Learning Conceptual-Contextual Embeddings for Medical Text

Xiao Zhang,[1*] **Dejing Dou,**[3,4] **Ji Wu**[1,2]
[1]Department of Electronic Engineering, Tsinghua University
[2]Institute for Precision Medicine, Tsinghua University
[3]Department of Computer and Information Science, University of Oregon
[4]Baidu Research
xzhang19@mails.tsinghua.edu.cn
dou@cs.uoregon.edu, doudejing@baidu.com
wuji_ee@mail.tsinghua.edu.cn

Abstract

External knowledge is often useful for natural language understanding tasks. We introduce a contextual text representation model called Conceptual-Contextual (CC) embeddings, which incorporates structured knowledge into text representations. Unlike entity embedding methods, our approach encodes a knowledge graph into a context model. CC embeddings can be easily reused for a wide range of tasks in a similar fashion to pre-trained language models. Our model effectively encodes the huge UMLS database by leveraging semantic generalizability. Experiments on electronic health records (EHRs) and medical text processing benchmarks showed our model gives a major boost to the performance of supervised medical NLP tasks.

Introduction

External knowledge is often useful for language understanding tasks. Especially in specialized domains like medicine, it is unlikely to attain human-level performance in text understanding without referring to external domain knowledge. Ontologies and knowledge graphs are the most common forms of domain knowledge, but due to their structured nature, it is not straightforward to incorporate them with representation-based neural models.

Current approaches usually bridge text and knowledge graphs with retrieval. Triplets or entities are retrieved based on occurrences of the text tokens in the entity descriptions. After retrieval, triplets can be treated as text sequences and be provided to the model as an extra input (Mihaylov and Frank 2018). Another method is to use the corresponding entity embeddings from a graph embedding model trained on knowledge graphs (Huang et al. 2019). However one still needs to deal with the aligning issue between entity embeddings and text representations.

In this paper, we take a novel approach which takes external knowledge into the realm of text representation learning. Word embeddings models like skip-gram (Mikolov et al. 2013a) and contextual embedding models like BERT (Devlin et al. 2018) have proved the crucial role of good text representations in NLP tasks. Our model aims to incorporate external knowledge into text representations, which makes it easy to apply external knowledge and makes it robust to variations of expression in text.

Our model, which we termed **C**onceptual-**C**ontextual (CC) Embeddings, is a contextual text representation model similar to BERT. Instead of providing general text representations, CC embeddings are specifically designed to be "concept aware." The model is trained to recognize concept and entity names in text and produce representations of those concepts and entities. Knowledge from knowledge graphs is encoded in the representations, which can be easily utilized in NLP tasks. Like other contextual representation models, CC embedding model can be used to generate embeddings as features or fine-tuned for a supervised learning task.

The rest of the paper is organized as follows: we first formulate our approach and discuss why it is particularly relevant for the medical domain. Then we detail our model and the process of encoding a large knowledge graph into contextual representations. Finally we evaluate on several tasks to validate the effects of our CC embeddings.

Methodology

Model

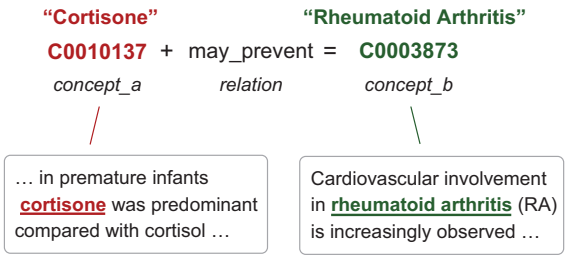

Figure 1: Encoding concept mentions in text

The core component of CC embedding model is an encoder which encodes structured knowledge. The encoder takes a written form of a concept as input, and outputs a vector representation of the concept. The idea is illustrated in Figure 1.

[*]Work done while visiting the University of Oregon

In this example, the encoder encodes a mention of a concept within a piece of text, and produces a concept embedding that satisfies a relationship defined in a knowledge graph:

$$Encode(\text{“Cortisone”}) + may\_prevent \approx Encode(\text{“Rheumatoid Arthritis”})$$

For simplicity we assume that the encoded concept embeddings and relation embeddings satisfy approximate translational relationship. Such a formulation is similar to TransE (Bordes et al. 2013), but instead of learning entity embeddings, we would like to learn an encoder that can “compose” the right concept representation from a mention found in text.

In this work we use a multi-layer bi-directional LSTM network as the encoder, similar to ELMo (Peters et al. 2018). Given an input sentence, it computes a representation vector at every word positions. The only difference is that we adopt a knowledge graph embedding objective, rather than a language modeling objective.

To embed knowledge into it, the model is trained on a graph embedding task. During training the model is exposed to a large text corpus to learn to recognize and encode concepts in text. After training, the model absorbs structured knowledge in its parameters and is ready for reuse.

