Improving Clinical Diagnosis Inference through Integration of Structured and Unstructured Knowledge

Yuan Ling, Yuan An
College of Computing & Informatics
Drexel University
Philadelphia, PA, USA
{yl638,ya45}@drexel.edu

Sadid A. Hasan
Artificial Intelligence Laboratory
Philips Research North America
Cambridge, MA, USA
sadid.hasan@philips.com

Abstract

This paper presents a novel approach to the task of automatically inferring the most probable diagnosis from a given clinical narrative. Structured Knowledge Bases (KBs) can be useful for such complex tasks but not sufficient. Hence, we leverage a vast amount of unstructured free text to integrate with structured KBs. The key innovative ideas include building a concept graph from both structured and unstructured knowledge sources and ranking the diagnosis concepts using the enhanced word embedding vectors learned from integrated sources. Experiments on the TREC CDS and HumanDx datasets showed that our methods improved the results of clinical diagnosis inference.

1 Introduction and Related Work

Clinical diagnosis inference is the problem of automatically inferring the most probable diagnosis from a given clinical narrative. Many health-related information retrieval tasks can greatly benefit from the accurate results of clinical diagnosis inference. For example, in recent Text REtrieval Conference (TREC) Clinical Decision Support track (CDS[1]), diagnosis inference from medical narratives has improved the accuracy of retrieving relevant biomedical articles (Roberts et al., 2015; Hasan et al., 2015; Goodwin and Harabagiu, 2016).

Solutions to the clinical diagnostic inferencing problem require a significant amount of inputs from domain experts and a variety of sources (Ferrucci et al., 2013; Lally et al., 2014). To address such complex inference tasks, researchers (Yao and Van Durme, 2014; Bao et al., 2014; Dong et al., 2015) have utilized structured KBs that store relevant information about various entity types and relation triples. Many large-scale KBs have been constructed over the years, such as WordNet (Miller, 1995), Yago (Suchanek et al., 2007), Freebase (Bollacker et al., 2008), DBpedia (Auer et al., 2007), NELL (Carlson et al., 2010), UMLS Metathesaurus (Bodenreider, 2004) etc. However, using KBs alone for inference tasks (Bordes et al., 2014) has certain limitations such as incompleteness of knowledge, sparsity, and fixed schema (Socher et al., 2013; West et al., 2014).

On the other hand, unstructured textual resources such as free texts from Wikipedia generally contain more information than structured KBs. As a supplementary knowledge to mitigate the limitations of structured KBs, unstructured text combined with structured KBs provides improved results for related tasks, for example, clinical question answering (Miller et al., 2016). For processing text, word embedding models (e.g. skip-gram model (Mikolov et al., 2013b; Mikolov et al., 2013a)) can efficiently discover and represent the underlying patterns of unstructured text. Word embedding models represent words and their relationships as continuous vectors. To improve word embedding models, previous works have also successfully leveraged structured KBs (Bordes et al., 2011; Weston et al., 2013; Wang et al., 2014; Zhou et al., 2015; Liu et al., 2015).

Motivated by the superior power of the integration of structured KBs and unstructured free text, we propose a novel approach to clinical diagnosis inference. The novelty lies in the ways of integrating structured KBs with unstructured text. Experiments showed that our methods improved clinical diagnosis inference from different aspects (Section 5.4). Previous work on diagnosis inference from clinical narratives either formulates the problem as a medical literature retrieval task (Zheng and Wan, 2016; Balaneshin-kordan and Kotov, 2016) or as a multiclass multilabel classi-

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

Proceedings of the 1st Workshop on Sense, Concept and Entity Representations and their Applications, pages 31–36,
Valencia, Spain, April 4 2017.

fication problem in a supervised setting (Hasan et al., 2016; Prakash et al., 2016). To the best of our knowledge, there is no work on diagnoses inference from clinical narratives conducted in an unsupervised way. Thus, we build such baselines for this task.

