NeighBERT: Medical Entity Linking Using Relation Induced Dense Retrieval

Ayush Singh (✉ ayush.singh@evicore.com)
Evernorth Health Services
Saranya Krishnamoorthy
Evernorth Health Services
John Ortega
Evernorth Health Services

Research Article

Keywords: Natural Language Processing, Knowledge Graph, Information Search and Retrieval, Deep Learning, Medical Entity Linking, Biomedical

Posted Date: September 15th, 2022

DOI: https://doi.org/10.21203/rs.3.rs-2050347/v1

NeighBERT: Medical Entity Linking Using Relation Induced Dense Retrieval

Ayush Singh[1*], Saranya Krishnamoorthy[1] and John Ortega[1]

[1]inQbator AI at eviCore Healthcare, Evernorth Health Services, Saint Louis, MO, USA.

*Corresponding author(s). E-mail(s): ayush.singh@evicore.com; Contributing authors: saranya.krishnamoorthy@evicore.com; john.ortega@evicore.com;

Abstract

One of the common tasks in clinical natural language processing is medical entity linking (MEL) which involves mention detection followed by linking the mention to an entity in a knowledge base. One reason that MEL has not been solved is due to a problem that occurs in language where ambiguous texts can be resolved to several named entities. This problem is exacerbated when processing text found in electronic health records. Recent work has shown that deep learning models based on transformers outperform previous methods on linking at higher rates of performance. We introduce NeighBERT, a custom pre-training technique which extends BERT [1] by encoding how entities are related within a knowledge graph. This technique adds relational context that has been traditionally missing in original BERT, helping resolve the ambiguity found in clinical text. In our experiments, NeighBERT improves the precision, recall and F1-score of the state of the art by 1–3 points for named entity recognition and 10–15 points for MEL on two widely known clinical datasets.

Keywords: Natural Language Processing, Knowledge Graph, Information Search and Retrieval, Deep Learning, Medical Entity Linking, Biomedical

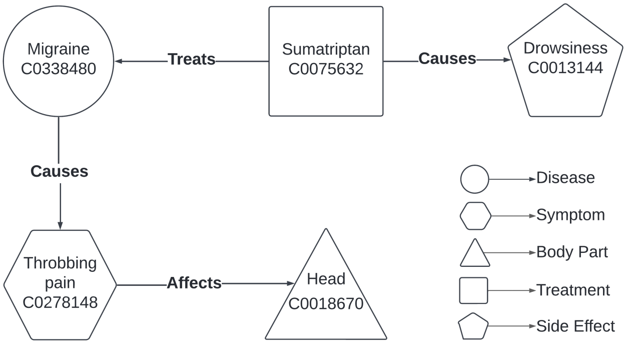

Fig. 1 An example of how different entities (shapes) are related (edges) to one another for a sentence: *A patient having throbbing pain due to Migraine was given Sumatriptan.* BERT is limited to only being able to contextualize from given text, on the other hand NeighBERT would not only give context from text but also add relational encoding e.g. even though *Sumatriptan* treats *Migraine* as evident from sentence, it also causes *Drowsiness* that can only be learned by having relational knowledge.

1 Introduction

Several tasks from the natural language processing (NLP) field are considered key to the success of recent transformer architectures like BERT [1]. While not unique to the clinical domain, two tasks based on inference are named-entity recognition (NER) and named-entity linking (NEL). Named entities in general are groups of words that represent persons, places, or companies. For clinical text, named entities identify concepts like diseases, body parts, and more. The process of discovering those entities in text and linking them to their corresponding entries in a knowledge base or term list is jointly known as medical entity linking (MEL) in the clinical domain.

MEL has remained a challenging problem to solve for a long time, even the state-of-the-art systems [2–4] find it hard to achieve good performance. This happens primarily because unlike general NEL where cardinality is low versus that of clinical which is very high e.g. UMLS has 4 million entities that can be potentially be assigned to a mention compared to less than 10 (LOC, MISC, ORG, PER) of CoNLL-2003 [5].

In this article, we present a clinical case study with a novel BERT-based technique called NeighBERT made publicly available[1] for both NER and MEL. Our work performs MEL by first identifying spans (mentions) from clinical text and then classifying those spans as clinical concepts (entities) from a common medical knowledge base known as the Unified Medical Language System (UMLS) [6] as shown in Figure 1.