Knowledge in medical KBs

Medical text processing is still quite challenging despite recent success of deep learning in other modalities like medical image and sequential measurement data. The difficulty of incorporating a large amount of domain knowledge to understand text is definitely a reason. Therefore we are interested in getting an overall picture of medical domain knowledge from the perspective of NLP, and find out to what extent can representation models capture structured knowledge.

Another reason we are interested in the medical domain is that there exists a good collection of well structured domain knowledge, maintained in the form of multiple ontologies and knowledge bases (KBs). And more importantly, a large portion of the knowledge base entries has central attributes (like concept names, relation names) expressed in written language, rather than merely symbols and proper nouns. This makes text processing extremely relevant in utilizing the domain knowledge.

Diving into one of the medical KBs, one can summarize the typical structured information there into two categories:

- **Language inferable (LI) knowledge**: these are triplets where the relation between two concepts can be at least partially inferred from the name of the concepts, e.g.,

  Pulmonary Fibrosis is_sibling *Cystic Disease of Lung*

  The relation can be inferred as likely because “pulmonary” means “relating to the lungs”.

- **Non-language inferable (Non-LI) knowledge**: facts that are independent of the meanings of the textual expression, for example:

  Iodine 10 mg/ml Topical Solution is_a *Ultracare Oral Product*

  In this example *Ultracare* is a brand name, and it is impossible to infer whether the relation holds solely based on the above text.

Research on knowledge bases usually do not make such distinctions and KB embedding models treat each concept as an individual entity. For text understanding, however, the first category deserves special attention because it represents **generalizable** knowledge. Such knowledge can be encoded in word representations or context representations, which can generalize to unseen expressions of concepts.

First, the knowledge can be generalized to different ways of writing the same concept. Medical concepts often have different names in different KBs, for example “*Enamel Dysplasia*,” “*Enamel Agenesis*,” and “*Enamel Hypoplasia*” can refer to the same concept. Second, knowledge can also generalize from one concept to other concepts, such as from “*Pulmonary hypertension*” to “*Pulmonary Fibrosis*.” As we will show in our experiments, exploiting such generalizations is a key to learning good medical concept embeddings.

Unlike entity embeddings, knowledge in text representations is generalizable and also directly available for neural NLP models and it can help text understanding in general. In the medical domain, applications that involve processing text such as doctor notes and electrical medical records could benefit from a representation model that incorporate generalizable domain knowledge.

Related Work

KB embedding models. In recent years a number of KB embedding models have been proposed that aim at learning entity embeddings on a knowledge graph (Cai, Zheng, and Chang 2018). Some models make use of textual information in KBs to improve entity embeddings, like using textual descriptions of entities as complement to triplet modeling (Wang and Li 2016; Xiao et al. 2017), or jointly learning structure-based embeddings and description-based embeddings (Xie et al. 2016; Xu et al. 2017). The latter approach learns an encoder which is similar to our work, but the encoder is only used to encode entity descriptions. These approaches are mainly concerned with KB representations rather than text processing. Using text also allows for inductive and zero-shot (Yang, Cohen, and Salakhutdinov 2016) entity representations, which is also a feature of our model.

Word embedding models. One way to incorporate external knowledge into text representations is by learning knowledge-enhanced word embeddings. Some use joint-objectives to train word embeddings to simultaneously satisfy co-occurrence relationships and external constraints, like in (Yu and Dredze 2014) and (Bian, Gao, and Liu 2014). Others rely on retrofitting, which fine-tunes word vectors in conventional word embeddings to reflect external knowledge (Faruqui et al. 2015; Nguyen, Schulte im Walde, and Vu 2016). (Glavaš and Vulić 2018) uses a technique called “explicit retrofitting” to learn a transformation which adds constraints to the embeddings. However, the external knowledge used in this line of work is mainly word-level lexical resources, like WordNet (Miller 1995; Liu et al. 2015), synonyms and antonyms (Nguyen, Schulte im Walde, and Vu

2016; Ono, Miwa, and Sasaki 2015). Integrating knowledge from a general knowledge graph is more difficult because entities and relations do not directly correspond to words.

Concept embedding models. In the NLP community sometimes concept embeddings are regarded as a form of phrase embeddings (Mikolov et al. 2013b), which can be learned by treating concepts as special words. One first annotate the concept mentions within a corpus, then use standard word embedding model to learn embeddings for those "special words" (Vu and Parker 2016; Shalaby, Zadrozny, and Jin 2018). In medical domain such method is widely explored with the help of automatic annotators and ontologies (De Vine et al. 2014; Finlayson, LePendu, and Shah 2014; Choi, Chiu, and Sontag 2016). (Mencía, de Melo, and Nam 2016) expands the method by also using relationships found in structured text.

