms and any demographic information of a patient. The clinical narrative is written in unstructured text, so these concepts need to be identified and extracted before we can query the knowledge graph.

For example, the clinical narrative may contain sentences such as "A 5-year-old boy has fever, cough, drooling, stridor, and dysphagia with voice change." We use a hybrid clinical NLP engine [32] to first identify and extract the symptoms: *fever, cough, drooling, stridor, dysphagia with voice change*, and demographic information: *5-year-old boy*. We also use the NLP engine to normalize the symptoms so that they can be mapped to their corresponding Wikipedia page. The NLP engine uses medical ontologies such as SNOMED [33], UMLS [34], and RadLex [35] for normalization. Components 1-2 in Figure 1 represent these steps.

2) Querying the Knowledge Graph: Next, we query the knowledge graph with the extracted symptoms for predicting a set of diagnoses. However, not all symptoms contribute equally in the prediction process. For example, symptoms such as *"fever"* is very common and can occur with many diseases, whereas *"stridor"* can be more uniquely associated to respiratory diseases. So from the list of extracted symptoms, we need to identify the symptoms that are the most distinctive for determining potential diagnoses and weigh them accordingly. For each extracted symptom, we query the PubMed corpus of the 2014 TREC CDS track using Elasticsearch[1] to retrieve its term frequency in the corpus, and use the inverse of the term frequency as its weight. This signifies that the symptoms that are relatively rare in the PubMed corpus will be assigned higher weights than the symptoms that are more frequent. Component 3 in Figure 1 represents this step. The calculated weights are used to activate these nodes in the knowledge graph at a later step.

3) Building the Solution Space: Our next step is to create a solution space within the knowledge graph that consists of nodes representing symptoms, leading to nodes representing candidate diagnosis. We conduct this in two stages: a) building a bare-bones sub-space, and b) expanding the subspace to have a connected path between any two nodes.

a) Building the initial subspace: The solution space initially contains only the nodes representing the input query symptoms. We further include all of the immediate neighbors of the input symptom nodes in the initial solution space. This process gives us several trees which may or may not (a scattered forest) be connected. If the trees are connected i.e. if there is an existing path between any two nodes of the graph, then we identify it as a connected forest. This is represented in component 5 of Figure 1. If there is no connected forest, then we perform the next step.

b) Expanding the initial subspace: If the resulting forest is not connected then we expand the subspace. For this, we identify nodes that share common entities such as diseases, medications, procedures and symptoms. We use a greedy approach to minimize the number of new nodes added to the solution space by expanding the nodes that would provide the maximum connectivity with a minimal number of nodes added to our solution space. For implementing the greedy approach we identify two nodes in the knowledge graph that have the smallest number of children and share at least one child between them. This common child acts like a path between the two unconnected trees making it a single connected graph. We repeat the process till the whole graph