While in this article our experiments focus on the clinical text, NeighBERT could be used for any text where there are text descriptions available of entities as well as how they are related to others, typical to a knowledge base for example. The main reason that this will work is due to the vector-space

[1] Code is available at https://anonymous_for_blind_review

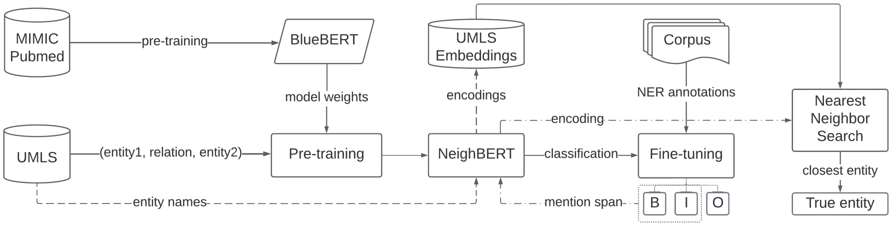

Fig. 2 An overview of how our proposed system works. Pre-training process is initialized with BlueBERT weights (already pre-trained on MIMIC dataset) and triplets from UMLS knowledge graph to induce relations. Pre-trained NeighBERT is then fine-tuned with corpus to detect mentions. Finally, entity linking process is done by encoding the mention spans using NeighBERT and performing a nearest neighbor search with an index of UMLS entity embeddings.

representations of text performed by BERT [1] which creates semantic representations by contextualizing the input text. In this work, we show that there is an advantage in including the relation between entities from a knowledge base into these pre-trained language models for MEL.

Knowledge graphs (KGs) like UMLS have been used for tasks like distance measurement and entity matching [7]. In this article, we compare NeighBERT to different state-of-the-art (SoTA) techniques for NER and MEL on the UMLS KG. To achieve this, we split MEL into two stages as shown in Figure 1: (1) *mention detection* which produces spans of text that could potentially be matched to a standardized entity followed by (2) *linking* mention spans to the correct entity based on semantically contextualized features. We compare NeighBERT to the SoTA approaches for two key corpora: (1) *i2b2-2010* [8] for NER and (2) *MedMentions* [9] for MEL.

sectionRelated Work

This section describes work related to NeighBERT to illustrate how our approach is universal and can be used for different NLP tasks and domains. Since we are specifically testing NeighBERT in the clinical domain, the investigative methods mentioned in this section mainly concentrate on clinical textual approaches. However, we present related work focused on broader domains that can be used for generic NLP tasks such as NER and NEL.

NeighBERT, formalized further in Section 2, focuses on encoding relationships found in a knowledge graph (i.e., UMLS). Different surveyed methods [7] have been used for encoding relationships and they range from the use of discrete features, such as the graph node degree or clustering coefficient inclusion, to shallow encoding approaches based on unsupervised neural network learning. In this section, we focus on three main topics when citing related work which are based on the development that NeighBERT takes to identify and classify named entities and concepts. The three developmental steps are: (1) the use of relationship extraction for clinical concepts, (2) the use of neural networks for NER, and (3) KGs in NER.

1.1 Clinical concept and relation extraction

Numerous works has experimented with different techniques for extracting clinical concepts from text. We mention those work that focuses specifically on extracting UMLS concepts through the use of NER and MEL. Work by [10] can be considered one of the main motivations for our work which is based on contextualized embeddings for enhancing clinical concept extraction. They compare several embedding methods like word2vec [11], GloVe [12], fastText [13], ELMo [14], and BERT [1] for use in clinical text to identify UMLS concepts. Their findings along with others [15–22] have shown that BERT-based models outperform other techniques for extracting clinical UMLS concepts. Based on their findings and other work [23–25] which show optimum performance is obtained for NER and MEL when the context is taken into account, we decide to experiment with a BERT-based contextualized embedding approach. In order to compare the performance of NeighBERT with the latest high-performing BERT-based techniques, we experiment with the following approaches: BioBERT [20], UmlsBERT [17], and BlueBERT [22]. We additionally provide insight in Section 4 on SapBERT [26] which uses UMLS concepts and their synonyms for MEL.

There are other approaches that either do not use or do not compare their approach to deep-learning approaches like we do. One approach called QuickUMLS [3] uses a string-matching algorithm which speeds up efficiency, but in our experiments it under-performs on MEL when compared to other BERT-based models. Another one, called MedCAT, uses a concept database built from biomedical text for detecting and linking concepts [2]. Since, to our knowledge, a comparison between QuickUMLS, MedCAT, and other deep learning approaches has not been performed for MEL, we add QuickUMLS and MedCAT as approaches to be compared in Section 4.