Contextual representation models. Recently contextual text representation models like ELMo (Peters et al. 2018), BERT (Devlin et al. 2018) and OpenAI GPT (Radford et al. 2018; 2019) have pushed the state-of-the-art results of various NLP tasks. Language modeling on a giant corpus learns powerful representations, which provides huge benefits to supervised tasks, especially where labeled data is scarce. These models use sequential or attention networks to generate word representations in context. In the biomedical domain there is also BioBERT (Lee et al. 2019), a BERT model trained on PubMed articles that offers competitive results on medical text processing tasks. More recently some enhanced BERT models propose to mark entities in training, to make models aware of entities in text (Zhang et al. 2019; Sun et al. 2019).

Relationship to other knowledge-enhanced NLP models. Some works have explored integrating knowledge representation into a specific task, like question answering (Hao et al. 2017; Mihaylov and Frank 2018) and language inference (Chen et al. 2017). These models include network components to match entities and combine entity embeddings with input at inference time. The model design is usually specific to the task formulation, for example, a model designed on WebQuestions cannot naturally generalize to QA tasks where the answers are not restricted to be entities. By contrast, our approach encodes knowledge into a general text representation model, and no specific network structure is needed to leverage knowledge.

Conceptual-Contextual Embeddings

In this section we detail the task used to train Conceptual-Contetxual embeddings and the training scheme, also performing evaluation within a knowledge graph for analyzing the effectiveness of training.

Task

To encode the knowledge into a text representation model, we use knowledge graph embedding task, like in (Bordes et al. 2013). To show our approach is scalable to large knowledge graphs in the medical domain, we use the UMLS database (Bodenreider 2004) for learning to encode medical concept embeddings.

UMLS. The Unified Medical Language System (UMLS) (Bodenreider 2004) Metathesaurus is a large biomedical thesaurus containing concepts and relations from nearly 200 vocabularies (knowledge bases). A statistic of the database is given in Table 1. A concept in UMLS has one or more names associated with it (because different source vocabulary can name a concept differently). Relationships are given as triplets *(head_concept, relation, tail_concept)*.

Table 1: UMLS dataset statistics

Item	#
Entities (concepts)	2,983,840
Relations (general label)	14
Relations (additional label)	936
Train triplets	23,029,716
Test triplets	8059

Concept names. For each concept we take all the names associated with it. Among all the names, some are labeled as "preferred name" by UMLS. We extract the first "preferred name" as a primary name for the concept, and all the other as name variations.

Relations. Each kind of relationship in UMLS has a general label (REL) and an optional additional label (RELA). General labels describe the basic nature of the relationship (e.g., *Broader, Narrower, Child of, Qualifier of*), while the additional labels explain the relationship more exactly (e.g., *is_a, branch_of, component_of*). We use additional labels as relationship labels whenever available, and use general labels when additional labels are absent.

All the triplets are extracted from the UMLS Metathesaurus Level 0 Subset and are split into a training set T and a testing set. For each triplet in testing set, the triplet that describes the inverse relationship is removed from the training set (if found). We further removed concepts with non-latin characters in their names for more meaningful results in text-based models.

Context corpus. Learning to recognize concepts in text requires "seeing concept names in context." We prepared a corpus from PubMed citations and MIMIC-III critical care database (Johnson et al. 2016). Text is extracted from PubMed article abstracts and clinical notes in MIMIC-III health records. The corpus contains roughly 192 million sentences. We employ Apache Solr™ to index the corpus, and use stemming normalization to increase recall for retrieval.

Model and training

The model is illustrated in Figure 2. The core of the model is a multi-layer bi-directional LSTM network. We make use of BioWordVec, a pre-trained biomedical domain word embedding from (Chen, Peng, and Lu 2019) to embed text inputs.

Given a triplet (h, r, t) in the training dataset T, we first lookup the name $h_{1...n}$ (with length n) of the head concept h: the primary name of the concept is used or, with probability α, randomly replaced with one of its name variations.

Next we use the name $h_{1...n}$ as keywords to retrieve sentences from the context corpus. We keep sentences with keyword occurrences lie adjacent to each other (forming