[1]https://www.elastic.co/products/elasticsearch

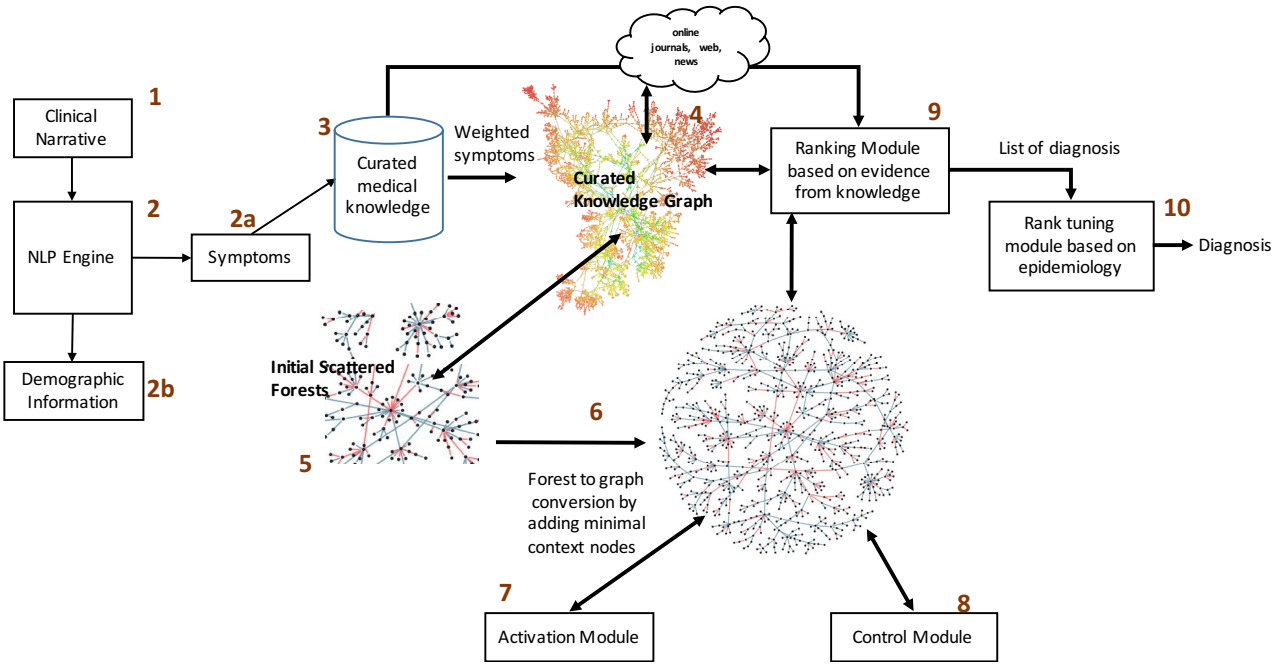

Figure 1. Knowledge graph based clinical diagnosis system

becomes connected.

The solution space can grow exponentially in size if the expanded node is a very common symptom. Also, expanding a very common symptom adds many unrelated diagnosis and procedures to our solution space. By following the expansion strategy mentioned above, we overcome the risk of adding the unrelated nodes to our solution space. Component 6 in Figure 1 represents this process.

4) Activating Nodes in the Solution Space: Next, we start activating nodes in the solution space to find a set of probable diagnoses. We start with the weights of the input query symptoms. These weights are then spread across the knowledge graph with an activation module and a control module.

The activation module takes the weight of a symptom node and propagates the weight to its immediate neighbor. The propagated weight is dampened by a factor, so that the weight propagated from a node weakens (i.e. lessens) when a propagation happens farther away from the initial symptom. All the nodes keep propagating the activation to their neighbors except to their parent node. The motivation for performing this step is to accumulate the weights from symptoms to the nodes containing diagnosis.

The control module is responsible for stopping an activation if the propagated weight is below a certain threshold. We use a very small value 0.001 as our threshold, below which the activation do not contribute much in the accumulation process. Also, the control module makes sure that there is no cyclic propagation of weights and keeps track of the nodes to which the current node has passed the activation.

The end result of this stage is a weighted graph, where each node is weighted based on the accumulation of the proportion of weights propagated from the symptoms. Components 7-8 in Figure 1 represent this stage.

5) Identifying and Ranking Diagnoses: Since our ultimate goal is to infer a set of diagnoses, any node in the weighted graph that is not a disease/syndrome node is filtered out from our solution space. The remaining nodes form the set of possible diagnoses for the input symptoms retrieved from the clinical narrative. For each disease node,

we check the signs and symptoms mentioned in Wikipedia for that disease, and score the node based on the overlap of the symptoms in the clinical narrative and the content of that disease/syndrome Wikipedia page. The diseases are then re-ranked based on the overlap score and they form the candidate set of diagnoses for the symptoms.