1.2 Classification approaches for NER

This task involves solely finding the named entities (or mentions), contemporary methods tend to use deep-learning techniques while more traditional methods do not. Several surveys cover the past thirty-year works that provide insight into different approaches. One survey by [27] covers work from year 1991 till 2006, and it shows that most NER works during that time are based on three types of statistical classifiers (supervised, semi-supervised, and unsupervised) and are evaluated using ACE [28] or GALE [29], shared tasks for the typical domains like news and gazetteer. Some of the methods are still used. A few examples of algorithms for NER that are still often used are: Hidden Markov Models [30] (Supervised), bootstrapping through syntactic and semantic relations [31] (supervised; this technique can be considered similar to our work for locating entities as a first step), and WordNet [32] clustering (unsupervised). Higher performing deep-learning approaches are covered in more recent surveys such as the one from [33]. They introduce nearly fifty approaches. The NER techniques that they have introduced are mostly measured using the CoNLL 2003 shared task [5] and presented architectures like the following: convolutional neural networks, recurrent neural networks, and deep transformers. Similar to our work, their best performing deep-learning architecture is based on the widely-used context encoder architecture. However, the best system reported from their survey is based on language-model embeddings and a hybrid long-term short-term (LSTM) neural network [34] mixed with BERT [1]. The introduction of an LSTM seems to have also performed well with nested NER detection [35]. For our work, despite the CoNLL performance found in the recent survey [33], we rely specifically on a BERT-based architecture called BlueBERT [22], which was found to outperform other techniques on clinical text.

1.3 Knowledge graphs use in NER

One of the fundamental innovations of our work is the inclusion of a KG for use in NER as a way to increase performance. Similar work has been used in different ways for embedding KG information, but NeighBERT uses the relations between clinical concepts found in UMLS like "is-a" and "part-of" (see Section 2, Figure 1). One similar work is from [36] which also uses relations from UMLS concepts, but, it uses Node2Vec [37] to store an entire KG. NeighBERT does not require graph storage, instead updates model weights to learn how entities are related. Another work [38] is similar to ours because it embeds information from UMLS and uses a spatial distance metric. Yet, its embedding approaches are not based on the transformer model (e.g. BERT [1]) like ours. Additionally, their work uses *cosine* similarity for measuring the distance between concepts while we use a Euclidean distance for measuring the difference. Other approaches [39–41] fall into similar categories where they either use different embedding approaches or a different form of classification such as more traditional neural networks like an LSTM [34].

2 Methodology

2.1 Problem formulation

The problem that we solve is based on maximizing the similarity between potential entities in the text and entries from the UMLS knowledge base, also called a KG since it has relations, nodes, and other attributes of a knowledge graph. In this section, we use the term "knowledge base" to loosely refer to a specific part of the UMLS concept KG in order to precisely transmit what NeighBERT is able to solve.

A knowledge base $\mathcal{G}$ can be represented as a set of pairs $\mathcal{G} = \{\mathcal{C}, \mathcal{R}\} = \{(e, r)\}$, where e is an entity from $\mathcal{C}$ (the universe of unique entities), r is a relation from $\mathcal{R}$ (the set of unique candidate relations). A document d from a collection of documents $\mathcal{D}$ is composed of several sequences where $\mathcal{X} = [x_0, x_1, \ldots, x_n]$ is represented as a span of tokens from d. A mention m is then represented as $m = x_{i:j}$ that begins with token i, ends with token j and can consist of a single token $(i = j)$ or a sequence of tokens $(j > i)$. An example follows.

In a document d where $\mathcal{X}$ = [*'The', 'patient', 'showed', 'up', 'to', 'the', 'ER', 'with', 'congestive', 'heart', 'failure', 'disease', 'and', 'has', 'complained', 'about', 'kidney', 'stones', 'in', 'the', 'past'*], let mention $m = x_{8:10}$ = [*'congestive', 'heart', 'failure'*]. Our task is to learn a function $f(m;\ \theta) : \mathcal{X} \rightarrow \mathbb{R}$ with parameters θ that maximizes the probability of finding the true concept e_{true} = *'heart failure'* from a universe of entities $\mathcal{C}$ as shown in Equation 1.

$$e_{true} = \text{argmax}_{e \in \mathcal{C}} f(P(m|e;\ \theta)) \tag{1}$$

2.2 Relation induction via pre-training

To pre-train NeighBERT, we extend the original work from [1] which uses cross-entropy minimization as the loss function for masked language modeling (MLM) where predictions are made on missing tokens using the `[MASK]` notation. Additionally, we abandon the next sentence prediction task since other work [42, 43] has shown it to be less effective. Since our task is to induce relations, we instead leverage context both before and after a masked token to predict relation. The relation is simply inserted as a text instead of a separate classification task which reduces the complexity of our approach as shown in Equation 2.