2 Overview of the Approach

Our approach includes four steps in general: 1) extracting source concepts, q, from clinical narratives, 2) iteratively identifying corresponding evidence concepts, a, from KBs and unstructured text, 3) representing both source and evidence concepts in a weighted graph via a regularizer-enhanced skip-gram model, and 4) ranking the relevant evidence concepts (i.e. diagnoses) based on their association with the source concepts, $S(q, a)$ (computed by weighted dot product of two vectors), to generate the final output. Figure 1 shows the overview using an illustrative example.

Given source concepts as input, we build an edge-weighted graph representing the connections among all the concepts by iteratively retrieving evidence concepts from both KBs and unstructured text. The weights of the edges represent the strengths of the relationships between concepts. Each concept is represented as a word embedding vector. We combine all the source concept vectors into a single vector representing a clinical scenario. Source concepts are differentiated according to the weighting scheme in Section 4.2. Evidence concepts are also represented as vectors and ranked according to their relevance to the source concepts. For each clinical case, we find the most probable diagnoses from the top-ranked evidence concepts.

3 Knowledge Sources of Evidence Concepts

In this study, we use UMLS Metathesaurus (Bodenreider, 2004) and Freebase (Bollacker et al., 2008) as the structured KBs. Both KBs provide semantic relation triples in the following format: <concept1, relation, concept2>. We select UMLS relation types that are relevant to the problem of clinical diagnosis inference. These types include disease-treatment, disease-prevention, disease-finding, sign or symptom, causes etc. Freebase contains a large number of triples from multiple domains. We select 61,243 triples from freebase that are classified as medicine relation types. There are 19 such relation types in total. Most of them fall under the "medicine.disease" category.

For unstructured text, we use articles from Wikipedia and MayoClinic corpus as the supplementary knowledge source. Important clinical concepts mentioned in a Wikipedia/MayoClinic page can serve as a critical clue to a clinical diagnosis. For example, in Figure 1, we see that "dyspnea", "shortness of breath", "tachypnea" etc. are the related signs and symptoms of the "Pulmonary Embolism" diagnosis. We select 37,245 Wikipedia pages under the clinical diseases and medicine category in this study. Most of the page titles represent disease names. In addition, MayoClinic[2] disease corpus contains 1,117 pages, which include sections of Symptoms, Causes, Risk Factors, Treatments and Drugs, Prevention, etc.

4 Methodology

4.1 Building Weighted Concept Graph

Both the source and the evidence concepts are represented as nodes in a graph. A clinical case is represented as a set of source concept nodes: $q = \{q_1, q_2, \ldots\}$. We build a weighted concept graph from source concepts using Algorithm 1.

Algorithm 1: Build Concept Graph

```
   Input  : source concept nodes q
   Output: graph G = (V, E)
1  S = q and V = q;
2  while S ≠ ∅ do
3      for each q_i in S do
4          if distance(q_i, q) > 2 then
5              continue;
6          end
7          if triple < q_i, r, a_j > in KBs then
8              w_ij = 1;
9              e = (q_i, a_j) and e.value = w_ij;
10             insert a_j to V and S;
11             insert e to E;
12         end
13         Use q_i as query, search in Unstructured Text Corpora, get
             Result R;
14         for each page-similarity pair (p, s_ij) in R do
15             e = (q_i, title(p)) and e.value = s_ij;
16             insert title(p) to V and S;
17             insert e to E;
18         end
19         remove q_i from S;
20     end
21 end
```

Two kinds of evidence concept nodes are added to the graph: 1) the entities from KBs (UMLS and Freebase) (step 7-12 in Algorithm 1), and 2) the entities from unstructured text pages (step 13-18). If there exists a triple $< q_i, r, a_j >$ in KBs, where