As the final ranking step, if the demographic information of the patient is retrieved from the clinical narrative, then we mine the epidemiology of the disease mentioned in Wikipedia to identify the prevalence of the current diagnosis in that age group. For example, if the disease is very prevalent in children but not in adults and the patient is mentioned as an adult, then the rank of the disease is pushed lower than the adult diseases in the list. Once the re-ranking process based on the epidemiology information is completed, we get our final ranked list of diagnoses that are inferred for the given clinical narrative. These final steps are represented in components 9-10 in Figure 1.

IV. Experiments and Evaluation

A. Dataset

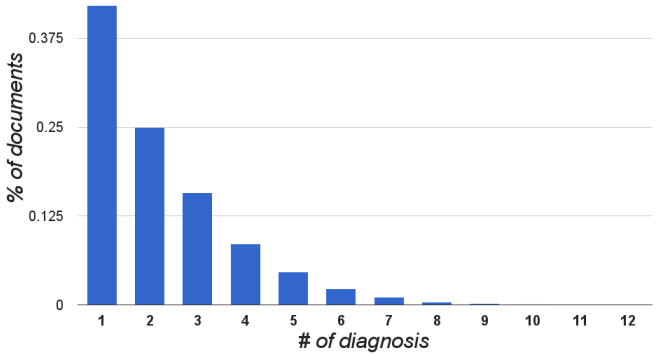

Figure 2. Distribution of diagnoses in MIMIC-III [8]

We evaluate the system on discharge notes in MIMIC-III database [36]. MIMIC-III contains physiological signals and various measurements captured from patient monitors, and comprehensive clinical data obtained from hospital medical information systems for over 58K hospital admissions. We use the note *events table* from MIMIC-III v1.3, which contains the free-text clinical notes for patients. We use 'discharge summaries,' instead of 'admission notes,' as former contains actual ground truth and free-text. Table I shows an example discharge note used in this paper. The diagnoses present in the MIMIC-III notes are very specific and are not evenly distributed as shown in Figure 2. Many diseases appear very few times inside the corpus.

For the experiments, we have used a subset of 14K notes. We processed the notes to extract the medical concepts from the MIMIC notes based on SNOMED [37] using our hybrid clinical NLP engine [32]. For example, after processing a discharge note (e.g. in Table I), we get the concepts shown in Table II.

There are 4,186 unique diagnoses in MIMIC-III discharge notes. However, many diagnoses (labels) occur in only a single note. The 50 most-common labels cover 97% of the notes, and the 100 most-common labels cover 99.97%. We present experiments for both the 50 most-common and 100-most common labels.

B. Comparison of Sections of EHR as Clinical Narrative

We conducted extensive experiments to understand the role of the content in a particular section of a clinical note to infer the correct diagnoses. Given the unstructured free-text in each section of the medical note as input, we measure the accuracy of the system in identifying the diagnoses. In this study, we consider the following sections of a MIMIC-III note for these experiments: *social history* (i.e., behavioral information such as smoking, drinking, diabetes etc.), *chief complaint* (i.e., symptoms such as chest pain, headache, dizziness, etc.), *history of present illness*, *past medical history*, *brief hospital course* (i.e. information about procedures and medications provided during the hospital stay), and *discharge medications*.

We also compare our system to a state-of-the-art clinical diagnosis inference system on the MIMIC-III dataset, which uses Condensed Memory Neural Networks (C-MemNNs) [8] to formulate the task as a multiclass-multilabel classification problem. Due to a large number of diagnoses (class labels) in the dataset, the C-MemNN model simplifies the task by considering the most frequent N diagnoses for training. We also adapt similar settings for our experiments.

C. Metrics

We use precision and recall to evaluate our systems. For a meaningful comparison, we consider two variations of these metrics: 1) strict (exact word match with the ground truth diagnosis), and 2) relaxed (allowing paraphrases and disease synonyms based on the human disease network [38]).