Our training objective provides the context of a relation to its mentions and vice-versa. To provide entity context to relations, we make sure that the following three components are available (unmasked) during training: (1) the head entity and relation when predicting the tail entity, (2) the tail entity and the relation when predicting the head entity, and (3) the head and tail when predicting relations.

Another importance of our work is the induction of relations from a UMLS knowledge graph. Since BERT [1] encodes sequences of tokens into contextualized (dense) representations, each token x_i in a sentence s will have its own representation $\vec{v}$. More importantly, the same token x_i will be represented differently if found in another sentence s^* such that $\vec{v}$ is not equal to $\vec{v}^*$. The pre-trained model used for allotting token representations is created using PubMed abstracts [44] and MIMIC-III clinical notes [45]. Their work is quite helpful since it establishes a baseline model for us to use from the clinical domain. However, it also presents a disadvantage in that BlueBERT [22] or BERT in general has not learned relationships between entities which is crucial for some tasks. To overcome this limitation, we introduce a custom pre-training technique to induce relational knowledge using relations from a KG ($\mathcal{G}$).

Our objective is to learn how entities are related to one another from $\mathcal{G}$. Input to NeighBERT is tokenized using the standard BERT [1] notation for a sentence s found in a document d (d is the document described in Section 2.1). Our notation uses a `[CLS]` token to mark the beginning of a classification example and a [SEP] token to identify the separation of entities, unlike typical BERT sentence classification where the [SEP] token is used for separating entire sentences. We can thus define X and Y as the input and target label respectively for NeighBERT. An example of how relational masking happens during pre-training process can be seen in Equation 2, where e_h is the entity head (*subject*) and e_t is the entity tail (*object*); r_{ht} represents the relation (*predicate*) between e_h and e_t; e^{name} and e^{desc} are the name and description of the entity e respectively denoted concatenated as E.

$$\begin{aligned} E &= e^{name} \oplus e^{desc} \\ X &= \texttt{[CLS]} \quad E_h \quad \texttt{[SEP]} \quad r_{ht} \quad \texttt{[SEP]} \quad E_t \quad \texttt{[EOS]} \\ Y &= \texttt{[CLS]} \; \texttt{[PAD]} \; \texttt{[SEP]} \; \texttt{[MASK]} \; \texttt{[SEP]} \; \texttt{[PAD]} \; \texttt{[EOS]} \end{aligned} \tag{2}$$

In order to prepare the data for pre-training, we label our target classes Y using a masking regime slightly different than the original BERT [1] training process. In our training, we only mask tokens of one of the semantic triplets (either of *subject, predicate, object*) at a time. Additionally, we mask the relations themselves (r) and predict them using information from the *subject* and *object* entity's context. This allows us to learn how the entities and their corresponding descriptions are related in the $\mathcal{G}$. In summary, we use the following three methods in NeighBERT to induce relations:

1. Use head E_h and relation r_{ht} to fill corrupted [MASK] tokens in tail E_t.
2. Use tail E_t and relation r_{ht} to fill corrupted [MASK] tokens in head E_h.
3. Use head E_h and tail E_t to fill corrupted [MASK] tokens of relation r_{ht}.

2.3 Fine-tuning for mention detection and linking

The experimental results and objective of this article are geared towards MEL. Our method relies on first detecting the span of words in a text (mention) followed by linking the predicted mention span with an entity in KG ($\mathcal{G}$). While the pre-training described in Section 2.2 helps induce relations, we further need to fine-tune our base model to detect mention(s) m. More specifically, we fine-tune for a NER task i.e. classify whether a token belongs to either of I, O, B (short for inside, outside, beginning) chunk tag set [46].

For linking, we encode the mention span using pre-trained model and use the encoding to find its nearest neighbor in an index that is comprised of encoding of all the entities from the UMLS. In order to do this, we extract the embedding of the [CLS] token as done in BERT [1] which is a representation of the encoding of the mention span. For our work, the target entity is predicted by determining the entity $e \in \mathcal{C}$ that has the closest Euclidean distance from the encoding of the predicted mention m as shown in Equation 3.

$$e_{true} = \text{argmin}_{e \in \mathcal{C}} \; \|\vec{v}_m - \vec{v}_e\|_2 \tag{3}$$

3 Experimental settings

In this section, we provide details about the corpora and its preparation along with the different parameters of training and evaluation.