[2]http://www.mayoclinic.org/diseases-conditions

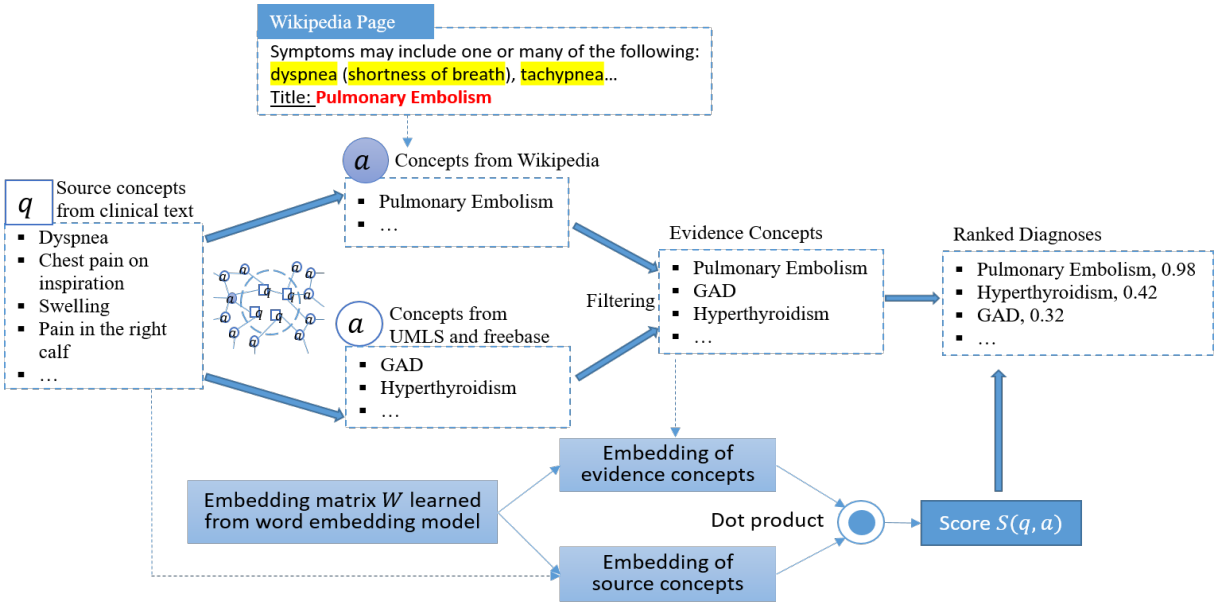

Figure 1: Overview of our system.

r refers to a relation, an edge is used to connect node q_i and node a_j. w_{ij} represents the weight for that edge, and let $w_{ij} = 1$, if the corresponding triple occurs at least once. Due to the incompleteness of the KBs, there may exist multiple missing connections between a potential evidence concept a_j and a source concept q_i. Unstructured knowledge from Wikipedia and MayoClinic can replenish these missing connections. For each page p, the page title represents an evidence concept a_j. We use each source concept q_i as a query, and page p as a document, and then calculate a query-document similarity to measure the edge weight w_{ij} between node a_j and node q_i. We only take evidence concepts as all nodes connected to source concepts in a distance of at most 2 (step 4-6).

4.2 Representing Clinical Case

We combine the source concepts q and get a single vector v_q to represent the clinical case narrative. The source concepts from narratives for clinical diagnosis inference should be differentiated. Some source concepts are major symptoms for a diagnosis, while others are less critical. These major source concepts should be identified and given higher weight values. We develop two kinds of weighting schema for the differential expression of the source concepts. The source concept is represented as $v_q = \frac{1}{N}\sum_{q_i \in q} \gamma_i v_{q_i}$. N is the total number of source concepts. v_{q_i} is the vector representation for one source concept q_i.

(1) A longer concept usually convey more information (e.g. *malar rash* vs. *rash*), so it should be given more weights. We define this weight value as $\gamma_1 = \#Words\,in\,Concept$.

(2) For some commonly seen concepts (e.g. *fever*), usually, there are more edges connected to them. Sometimes, a common concept is less important for diagnosis inference, while some unique concepts are critical to infer a specific diagnosis. We define this weight value for each concept as $\gamma_2 = \frac{1}{\#Connected\,Edges}$. A higher weight value means the source concept is more unique.