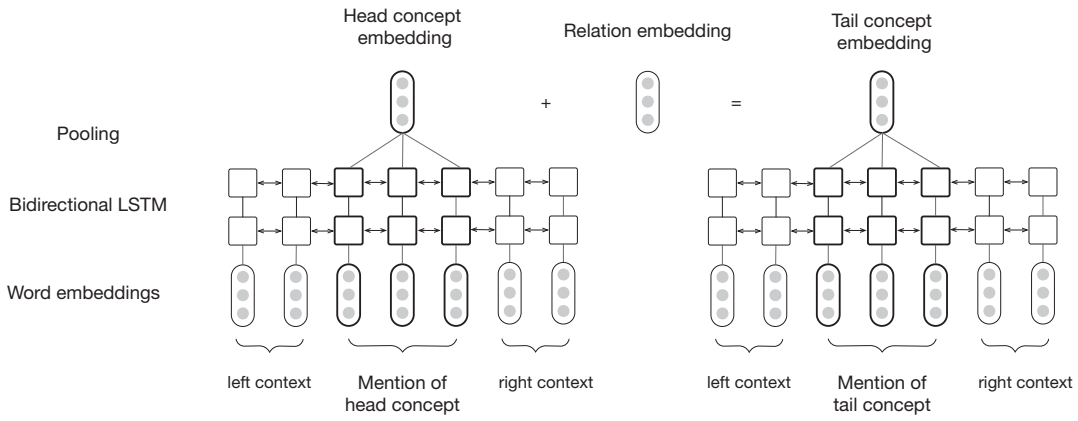

Figure 2: Training CC embedding model to embed concepts

Algorithm 1 Training CC embedding model

Require: Training set of triplets $T = \{(h, r, t)\}$, relations L and concept names $C = \{c_{1...n}\}$. Vocabulary V and word embeddings E. Context corpus $S = \{s_{1...m}\}$

loop
  for $(h, r, t) \in T$ **do**
    $(h', r, t') \leftarrow$ sample$(T, (h, r, t))$
      // sample a corrupted triplet
    for $c \in \{h, t, h', t'\}$ **do**
      $c_{1...n} \leftarrow$ lookup(C, c)
        // lookup concept names
      $c^{ct}_{1...m} \leftarrow$ retrieve$(S, c_{1...n})$
        // retrieve context sentences
      $\boldsymbol{c}^{ct}_{1...m} \leftarrow$ LSTM$(E(c^{ct}_{1...m}))$
      $\boldsymbol{c} \leftarrow$ selective_pool$(\boldsymbol{c}^{ct}_{1...m})$
      $\boldsymbol{c} \leftarrow \boldsymbol{c}/||\boldsymbol{c}||$
    end for
    Update network w.r.t.
      $\nabla[\gamma + d(\boldsymbol{h} + \boldsymbol{r}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{r}, \boldsymbol{t'})]_+$
  end for
end loop
return Trained LSTM network (including modified word embeddings)

phrases). A random sentence is selected from the top 10 ranked retrieval results as the context $h^{ct}_{1...m}$ for concept h.

The context sentence $h^{ct}_{1...m}$ (with length m) is then encoded by the LSTM network. The output sequence $\boldsymbol{c}^{ct}_{1...m}$ of the LSTM network is multiplied with a mask, which only keeps the output on the positions corresponding to the concept name in the sentence. The output is then max-pooled into a single vector $\boldsymbol{h}$ and normalized to unit length, which serves as a representation of the head concept h. The same is performed to generate a representation $\boldsymbol{t}$ of the tail concept.

Once the head and the tail concepts are encoded into vectors, we use vector addition in embedding space to model the relationship between the concepts. The formulation in this step is similar to TransE except that we use LSTM outputs in place of entity embeddings. For training the model, negative triplets (h', r, t') are sampled by replacing the head or tail with a random concept, which are then processed by the model in the same fashion. The model is trained by minimizing a margin-based ranking loss:

$$\mathcal{L} = \nabla[\gamma + d(\boldsymbol{h} + \boldsymbol{r}, \boldsymbol{t}) - d(\boldsymbol{h'} + \boldsymbol{r}, \boldsymbol{t'})]_+ \quad (1)$$

In experiments we use 200-dimensional word embeddings and 2-layer bi-directional LSTM network with also 200 dimensions. In the ranking loss $\mathcal{L}$, Euclidean norm is used in distance function d and margin $\gamma = 0.1$. Vanilla stochastic gradient descent with learning rate $l = 1.0$ is used to optimize the network. A total of 10 epochs is trained on 23 million training triplets. Note that the hyper-parameter values are largely chosen heuristically and are not sufficiently tuned, due to efficiency reasons of the LSTM network and the size of the UMLS.

Discriminative Training

To encode concept names into higher fidelity concept representations, the model needs to recognize subtle differences between terms. We add a discriminative training step for this purpose. During training, when corrupted triplets are sampled, we sample concepts with names that are similar to the true concept instead of random sampling. This is done with probability $\beta = 0.5$. For example, given concept "*Myeloid Leukemia*" as a true tail, concept "*Lymphocytic Leukemia*" would be more likely to be sampled as a corrupted tail under discriminative training. To avoid calculating the full similarity matrix between 3 million concepts, we take a crude but fast approximation: when sampling a negative concept c', we randomly choose a word w from the name $c_{1...n}$ of the true concept c, then randomly choose a concept c' that also has w in its name. The sampled negative concept c' at least shares one common word in its name with the true concept.