3.1 Pre-training and fine-tuning

We pre-train using method described in Section 2.2 using the latest checkpoint released by BlueBERT[2] for 16704 steps and a batch size of 512. We use an Adam optimizer with a learning rate of `5e-5` and a weight decay rate of `1e-2`. We also used a cosine learning rate schedule for the first 500 steps of the training. In NeighBERT, a random span, used 15% of the time (replacing 80% of the masked tokens with `[MASK]`, 10% with random spans, and 10% with the original spans), is corrupted and the model is tasked to predict the masked token. All `[CLS]`, `[MASK]` and `[SEP]` are special tokens in BERT that are neither tokenized nor processed. As described in Section 2.3, we fine-tuned our models on mention detection task and used embeddings from last layer for entity linking. We kept the learning rates and optimizer the same as pre-training. Additionally, we used early stopping to avoid over-fitting for 30 epochs (on average the training stopped by 12^{th} epoch). Hardware comprised of 4 NVIDIA V100 GPU on a machine with 24 CPU cores and 448GB RAM.

3.2 Corpora

We used the full release of UMLS 2021AB[3] to build KG for entity retrieval. This KG including entities along with their names, descriptions and relations to other entities. Our final pre-training data 3,167,826 triples generated from a total of 3,883,156 entities related to each other via 853 different relation types forming over 25 million relations. Appendix A has a list of ontologies and details used for preparing UMLS.

For NER and MEL, we used two main data sources: (1) **i2b2-2010** [8] as a corpus for NER as it contains de-identified discharge summaries comprising of concepts, assertions and their relations and (2) **MedMentions** [9] corpus to measure performance on MEL as it contains 350,000 UMLS concepts mentions which are manually linked to 4,392 PubMed abstracts [44]. For comparing approaches using MedMentions, we used ST21pv subset as recommended by the original work. We limited our study to only MedMentions because earlier entity retrieval datasets [47, 48] are more than an order of magnitude smaller. Additionally, these datasets have lower coverage in the type of entities and the number of samples making them less than ideal for validating the efficacy of our approach.

Both corpora have been used for comparing systems similar to ours (see Section 1) and are considered SoTA for the clinical domain. The corpora used during development for pre-training and fine-tuning, when not defined already by the original task, were randomly split as 10% of the training set. All of

[2] https://huggingface.co/bionlp/bluebert_pubmed_mimic_uncased_L-24_H-1024_A-16

[3] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html#2021AB

Table 1 MedMentions [9] and i2b2-2010 [8]: document, mention, and unique entity count used for training and test.

Corpus	Documents		Mentions		Unique entities
	Train	Test	Train	Test	
MedMentions	3513	879	162908	40101	25419
i2b2-2010	170	256	16520	31161	7

the preprocessing of the corpora, including text used during pre-training and fine-tuning, was kept same as that of original work by [1, 22].

3.3 Evaluation

We divided our experiments into two main categories: (1) a NER comparison and (2) a MEL comparison. In the NER comparison, we included three SoTA approaches (BioBERT [20], UmlsBERT [17], and BlueBERT [22]) which were pre-trained on clinical text similar to NeighBERT. For the MEL comparison, we used MedCAT [2] which uses a novel unsupervised approach and QuickUMLS [3] which uses a dictionary-matching algorithm. For the MedCAT experiment, we built a concept database using MedCAT standards[4] from UMLS and MedMentions ST21pv training and dev sets.

The other BERT-based approaches introduced for NER did not link the entities to UMLS concepts on the MedMentions [9] corpus. Therefore, to compare against SoTA biomedical entity retrieval algorithm SapBERT [26], we extended BlueBERT [22] for NER and used its output as input for SapBERT [26] for MEL. Extending other BERT-based models like BioBERT [20], UmlsBERT [17], and BlueBERT [22] for evaluation on MedMentions [9] on the MEL task is something we leave for future work.

We measure performance using *precision*, *recall* and *micro* F1-scores on exact match linking at the mention level. In our results, an exact match is considered a true positive only when the start and end index of a mention span along with its linked entity type are predicted correctly.

4 Results and discussion

In this section, we present results from two main experiments covering NER using the i2b2-2010 corpus [8] and MEL using the MedMentions [9] corpus. Since NER has been previously reported with published results and codes for BioBERT [20], UmlsBERT [17], and BlueBERT [22] and they are considered SoTA, we compared NeighBERT to these three approaches. For the MEL task, we specifically reported on MedCAT [2] and QuickUMLS [3] to show SoTA results that were performed with distinct paradigms. Additionally, we reported on another BERT-based combination (BlueBERT [22] and SapBERT [26])

[4] https://github.com/CogStack/MedCAT

which was not found in the literature, a novel combination created during the development phase.