Recall that, our knowledge graph is built using the medical concepts in Wikipedia, where the clinical concepts are mostly standardized and may be different from the abbreviated/colloquial usage of medical terms in a clinical note. For example, a MIMIC note may refer to "diabetes mellitus type 2" by mentioning "dm type 2", "diabetes type 2", "db 2" or "diabetes 2". Since our approach considers diagnosis concepts based on the Wikipedia page titles, a strict measure of precision and recall based on exact word overlap with the ground truth diagnosis may be insufficient to measure the effectiveness of our systems. Hence, we introduced the relaxed alternatives of the metrics.

The precision at 5 represented as P@5 is the ratio of correct diagnoses over the top five predictions. It should be noted that a MIMIC chart note can have many diagnoses,

Table I
EXAMPLE OF A DISCHARGE NOTE IN MIMIC-III

Discharge Note (partially shown)
CHIEF COMPLAINT: Chest pain. HISTORY OF PRESENT ILLNESS:A 41-year-old female with a history of coronary artery bypass graft x3 in [**3216**] who has experienced substernal chest pain over the past two days. Patient initially attributed her discomfort to a cold. This afternoon pain worsened then spread to her arms and neck. She planned to see her doctor tomorrow, but due to this worsening of the pain, the patient decided to come to the Emergency Department.
Discharge Diagnosis
1. Three vessel coronary artery disease. 2. Occluded saphenous vein graft to obtuse marginal. 3. Mild systolic and diastolic left ventricular dysfunction. 4. Acute inferior myocardial infarction managed by acutePTCA. 5. Successful Angio-Jet and stenting of the distal right coronary artery beyond the saphenous vein graft-right coronary artery anastomosis.

Table II
MEDICAL CONCEPTS EXTRACTED USING A HYBRID CLINICAL NLP ENGINE [32]

Discharge Note
CHIEF COMPLAINT: Chest pain. HISTORY OF PRESENT ILLNESS: 41-year-old female , coronary artery bypass, chest pain, discomfort,cold, pain
Diagnosis
coronary artery disease, saphenous vein graft, myocardial infarction, stenting, coronary artery, coronary artery anastomosis.

hence making this a very strict measure. The recall at 5 represented as R@5 measures how well we covered all the possible diagnoses in the MIMIC note. In the relaxed setting of a measure we consider that the predicted diagnosis is correct even when the ground truth diagnoses and the predicted diagnoses are synonyms or paraphrases of each other.

D. Results and Discussion

Table III shows the results of our experiments with various sections of the medical chart note. From these results, we can understand the role of content to infer the correct diagnoses.

We can see that the individual sections perform better than the combined sections (*All*). This can be attributed to the generality of some of the sections in the MIMIC notes, where the procedures/medications apply to many diseases. Specifically, the *brief hospital course* section has many procedures that are common among several diseases, which may have led it to lower scores. On the other hand, the *discharge medications* section only covers the pain medications and may not be representative of the surgery or complications the patient had due to pre-existing chronic conditions.

Further analysis shows that *social history* has a higher score in the relaxed measure of precision when we consider the top 100 classes. Also, this can be an aberration as people with a social history of alcohol and smoking had more chances of having diabetes, hypertension and other lifestyle diseases that are among the most common diseases in MIMIC notes. Not surprisingly, the results show that *history of present illness* and *past medical history* have the most relevant information for identifying the diagnoses.

We also compare our results with the C-MemNN [8] model. In their paper, the authors report the results using three metrics: P@5, Area Under the Curve (AUC), and Hamming loss. AUC and Hamming loss are not the appropriate metrics for our experimental settings, so we use precision- and recall-based metrics for this comparison. Results show that our systems have lower strict precision scores for the "top-50 classes" experiments. However, when we consider the top 100 classes, the *All* sections variant performs better than the C-MemNN system, which also uses all sections (except the *diagnosis* section) as the input of their model. Considering the relaxed precision metric, we find that the proposed KG-based system can perform better than the C-MemNN model with the selective use of content from various sections.