4.3 Inferring Concepts for Diagnosis

Extracting Potential Evidence Concepts: From source concept nodes q, we find their connected concepts in the graph as evidence concepts. Traversing all edges in a graph is computationally expensive and often unnecessary for finding potential diagnoses. The solution is to use a subgraph. We follow the idea proposed in Bordes et al. (2014). The evidence concepts are defined as all nodes connected to source concepts in a distance of at most 2.

Ranking Evidence Concepts: We rank each evidence concept a' according to its matching score $S(q, a')$ to the source concepts. The matching score $S(q, a')$ is a dot product of embedding representation of the evidence concept a' and the source concept q by taking the edge weights w_{ij} into consideration. $S(q, a') = w_{ij} v_{a'} \cdot v_q$. $v_{a'}$ and

v_q are embedding representations for a' and q. The embedding $E \in R^{k \times N}$ for concepts are trained using embedding models (Section 4.4). N is the total number of concepts and k is the predefined dimensions for the embedding vector. Each concept in the graph can find a k dimensional vector representation in E. For a set of source concepts and evidence concepts $A(q)$, the top-ranked evidence concept can be computed as:

$$a = argmax_{(a' \in A(q))} S(q, a') \quad (1)$$

4.4 Word Embedding Models

We use the skip-gram model as the basic model. The skip-gram model predicts surrounding words $w_{t-c}, \ldots, w_{t-1}, w_{t+1}, \ldots, w_{t+c}$ given the current center word w_t. We further enhance the skip-gram model by adding a graph regularizer. Given a sequence of training words $w_1, w_2, \ldots, w_T$, the objective function is:

$$J = \max \frac{1}{T} \sum_{t=1}^{T} (1-\lambda) \sum_{-c \leq j \leq c, j \neq 0} \log p(w_{t+j}|w_t) - \lambda \sum_{r=1}^{R} D(v_t, v_r), \quad (2)$$

where v_t and v_r are the representation vectors for word w_t and word w_r. λ is a parameter to leverage the graph regularizer and original objective. Suppose, word w_t is mentioned having relations with a set of other words $w_r, r \in \{1, \ldots, R\}$ in KBs. The graph regularizer $\lambda \sum_{r=1}^{R} D(v_t, v_r)$ integrates extra knowledge about semantic relationships among words within the graph structure. $D(v_t, v_r)$ represents the distance between v_t and v_r. In our experiments, the distance between two concepts is measured using KL-Divergence. $D(v_t, v_r)$ can be calculated using any other types of distance metrics. By minimizing $D(v_t, v_r)$, we expect if two concepts have a close relation in KBs, their vector representations will also be close to each other.

5 Experiments

5.1 Datasets for Clinical Diagnosis Inference

Our first dataset is from the 2015 TREC CDS track (Roberts et al., 2015). It contains 30 topics, where each topic is a medical case narrative that describes a patient scenario. Each case is associated with the ground truth diagnosis. We use MetaMap[3] to extract the source concepts from a narrative and then manually refine them to remove redundancy.

[3]https://metamap.nlm.nih.gov/

Our second dataset is curated from HumanDx[4], a project to foster integrating efforts to map health problems to their possible diagnoses. We curate diagnosis-findings relationships from HumanDx and create a dataset with 459 diagnosis-findings entries. Note that, the findings from this dataset are used as the given source concepts for a clinical scenario.

5.2 Training Data for Word Embeddings

We curate a biomedical corpus of around 5M sentences from two data sources: PubMed Central[5] from the 2015 TREC CDS snapshot[6] and Wikipedia articles under the "Clinical Medicine" category[7]. After sentence splitting, word tokenization, and stop words removal, we train our word embedding models on this corpus. UMLS Metathesaurus and Freebase are used as KBs to train the graph regularizer. We use stochastic gradient descent (SGD) to maximize the objective function and set the parameters empirically.

5.3 Results

We use Mean Reciprocal Rank (MRR) and Average Precision at 5 (P@5) to evaluate our models. MRR is a statistical measure to evaluate a process that generates a list of possible responses to a sample of queries, ordered by probability of correctness. Average P@5 is calculated as precision at top 5 predicted results divided by the total number of topics. Since our dataset only has one correct diagnosis for each topic, all results have poor Average P@5 scores.