Table 2 A comparison to NeighBERT of precision (P), recall (R), and micro F1-score for NER on i2b2-2010 with other BERT-based techniques.For MEL, NeighBERT is compared with rule-based (MedCAT and QuickUMLS) as well as adapted (BlueBERT and SapBERT) approaches.

NER Model	P	R	F1
BioBERT	.868	.856	.867
UmlsBERT	.873	**.896**	.885
BlueBERT	.836	.835	.836
NeighBERT	**.902**	.884	**.893**

MEL Model	P	R	F1
MedCAT	.273	.281	.277
QuickUMLS	.254	.202	.225
Blue+SapBERT	**.458**	.328	**.381**
NeighBERT	.323	**.453**	.377

Table 2 contains the comparisons for both NER and MEL on the different approaches. It is important to note that, although the BlueBERT [22] and SapBERT [26] implementation scores higher than NeighBERT for the MEL task, it comes with the trade-off of additional complexity by having to work with adapting input-outputs of two large models. On the other hand, when NeighBERT is pre-trained and fine-tuned, it could be used end-to-end for both NER and MEL thus removing the need for combining and adapting models like BlueBERT and SapBERT in turn eradicating complexity.

There are other important findings to note: NeighBERT considerably improves precision for the i2b2-2010 corpus [8], this is due to the assumptions that are made when predicting ambiguous entities. It also scores nearly identical to UmlsBERT and 3 to 6 points better than both BioBERT [20] and BlueBERT [22] (SoTA for NER), respectively. For the MEL task, NeighBERT improves precision by 5 points when compared to MedCAT [2], considered SoTA for MedMentions. Additionally, NeighBERT improves recall over all approaches for MEL, including the adapted version of BlueBERT and SapBERT.

During our analysis, we found that the method of measuring true positives for MedMentions [9] was mixed. MedCAT's implementation [2], to our knowledge, only takes into account the starting index of span and entity type, while another work by [4] only counted sets of entities ignoring duplicates in the process. Those types of evaluation could lead to conflicting evaluations for clinical text phrases e.g. *Correlation with severity* is counted correct even if the correct mention span is *Correlation* because end index is ignored. For evaluation of NeighBERT, we included the start and end index along with entity ID when comparing our results to the MedMentions test set. This increases the rigor and hence results in realistic scores.

Granted that if the knowledge graph is directed as can be seen present in Figure 1, NeighBERT offers another advantage where it not only learns how things are correlated in real-world but also the directionality of correlation i.e. *causation*. This paves the way for machine learning models to also have a sense of causality, which is missing in even the largest of language models out there.

Two important drawbacks were discovered during experimentation. First, since NeighBERT predicts mention spans given a sentence, it is not able to correctly label overlapping entities. Second, NeighBERT uses UMLS definitions for measuring the distance between mentions and concepts; since only 60% of the MedMentions concepts have corresponding UMLS definitions, the full performance potential was not achieved for the MEL task.

5 Conclusion and future work

In this work, we introduce a novel deep-learning method called NeighBERT that can be used on any type of digital text to perform end-to-end MEL. We experimented with several SoTA approaches and found that our approach is robust and able to improve performance on both tasks. NeighBERT is distributed publicly for the inclusion in SoTA experiments.

During the development of NeighBERT, we decided to leave the following investigative lines for future work. First, we plan to try NeighBERT for non-clinical text tasks such as those presented at CoNLL 2003 [5]. Second, along the same lines we also plan to evaluate or develop datasets that validate the ability of models to leverage relational and causal information. Third, previous work outside of the clinical domain has shown to work well by mixing deep learning methods such as LSTMs [34] with BERT [1], we plan on trying these hybrid techniques for improving mention detection. As noted in Section 4, since there is limited coverage for UMLS definitions, we plan on developing a technique based on taxonomies or augmentation [49] for retrieving even those concepts that do not have descriptions by default. Lastly, we plan on addressing the lack of ability to predict overlapping entities by introducing a re-sampling technique that uses a threshold for mention detection inspired by work of [50–52].

Supplementary information. The code for reproducing the models along with datasets are provided as supplementary.

Declarations

5.1 Ethical Approval

There are no human subjects in this article and hence informed consent does not apply. Furthermore, our work does not involve any kind of sensitive data.