This sampling step increases the difficulty of negative samples by making them more similar to the true triplets and thus more challenging for ranking. This forces the model to discriminate the semantic meaning of similar named concepts. To keep the model exposed to the whole set of possible concepts, there is still $1 - \beta$ probability to sample from any concepts.

Table 2: Entity prediction results

Model	Mean rank		Mean log(rank)		Hits@10 (%)		Hits@1 (%)	
	raw	*filtered*	*raw*	*filtered*	*raw*	*filtered*	*raw*	*filtered*
TransE	213010	212298	2.90	2.90	16.7	16.7	3.3	3.3
CC-DNN	24441	23955	2.45	2.29	22.1	27.7	9.2	13.7
CC-LSTM	22888	22685	1.65	1.61	50.8	51.9	44.7	45.7
CC-LSTM (DT)	43637	43518	**1.27**	**1.22**	**64.0**	**65.4**	**56.8**	**58.7**

*DT: discriminative training

Intrinsic Evaluation

Before evaluating the learned representation on downstream tasks, we want to first analyze to what extent our model encodes structured knowledge in the UMLS, also validate the generalizability of the model.

We use the entity prediction task to measure the quality of the embeddings produced for concept names. Entity prediction is a standard task for evaluating entity embeddings, but here we only use it for analysis purposes rather than as a goal. For each triplet from the testing set we split from the UMLS, either the head or the tail is replaced with every concept in the UMLS. The true triplet is then ranked against corrupted triplets by the model. Results of ranking performance are shown in Table 2. We follow common practice to report raw and filtered ranks.

TransE is listed in the table as a reference because we use the same translational formula to model relationships. CC-LSTM model performs surprisingly well on entity prediction, given that it is ranking among 3 million concepts. Especially for the Hits@1 metric, which is equivalent to making the "correct" prediction. The best CC model makes the correct prediction more than half of the time, indicating its ability at fine-grained differentiation of concept semantics.

We also include a mean log(rank) metric, for a better representation of "the average ranking position." When the number of ranking candidates is extremely large, one badly ranked example could make an otherwise good "mean rank" drop a lot, making the metric less intuitive. It can be seen from the mean log(rank) column, roughly, the ranks of CC-LSTM model are generally of order 10^1-10^2 and the ranks of TransE are generally of order 10^3.

In place of the LSTM network, we experimented with using DNN to generate concept embeddings, but results are far inferior. Contextual information is important to correctly represent a concept based on its name. Discriminative training also substantially enhanced the performance of CC-LSTM model.

Table 3: Performance on language-inferable and non-language-inferable knowledge

	# of examples	Hits@10 (%)
LI	76	77.6
Non-LI	24	20.8
Total	100	64.0

Break-down analysis To measure the effect of semantic generalizability on model performance, and to understand the performance gap between the CC model and TransE, we first sampled 100 examples from the testing set, and labeled them to two categories: Language-inferable (LI) and Non-language-inferable (Non-LI), following our previous definition. Performance of the CC model on each category is shown in Table 3. First we observe that 3/4 of the triplets contain knowledge that can be inferred from text. This shows that in medical knowledge graphs, a majority of structured knowledge can potentially be carried by text representations. Making use of concept names can be difference-making in medical knowledge embeddings. In the case of the CC model, on LI type examples it gets to 77 percent hit at top10, while for Non-LI type the performance is much lower and is on par with the TransE model.

Table 5: Error analysis by category

Category	Percentage
Policy	3.0
Long name	2.5
UNK	7.5
SIB	5.0
Facts (Non-LI)	14.5
Other errors	5.5
Correct	62.0

On LI type knowledge the model is still quite far from perfect. We summarize the reasons for model failure in Table 5. After examining 200 examples in the testing set, we arrive at five common categories of difficult examples, which are:

- Policy: this category is for triplets describing knowledge on medical policy or administration. These are not medical knowledge in the very strict sense, and the poor result could possibly be attributed to domain mismatch of pre-trained word embeddings.
- Long name: the name of one of the concepts in the triplet is longer than 10 words. Because we truncate long names to 10 words for faster training, some information is missing from the input.
- UNK: one of the concepts has more than half of out-of-vocabulary words in its name. This typically makes the concept indiscernible to the model.
- SIB: this category of triplets all have *is_sibling* relationship. The model seems to have some difficulty judging whether some closely related concepts are at the same level in the hierarchy.

Table 4: Readmission prediction performance

Model	Acc	Pre-0	Pre-1	Re-0	Re-1	A.R.	A.P.
(Lin et al. 2018)	0.698	0.916	0.367	0.687	0.742	0.791	0.513
LSTM	0.840	0.956	0.366	0.859	0.704	0.794	0.600
CC-LSTM	0.848	0.978	0.321	0.854	0.786	**0.804**	0.613

*Acc: Accuracy, Pre: Precision, Re: Recall, A.R: Area under ROC, A.P: area under PRC

- Facts (Non-LI): factual knowledge that is not inferable form text.