From our experiments, it is clear that all sections do not contribute equally for clinical diagnosis inference. Hence, it might be difficult for a machine learning system to learn the complex relationships among the medical concepts and the diagnoses present in a clinical note. For the MIMIC-III dataset, our experiments suggest that training the model on the *past medical history* and *history of present illness* sections could help a machine learning system improve the accuracy of clinical diagnosis inference compared to considering the full clinical note.

Table III
EVALUATION RESULTS (ALL=COMBINATION OF ALL CONSIDERED SECTIONS; P=PRECISION; R=RECALL; S=STRICT; R=RELAXED; TOP SCORES ARE **BOLDFACED**).

Sections of EHR	Top 50				Top 100			
	R@5(r)	R@5(s)	P@5(r)	P@5(s)	R@5(r)	R@5(s)	P@5(r)	P@5(s)
Social history	0.667	0.177	0.403	0.082	**0.798**	0.104	**0.469**	0.048
Chief complaint	0.717	0.289	0.426	0.111	0.729	0.229	0.432	0.087
History of present illness	**0.735**	0.309	**0.460**	0.135	0.731	0.235	0.449	0.103
Past medical history	0.728	**0.312**	0.453	**0.152**	0.726	**0.257**	0.440	**0.128**
Brief hospital course	0.663	0.236	0.380	0.091	0.714	0.177	0.404	0.068
Discharge medications	0.589	0.183	0.343	0.087	0.533	0.113	0.300	0.055
All	0.669	0.286	**0.436**	0.135	0.799	0.219	**0.429**	0.103
C-MemNN [8]	-	-	-	**0.420**	-	-	-	**0.320**

V. LIMITATIONS

Due to the fact that a thorough evaluation of the NLP engine [32] used in our KG-based clinical diagnosis system was not conducted, especially with respect to information extraction from the MIMIC III discharge summaries, we could not adequately control for the errors in predicting the most likely diagnoses that may be attributable to errors in the extracted clinical concepts. However, we rely on the fact that the same NLP engine was used to process all sections of the discharge summaries in verifying that the results from our experiments are most likely unbiased in terms of errors in the extracted clinical concepts. We intend to evaluate the accuracy of the NLP engine on information extraction from the MIMIC III discharge summaries in the near future.

VI. CONCLUSION

In this paper, we described our Knowledge Graph (KG)-based clinical diagnosis inference system. We conducted extensive experiments on the MIMIC-III benchmark dataset considering various sections of a clinical note. Results demonstrated that the content of the *history of present illness* and *past medical history* sections can contribute the most for clinical diagnosis inference compared to all sections. Furthermore, we showed that the proposed KG-based system can perform well in comparison to the state-of-the-art C-MemNN model for a relaxed precision metric.

In future, we would improve the current KG-based diagnosis inference system by adding more properties (e.g. relationships among the clinical concepts) to the edges of the knowledge graph. Also, we would like to utilize the findings of this study to improve the training data sets for machine learning models that help infer the clinical diagnoses from free-text narratives.

REFERENCES

[1] G. Norman, M. Young, and L. Brooks, "Non-analytical models of clinical reasoning: the role of experience," *Medical Education*, vol. 41, no. 12, pp. 1140–1145, 2007. [Online]. Available: http://dx.doi.org/10.1111/j.1365-2923.2007.02914.x

[2] Y. Ling, S. A. Hasan, V. Datla, A. Qadir, K. Lee, J. Liu, and O. Farri, "Learning to diagnose: Assimilating clinical narratives using deep reinforcement learning," in *Proceedings of the 8th International Joint Conference on Natural Language Processing (IJCNLP)*, 2017.

[3] Y. Ling, S. A. Hasan, V. Datla, A. Qadir, K. Lee, J. Liu, and O. Farri, "Diagnostic inferencing via improving clinical concept extraction with deep reinforcement learning: A preliminary study," in *Proceedings of the 2nd Conference on Machine Learning for Health Care (MLHC)*, 2017.