5.2 Competing interests

The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

5.3 Authors' contributions

Ayush Singh did the conceptualization and implementation of the technique. Ayush, Saranya and John conceived and planned the experiments. Saranya performed additional studies for comparison. Ayush and John wrote the main manuscript text. Ayush and Saranya prepared all the figures as well as curated the datasets required for the study. All authors provided critical feedback and helped shape the research, analysis and manuscript.

5.4 Funding

This funding was in-part supported by Evernorth, a health-care related enterprise with the intention of creating freely available software and contributing to the medical community. Funding for this project was not issued on a grant basis nor on a for-profit basis. Contributors intent to publish all results and other materials on a public forum with Apache 2.0 and/or Creative Commons license.

5.5 Availability of data and materials

We make our code and data[5] used for pre-training publicly available. Note that all datasets i.e. i2b2[6] and UMLS[7] with the exception of MedMentions require the end user to licensed. More information can be found in readme of our shared codebase.

Appendix A UMLS setup

We used the full release of UMLS 2021AB[8] to build KG for entity retrieval. Specifically, we used `MRCONSO.RRF`, `MRDEF.RRF`, `MRSTY.RRF` and `MRREL.RRF` for extracting entities, synonyms, descriptions and relations correspondingly.

References

[1] Devlin, J., Chang, M.-W., Lee, K., Toutanova, K.: BERT: Pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers), pp. 4171–4186. Association for Computational Linguistics, Minneapolis, Minnesota (2019). https://doi.org/10.18653/v1/N19-1423. https://aclanthology.org/N19-1423

[2] Kraljevic, Z., Searle, T., Shek, A., Roguski, L., Noor, K., Bean, D., Mascio, A., Zhu, L., Folarin, A.A., Roberts, A., Bendayan, R., Richardson, M.P., Stewart, R., Shah, A.D., Wong, W.K., Ibrahim, Z., Teo, J.T., Dobson,

[5] Code and data available at https://anonymous for blind review

[6] https://portal.dbmi.hms.harvard.edu/projects/n2c2-nlp/

[7] https://www.nlm.nih.gov/research/umls

[8] https://www.nlm.nih.gov/research/umls/licensedcontent/umlsarchives04.html#2021AB

Fig. A1 Distribution of length of name of entities in UMLS. Most entity names are less than 5 words. We sample from this distribution when deciding how many tokens to mask in a given sequence.

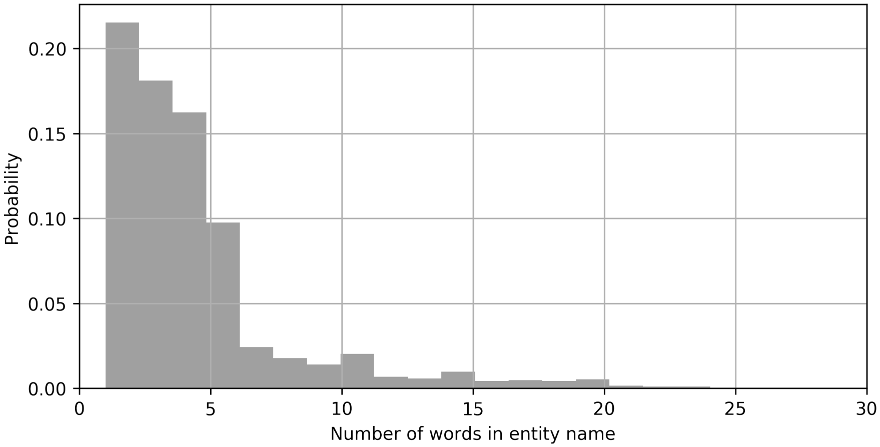

Table A1 The restricted set of source ontologies used for preparation of our UMLS Knowledge Base. We use the same ontologies used in original work by [9] with the difference of year being 2021 instead of 2017 and NDFRT being officially superceded by MED-RT.

Ontology Abbrev.	Name
CPT	Current Procedural Terminology
FMA	Foundational Model of Anatomy
GO	Gene Ontology
HGNC	HUGO Gene Nomenclature Committee
HPO	Human Phenotype Ontology
ICD10	International Classification of Diseases, Tenth Revision
ICD10CM	ICD10 Clinical Modification
ICD9CM	ICD9 Clinical Modification
MDR	Medical Dictionary for Regulatory Activities
MSH	Medical Subject Headings
MTH	UMLS Metathesaurus Names
NCBI	National Center for Biotechnology Information Taxonomy
NCI	National Cancer Institute Thesaurus
NDDF	First DataBank MedKnowledge
MED-RT	National Drug File–Reference Terminology
OMIM	Online Mendelian Inheritance in Man
RXNORM	NLM's Nomenclature for Clinical Drugs for Humans
SNOMEDCT_US	US edn. of the Systematized Nomenclature of Medicine-Clinical Terms