These categories account for most of the errors of the CC model on the entity prediction task. Except for the Non-LI category, these errors are in principle resolvable with proper modifications to the model. Overall, the CC model captures language-inferable medical knowledge quite effectively, and next we will show it serves as a useful text representation.

Downstream Applications

As a contextual text representation model, the CC embedding model can be fine-tuned to various NLP tasks. By doing so the CC embeddings introduce concept awareness and external structured knowledge into the task model. We first present results on two real-world medical tasks then on another medical NLP benchmark task.

MIMIC-III and Derived Datasets

The MIMIC-III Critical Care Database (Johnson et al. 2016; Goldberger et al. 2000) is a large database of electronic health records (EHRs) of over 40,000 patients in Intensive Care Unit. Various kinds of numerical and report data is provided. In this study we are only concerned with textual data in EHRs. Specifically, we use the "Discharge Summary" included in each ICU admission, which is a note written by doctors when the patient is discharged from ICU. Here is a snippet from one such note:

```
This is a 65 year old female with re-
cent history of C. diff colitis (06')
and recent mult abx use for UTI/PNA
past couple months who presented after
a syncopal episode in the setting of
diarrhea/dehydration ...
```

Data pre-processing follows (Harutyunyan et al. 2017) and (Lin et al. 2018): after data screening there are 35,334 patients and 48,393 ICU stays. The patients are split into training (80%), validation (10%) and testing (10%) sets with 5-fold cross validation.

In the following two tasks, we add a pooling and a linear layer on top of the CC-LSTM model to perform classification. A plain LSTM classifier with identical structure is used as a baseline. All models use BioWordVec as word embeddings. We use early-stopping on validation set to select the best model. Reported results are averages over 5-fold splits.

Readmission Prediction

Unplanned ICU readmission rate is an important metric in hospital operation. Readmission prediction can help identify high-risk patients and reduce premature discharge (Kansagara et al. 2011). Our model predicts whether a patient is likely to be readmitted into ICU within 30 days, upon his/her discharge.

In Table 4, we present our model results and state-of-the-art result from (Lin et al. 2018). Lin et al. uses chart events, demographic information and diagnosis as input to a LSTM+CNN model. We only use written note text and none of the numerical and time-series information. The primary metric *area under ROC* clearly shows that the CC model produces a performance boost over the baseline and surpassed state-of-the-art results.

Table 6: Post-discharge mortality prediction performance

Model	30-day A.R.	1-year A.R.
(Ghassemi et al. 2014)	0.80	0.77
(Ghassemi et al. 2014) (retrospective)	0.82	0.81
(Grnarova et al. 2016)	0.858	0.853
LSTM	0.823	0.820
CC-LSTM	0.839	0.837

Mortality Prediction

In this task we predict post ICU discharge mortality. Mortality prediction can help make better management and treatment decisions in costly ICU operations (Pirracchio et al. 2015). Table 6 gives the prediction results of patient mortality within 30-day and 1-year after discharge. Note that like in the previous task, results from other works are listed mainly for reference rather than direct comparison, for these models use different information from EHR as input. Although not matching with state-of-the-art, the performance gain of the CC model over an LSTM model is consistent.

Medical Language Inference

Natural language inference (NLI) is a task determining the entailment relationship between two pieces of text. We use the MedNLI dataset (Romanov and Shivade 2018) to evaluate language inference in the medical domain. Original dataset contains 11232 sentence pairs for training and 1395 and 1422 pairs for development and testing. Results are listed in Table 7.

We implemented the ESIM model (Harutyunyan et al. 2017) for NLI task, which consists of an LSTM encoder layer and an LSTM composition layer. CC-ESIM simply replaces the LSTM network in encoding layer with our trained CC-LSTM. The performance gain indicates the CC embeddings successfully introduces external knowledge into the

Table 7: Performance on medical language inference

Model	Dev Acc	Test Acc
ESIM (Romanov and Shivade 2018)	74.4	73.1
ESIM (our implementation)	74.8	71.3
CC-ESIM	**77.1**	**75.2**

model and benefits the task.

Conclusion

We have presented Conceptual-Contextual embeddings, a contextual text representation model which introduces structured external knowledge into text representations. The effectiveness of the model is validated on the medical domain, where domain knowledge is substantially associated with text understanding. Our work serves as a bridging perspective between knowledge graph representations and unsupervised text representation models.