Table 1 presents the results for our experiments. We report two baselines: *Skip-gram* refers to the basic word embedding model, and *Skip-gram** refers to the graph-regularized model using KBs. We also show the results for using different unstructured knowledge sources and different weighting schema. We can see that the best scores are obtained by the graph-regularized models with both the unstructured knowledge sources with variable weighting schema (Section 4.2).

5.4 Discussion

Unstructured text is a critical supplement: We analyze the source concepts and the corresponding evidence concepts for CDS topics, and investigate

[4]https://www.humandx.org/

[5]https://www.ncbi.nlm.nih.gov/pmc/

[6]http://www.trec-cds.org/2015.html#documents

[7]https://en.wikipedia.org/wiki/Category:Clinical_medicine

	TREC CDS		HumanDx	
Method	**MRR**	**Average P@5**	**MRR**	**Average P@5**
Baselines				
Skip-gram	21.66	8.88	18.56	5.08
Skip-gram*	22.60	8.88	18.63	5.15
Skip-gram* + Different Unstructured Text Datasets				
Wikipedia	26.01	8.96	19.42	5.76
MayoClinic	32.64	9.52	19.46	5.80
Both	32.29	9.60	19.12	5.76
Skip-gram* + Both Text Datasets + Different Weights				
γ_1	32.22	10.40	**21.09**	5.88
γ_2	**32.77**	**12.00**	20.86	**5.93**

Table 1: Evaluation results.

the origin of the correct diagnoses. 70% of the correct diagnoses can be inferred from Wikipedia, 60% of the correct diagnoses from MayoClinic, 56% of the correct diagnoses from Freebase, and only 7% are from UMLS. Hence, Wikipedia and MayoClinic are very important sources for finding the correct diagnoses.

Source concepts should be differentiated: In clinical narratives, some concepts are more critical than others for the clinical diagnosis inference. We developed two weighting schema to assign higher weight values to more important concepts. The results in Table 1 show that differentiating the source concepts with different weight values has a large impact on the model performance.

Enhanced skip-gram is better: We propose the enhanced skip-gram model by using a graph regularizer to integrate the semantic relationships among concepts from KBs. Experimental results show that diagnosis inference is improved by using word embedding representations from the enhanced skip-gram model.

6 Conclusion

We proposed a novel approach to the task of clinical diagnosis inference from clinical narratives. Our method overcomes the limitations of structured KBs by making use of the integrated structured and unstructured knowledge. Experimental results showed that the enhanced skip-gram model with differential expression of source concepts improved the performance on two benchmark datasets.

References

Sören Auer, Christian Bizer, Georgi Kobilarov, Jens Lehmann, Richard Cyganiak, and Zachary Ives. 2007. Dbpedia: A nucleus for a web of open data. In *The semantic web*, pages 722–735. Springer.

Saeid Balaneshin-kordan and Alexander Kotov. 2016. 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*, pages 241–250. ACM.

Junwei Bao, Nan Duan, Ming Zhou, and Tiejun Zhao. 2014. Knowledge-based question answering as machine translation. *Cell*, 2(6).

Olivier Bodenreider. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research*, 32(suppl 1):D267–D270.

Kurt Bollacker, Colin Evans, Praveen Paritosh, Tim Sturge, and Jamie Taylor. 2008. Freebase: a collaboratively created graph database for structuring human knowledge. In *Proceedings of the 2008 ACM SIGMOD international conference on Management of data*, pages 1247–1250. ACM.

Antoine Bordes, Jason Weston, Ronan Collobert, and Yoshua Bengio. 2011. Learning structured embeddings of knowledge bases. In *Conference on artificial intelligence*.

Antoine Bordes, Sumit Chopra, and Jason Weston. 2014. Question answering with subgraph embeddings. *arXiv preprint arXiv:1406.3676*.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R Hruschka Jr, and Tom M Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*, volume 5, page 3.