[4] F. Wu and D. S. Weld, "Open information extraction using wikipedia," in *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, ser. ACL '10. Stroudsburg, PA, USA: Association for Computational Linguistics, 2010, pp. 118–127. [Online]. Available: http://dl.acm.org/citation.cfm?id=1858681.1858694

[5] D. Milne and I. H. Witten, "An open-source toolkit for mining wikipedia," *Artificial Intelligence*, vol. 194, pp. 222–239, 2013.

[6] B. Katz, G. Marton, G. C. Borchardt, A. Brownell, S. Felshin, D. Loreto, J. Louis-Rosenberg, B. Lu, F. Mora, S. Stiller *et al.*, "External knowledge sources for question answering." in *TREC*, 2005.

[7] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, 2016.

[8] A. Prakash, S. Zhao, S. A. Hasan, V. Datla, K. Lee, A. Qadir, J. Liu, and O. Farri, "Condensed memory networks for clinical diagnostic inferencing," *AAAI*, 2016.

[9] Z. C. Lipton, D. C. Kale, C. Elkan, and R. Wetzell, "Learning to diagnose with lstm recurrent neural networks," *arXiv preprint arXiv:1511.03677*, 2015.

[10] E. Choi, M. T. Bahadori, and J. Sun, "Doctor ai: Predicting clinical events via recurrent neural networks," *arXiv preprint arXiv:1511.05942*, 2015.

[11] E. Choi, M. T. Bahadori, J. Sun, J. Kulas, A. Schuetz, and W. Stewart, "Retain: An interpretable predictive model for healthcare using reverse time attention mechanism," in *Advances in Neural Information Processing Systems*, 2016, pp. 3504–3512.

[12] M. S. Simpson, E. M. Voorhees, and W. Hersh, "Overview of the trec 2014 clinical decision support track," DTIC Document, Tech. Rep., 2014.

[13] K. Roberts, M. S. Simpson, E. Voorhees, and W. R. Hersh, "Overview of the trec 2015 clinical decision support track," in *TREC*, 2015.

[14] S. A. Hasan, S. Zhao, V. Datla, J. Liu, K. Lee, A. Qadir, A. Prakash, and O. Farri, "Clinical question answering using key-value memory networks and knowledge graph." TREC, 2016.

[15] S. A. Hasan, Y. Ling, J. Liu, and O. Farri, "Using neural embeddings for diagnostic inferencing in clinical question answering," 2015.

[16] T. R. Goodwin and S. M. Harabagiu, "Medical question answering for clinical decision support," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 297–306.

[17] Y. Ling, Y. An, and S. A. Hasan, "Improving clinical diagnosis inference through integration of structured and unstructured knowledge," in *Proceedings of the 1st EACL Workshop on Sense, Concept and Entity Representations and their Applications (SENSE)*, 2017.

[18] Y. Ling, Y. An, M. Liu, S. A. Hasan, Y. Fan, and X. Hu, "Integrating extra knowledge into word embedding models for biomedical nlp tasks," in *Proceedings of the 30th International Joint Conference on Neural Networks (IJCNN)*, 2017.

[19] Z. Zheng and X. Wan, "Graph-based multi-modality learning for clinical decision support," in *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*. ACM, 2016, pp. 1945–1948.

[20] S. Balaneshin-kordan and A. Kotov, "Optimization method for weighting explicit and latent concepts in clinical decision support queries," in *Proceedings of the 2016 ACM on International Conference on the Theory of Information Retrieval*. ACM, 2016, pp. 241–250.

[21] L. Shi, S. Li, X. Yang, J. Qi, G. Pan, and B. Zhou, "Semantic health knowledge graph: Semantic integration of heterogeneous medical knowledge and services."