Future work include incorporating more powerful relationship models like TransR (Lin et al. 2015) into the CC embedding model. As our model only captures conceptual knowledge, combining CC embeddings with general representation models like BERT is also an interesting investigation. Under our formulation it is also straightforward to combine the two into a single model with multi-task learning, to further improve state-of-the-art text representation models.

Acknowledgments

This research is partially supported by the National Key Research and Development Program of China (No.2018YFC0116800) and the NSF grant CNS-1747798 to the IUCRC Center for Big Learning.

References

Bian, J.; Gao, B.; and Liu, T.-Y. 2014. Knowledge-powered deep learning for word embedding. In *Machine Learning and Knowledge Discovery in Databases*, 132–148. Springer Berlin Heidelberg.

Bodenreider, O. 2004. The unified medical language system (umls): integrating biomedical terminology. *Nucleic acids research* 32(suppl_1):D267–D270.

Bordes, A.; Usunier, N.; Garcia-Duran, A.; Weston, J.; and Yakhnenko, O. 2013. Translating embeddings for modeling multi-relational data. In *Advances in Neural Information Processing Systems 26*. 2787–2795.

Cai, H.; Zheng, V. W.; and Chang, K. C.-C. 2018. A comprehensive survey of graph embedding: Problems, techniques, and applications. *IEEE Transactions on Knowledge and Data Engineering* 30(9):1616–1637.

Chen, Q.; Zhu, X.; Ling, Z.-H.; Wei, S.; Jiang, H.; and Inkpen, D. 2017. Enhanced LSTM for natural language inference. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 1657–1668.

Chen, Q.; Peng, Y.; and Lu, Z. 2019. Biosentvec: creating sentence embeddings for biomedical texts. In *Proceedings of the 7th IEEE International Conference on Healthcare Informatics (ICHI)*.

Choi, Y.; Chiu, C. Y.-I.; and Sontag, D. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings* 2016:41.

De Vine, L.; Zuccon, G.; Koopman, B.; Sitbon, L.; and Bruza, P. 2014. Medical semantic similarity with a neural language model. In *Proceedings of the 23rd ACM International Conference on Conference on Information and Knowledge Management*, 1819–1822.

Devlin, J.; Chang, M.-W.; Lee, K.; and Toutanova, K. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Faruqui, M.; Dodge, J.; Jauhar, S. K.; Dyer, C.; Hovy, E.; and Smith, N. A. 2015. Retrofitting word vectors to semantic lexicons. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1606–1615.

Finlayson, S. G.; LePendu, P.; and Shah, N. H. 2014. Building the graph of medicine from millions of clinical narratives. *Scientific data* 1:140032.

Ghassemi, M.; Naumann, T.; Doshi-Velez, F.; Brimmer, N.; Joshi, R.; Rumshisky, A.; and Szolovits, P. 2014. Unfolding physiological state: Mortality modelling in intensive care units. In *Proceedings of the 20th ACM SIGKDD international conference on Knowledge discovery and data mining*, 75–84.

Glavaš, G., and Vulić, I. 2018. Explicit retrofitting of distributional word vectors. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 34–45.

Goldberger, A. L.; Amaral, L. A.; Glass, L.; Hausdorff, J. M.; Ivanov, P. C.; Mark, R. G.; Mietus, J. E.; Moody, G. B.; Peng, C.-K.; and Stanley, H. E. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *Circulation* 101(23):e215–e220.

Grnarova, P.; Schmidt, F.; Hyland, S. L.; and Eickhoff, C. 2016. Neural document embeddings for intensive care patient mortality prediction. *arXiv preprint arXiv:1612.00467*.

Hao, Y.; Zhang, Y.; Liu, K.; He, S.; Liu, Z.; Wu, H.; and Zhao, J. 2017. An end-to-end model for question answering over knowledge base with cross-attention combining global knowledge. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 221–231.

Harutyunyan, H.; Khachatrian, H.; Kale, D. C.; Ver Steeg, G.; and Galstyan, A. 2017. Multitask Learning and Benchmarking with Clinical Time Series Data. *arXiv preprint arXiv:1703.07771*.

Huang, X.; Zhang, J.; Li, D.; and Li, P. 2019. Knowledge graph embedding based question answering. In *Proceedings of the Twelfth ACM International Conference on Web Search and Data Mining*, 105–113.

Johnson, A. E. W.; Pollard, T. J.; Shen, L.; Lehman, L.-w. H.; Feng, M.; Ghassemi, M.; Moody, B.; Szolovits, P.; Anthony Celi, L.; and Mark, R. G. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data* 3:160035.

Kansagara, D.; Englander, H.; Salanitro, A.; Kagen, D.; Theobald, C.; Freeman, M.; and Kripalani, S. 2011. Risk Prediction Models for Hospital Readmission: A Systematic Review. *JAMA* 306(15):1688–1698.