Li Dong, Furu Wei, Ming Zhou, and Ke Xu. 2015. Question answering over freebase with multi-column convolutional neural networks. In *Proceedings of Association for Computational Linguistics*, pages 260–269.

David Ferrucci, Anthony Levas, Sugato Bagchi, David Gondek, and Erik T Mueller. 2013. Watson: beyond jeopardy! *Artificial Intelligence*, 199:93–105.

Travis R. Goodwin and Sanda M. Harabagiu. 2016. Medical question answering for clinical decision

support. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 297–306. ACM.

Sadid A. Hasan, Yuan Ling, Joey Liu, and Oladimeji Farri. 2015. Using neural embeddings for diagnostic inferencing in clinical question answering. In *TREC*.

Sadid A. Hasan, Siyuan Zhao, Vivek Datla, Joey Liu, Kathy Lee, Ashequl Qadir, Aaditya Prakash, and Oladimeji Farri. 2016. Clinical question answering using key-value memory networks and knowledge graph. In *TREC*.

Adam Lally, Sugato Bachi, Michael A. Barborak, David W. Buchanan, Jennifer Chu-Carroll, David A. Ferrucci, Michael R. Glass, Aditya Kalyanpur, Erik T. Mueller, J. William Murdock, et al. 2014. Watsonpaths: scenario-based question answering and inference over unstructured information. *Yorktown Heights: IBM Research*.

Quan Liu, Hui Jiang, Si Wei, Zhen-Hua Ling, and Yu Hu. 2015. Learning semantic word embeddings based on ordinal knowledge constraints. In *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (ACL-IJCNLP)*, pages 1501–1511.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013a. Efficient estimation of word representations in vector space. *arXiv preprint arXiv:1301.3781*.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013b. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. *CoRR*, abs/1606.03126.

George A. Miller. 1995. Wordnet: a lexical database for english. *Communications of the ACM*, 38(11):39–41.

Aaditya Prakash, Siyuan Zhao, Sadid A. Hasan, Vivek Datla, Kathy Lee, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Condensed memory networks for clinical diagnostic inferencing. *arXiv preprint arXiv:1612.01848*.

Kirk Roberts, Matthew S. Simpson, Ellen Voorhees, and William R. Hersh. 2015. Overview of the trec 2015 clinical decision support track. In *TREC*.

Richard Socher, Danqi Chen, Christopher D. Manning, and Andrew Ng. 2013. Reasoning with neural tensor networks for knowledge base completion. In *Advances in Neural Information Processing Systems*, pages 926–934.

Fabian M. Suchanek, Gjergji Kasneci, and Gerhard Weikum. 2007. Yago: a core of semantic knowledge. In *Proceedings of the 16th international conference on World Wide Web*, pages 697–706. ACM.

Zhen Wang, Jianwen Zhang, Jianlin Feng, and Zheng Chen. 2014. Knowledge graph and text jointly embedding. In *EMNLP*, pages 1591–1601. Citeseer.

Robert West, Evgeniy Gabrilovich, Kevin Murphy, Shaohua Sun, Rahul Gupta, and Dekang Lin. 2014. Knowledge base completion via search-based question answering. In *Proceedings of the 23rd international conference on World wide web*, pages 515–526. ACM.

Jason Weston, Antoine Bordes, Oksana Yakhnenko, and Nicolas Usunier. 2013. Connecting language and knowledge bases with embedding models for relation extraction. *arXiv preprint arXiv:1307.7973*.

Xuchen Yao and Benjamin Van Durme. 2014. Information extraction over structured data: Question answering with freebase. In *ACL (1)*, pages 956–966. Citeseer.

Ziwei Zheng and Xiaojun Wan. 2016. Graph-based multi-modality learning for clinical decision support. In *Proceedings of the 25th ACM International on Conference on Information and Knowledge Management*, pages 1945–1948. ACM.

Guangyou Zhou, Tingting He, Jun Zhao, and Po Hu. 2015. Learning continuous word embedding with metadata for question retrieval in community question answering. In *Proceedings of ACL*, pages 250–259.