[22] S. Geng and Q. Zhang, "Clinical diagnosis expert system based on dynamic uncertain causality graph," in *Information Technology and Artificial Intelligence Conference (ITAIC), 2014 IEEE 7th Joint International*. IEEE, 2014, pp. 233–237.

[23] D. Ferrucci, A. Levas, S. Bagchi, D. Gondek, and E. T. Mueller, "Watson: beyond jeopardy!" *Artificial Intelligence*, vol. 199, pp. 93–105, 2013.

[24] A. Lally, S. Bachi, M. A. Barborak, D. W. Buchanan, J. Chu-Carroll, D. A. Ferrucci, M. R. Glass, A. Kalyanpur, E. T. Mueller, J. W. Murdock *et al.*, "Watsonpaths: scenario-based question answering and inference over unstructured information," *Yorktown Heights: IBM Research*, 2014.

[25] K. Bollacker, C. Evans, P. Paritosh, T. Sturge, and J. Taylor, "Freebase: a collaboratively created graph database for structuring human knowledge," in *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*. AcM, 2008, pp. 1247–1250.

[26] F. M. Suchanek, G. Kasneci, and G. Weikum, "Yago: a core of semantic knowledge," in *Proceedings of the 16th international conference on World Wide Web*. ACM, 2007, pp. 697–706.

[27] S. Auer, C. Bizer, G. Kobilarov, J. Lehmann, R. Cyganiak, and Z. Ives, "Dbpedia: A nucleus for a web of open data," in *The semantic web*. Springer, 2007, pp. 722–735.

[28] S. Yang, Y. Xie, Y. Wu, T. Wu, H. Sun, J. Wu, and X. Yan, "Slq: a user-friendly graph querying system," in *Proceedings of the 2014 ACM SIGMOD International Conference on Management of Data*. ACM, 2014, pp. 893–896.

[29] S. Jonnalagadda, T. Cohen, S. Wu, and G. Gonzalez, "Enhancing clinical concept extraction with distributional semantics," *J. of Biomedical Informatics*, vol. 45, no. 1, pp. 129–140, 2012.

[30] Y. Li, S. Lipsky Gorman, and N. Elhadad, "Section classification in clinical notes using supervised hidden markov model," in *Proceedings of the 1st ACM International Health Informatics Symposium*. ACM, 2010, pp. 744–750.

[31] R. Pivovarov and N. Elhadad, "A hybrid knowledge-based and data-driven approach to identifying semantically similar concepts," *Journal of biomedical informatics*, vol. 45, no. 3, pp. 471–481, 2012.

[32] S. A. Hasan, X. Zhu, Y. Dong, J. Liu, and O. Farri, "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*, 2014.

[33] K. A. Spackman, K. E. Campbell, and R. A. Côté, "Snomed rt: a reference terminology for health care." in *Proceedings of the AMIA annual fall symposium*. American Medical Informatics Association, 1997, p. 640.

[34] O. Bodenreider, "The unified medical language system (umls): integrating biomedical terminology," *Nucleic acids research*, vol. 32, no. suppl 1, pp. D267–D270, 2004.

[35] C. P. Langlotz, "Radlex: a new method for indexing online educational materials 1," 2006.

[36] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. A. Celi, and R. G. Mark, "Mimic-iii, a freely accessible critical care database," *Scientific data*, vol. 3, 2016.

[37] M. Q. Stearns, C. Price, K. A. Spackman, and A. Y. Wang, "Snomed clinical terms: overview of the development process and project status." in *Proceedings of the AMIA Symposium*. American Medical Informatics Association, 2001, p. 662.

[38] L. M. Schriml, C. Arze, S. Nadendla, Y.-W. W. Chang, M. Mazaitis, V. Felix, G. Feng, and W. A. Kibbe, "Disease ontology: a backbone for disease semantic integration," *Nucleic acids research*, vol. 40, no. D1, pp. D940–D946, 2012.