Lee, J.; Yoon, W.; Kim, S.; Kim, D.; Kim, S.; So, C. H.; and Kang, J. 2019. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *arXiv preprint arXiv:1901.08746*.

Lin, Y.; Liu, Z.; Sun, M.; Liu, Y.; and Zhu, X. 2015. Learning

entity and relation embeddings for knowledge graph completion. In *Twenty-ninth AAAI conference on artificial intelligence*.

Lin, Y.-W.; Zhou, Y.; Faghri, F.; Shaw, M. J.; and Campbell, R. H. 2018. Analysis and prediction of unplanned intensive care unit readmission using recurrent neural networks with long short-term memory. *bioRxiv* 385518.

Liu, Q.; Jiang, H.; Wei, S.; Ling, Z.-H.; and Hu, Y. 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 (Volume 1: Long Papers)*, 1501–1511.

Mencía, E. L.; de Melo, G.; and Nam, J. 2016. Medical concept embeddings via labeled background corpora. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, 4629–4636.

Mihaylov, T., and Frank, A. 2018. Knowledgeable Reader: Enhancing Cloze-Style Reading Comprehension with External Commonsense Knowledge. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, 821–832.

Mikolov, T.; Chen, K.; Corrado, G.; and Dean, J. 2013a. Efficient estimation of word representations in vector space. In *International Conference on Learning Representations 2013*.

Mikolov, T.; Sutskever, I.; Chen, K.; Corrado, G.; and Dean, J. 2013b. Distributed representations of words and phrases and their compositionality. In *Proceedings of the 26th International Conference on Neural Information Processing Systems - Volume 2*, 3111–3119.

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

Nguyen, K. A.; Schulte im Walde, S.; and Vu, N. T. 2016. Integrating distributional lexical contrast into word embeddings for antonym-synonym distinction. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 454–459.

Ono, M.; Miwa, M.; and Sasaki, Y. 2015. Word embedding-based antonym detection using thesauri and distributional information. In *Proceedings of the 2015 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 984–989.

Peters, M. E.; Neumann, M.; Iyyer, M.; Gardner, M.; Clark, C.; Lee, K.; and Zettlemoyer, L. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, 2227–2237.

Pirracchio, R.; Petersen, M. L.; Carone, M.; Rigon, M. R.; Chevret, S.; and van der Laan, M. J. 2015. Mortality prediction in intensive care units with the super icu learner algorithm (sicula): a population-based study. *The Lancet Respiratory Medicine* 3(1):42–52.

Radford, A.; Narasimhan, K.; Salimans, T.; and Sutskever, I. 2018. Improving language understanding by generative pre-training. Technical report, OpenAI.

Radford, A.; Wu, J.; Child, R.; Luan, D.; Amodei, D.; and Sutskever, I. 2019. Language models are unsupervised multitask learners.

Romanov, A., and Shivade, C. 2018. Lessons from natural language inference in the clinical domain. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, 1586–1596.

Shalaby, W.; Zadrozny, W.; and Jin, H. 2018. Beyond word embeddings: learning entity and concept representations from large scale knowledge bases. *Information Retrieval Journal* 1–18.

Sun, Y.; Wang, S.; Li, Y.; Feng, S.; Chen, X.; Zhang, H.; Tian, X.; Zhu, D.; Tian, H.; and Wu, H. 2019. Ernie: Enhanced representation through knowledge integration. *arXiv preprint arXiv:1904.09223*.

Vu, T., and Parker, D. S. 2016. k-embeddings: Learning conceptual embeddings for words using context. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, 1262–1267.

Wang, Z., and Li, J. 2016. Text-enhanced representation learning for knowledge graph. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, 1293–1299.

Xiao, H.; Huang, M.; Meng, L.; and Zhu, X. 2017. Ssp: semantic space projection for knowledge graph embedding with text descriptions. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Xie, R.; Liu, Z.; Jia, J.; Luan, H.; and Sun, M. 2016. Representation learning of knowledge graphs with entity descriptions. In *Thirtieth AAAI Conference on Artificial Intelligence*.

Xu, J.; Qiu, X.; Chen, K.; and Huang, X. 2017. Knowledge graph representation with jointly structural and textual encoding. In *Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, IJCAI-17*, 1318–1324.

Yang, Z.; Cohen, W. W.; and Salakhutdinov, R. 2016. Revisiting semi-supervised learning with graph embeddings. In *Proceedings of the 33rd International Conference on International Conference on Machine Learning - Volume 48*, 40–48.

Yu, M., and Dredze, M. 2014. Improving lexical embeddings with semantic knowledge. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, 545–550.

Zhang, Z.; Han, X.; Liu, Z.; Jiang, X.; Sun, M.; and Liu, Q. 2019. ERNIE: Enhanced language representation with informative entities. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, 1441–1451.