R.J.B.: Multi-domain clinical natural language processing with MedCAT: The medical concept annotation toolkit. Artif. Intell. Med. **117**, 102083 (2021). https://doi.org/10.1016/j.artmed.2021.102083

[3] Soldaini, L., Goharian, N.: Quickumls: a fast, unsupervised approach for medical concept extraction. In: MedIR Workshop, Sigir, pp. 1–4 (2016)

[4] Mohan, S., Angell, R., Monath, N., McCallum, A.: Low resource recognition and linking of biomedical concepts from a large ontology. In: Proceedings of the 12th ACM Conference on Bioinformatics, Computational Biology, and Health Informatics, pp. 1–10 (2021)

[5] Sang, E.T.K., De Meulder, F.: Introduction to the conll-2003 shared task: Language-independent named entity recognition. In: Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, pp. 142–147 (2003)

[6] Bodenreider, O.: The unified medical language system (umls): integrating biomedical terminology. Nucleic acids research **32**(suppl_1), 267–270 (2004)

[7] Wang, Q., Mao, Z., Wang, B., Guo, L.: Knowledge graph embedding: A survey of approaches and applications. IEEE Transactions on Knowledge and Data Engineering **29**(12), 2724–2743 (2017)

[8] Uzuner, O., South, B.R., Shen, S., DuVall, S.L.: 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. J. Am. Medical Informatics Assoc. **18**(5), 552–556 (2011). https://doi.org/10.1136/amiajnl-2011-000203

[9] Mohan, S., Li, D.: Medmentions: A large biomedical corpus annotated with umls concepts. In: Automated Knowledge Base Construction (AKBC) (2018)

[10] Si, Y., Wang, J., Xu, H., Roberts, K.: Enhancing clinical concept extraction with contextual embeddings. Journal of the American Medical Informatics Association **26**(11), 1297–1304 (2019)

[11] Mikolov, T., Sutskever, I., Chen, K., Corrado, G.S., Dean, J.: Distributed representations of words and phrases and their compositionality. Advances in neural information processing systems **26** (2013)

[12] Pennington, J., Socher, R., Manning, C.D.: Glove: Global vectors for word representation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP), pp. 1532–1543 (2014)

[13] Bojanowski, P., Grave, E., Joulin, A., Mikolov, T.: Enriching word vectors with subword information. Transactions of the association for computational linguistics **5**, 135–146 (2017)

[14] Peters, M.E., Neumann, M., Iyyer, M., Gardner, M., Clark, C., Lee, K.,

Zettlemoyer, L.: 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), pp. 2227–2237. Association for Computational Linguistics, New Orleans, Louisiana (2018). https://doi.org/10.18653/v1/N18-1202. https://aclanthology.org/N18-1202

[15] Yang, X., Bian, J., Hogan, W.R., Wu, Y.: Clinical concept extraction using transformers. Journal of the American Medical Informatics Association **27**(12), 1935–1942 (2020)

[16] Fu, S., Chen, D., He, H., Liu, S., Moon, S., Peterson, K.J., Shen, F., Wang, L., Wang, Y., Wen, A., *et al.*: Clinical concept extraction: a methodology review. Journal of biomedical informatics **109**, 103526 (2020)

[17] Michalopoulos, G., Wang, Y., Kaka, H., Chen, H., Wong, A.: UmlsBERT: Clinical domain knowledge augmentation of contextual embeddings using the Unified Medical Language System Metathesaurus. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 1744–1753. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.naacl-main.139. https://aclanthology.org/2021.naacl-main.139

[18] Ji, Z., Wei, Q., Xu, H.: Bert-based ranking for biomedical entity normalization. AMIA Summits on Translational Science Proceedings **2020**, 269 (2020)

[19] Joshi, M., Chen, D., Liu, Y., Weld, D.S., Zettlemoyer, L., Levy, O.: Spanbert: Improving pre-training by representing and predicting spans. Transactions of the Association for Computational Linguistics **8**, 64–77 (2020)

[20] Lee, J., Yoon, W., Kim, S., Kim, D., Kim, S., So, C.H., Kang, J.: Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics **36**(4), 1234–1240 (2020)

[21] Nejadgholi, I., Fraser, K.C., De Bruijn, B., Li, M., LaPlante, A., El Abidine, K.Z.: Recognizing umls semantic types with deep learning. In: Proceedings of the Tenth International Workshop on Health Text Mining and Information Analysis (LOUHI 2019), pp. 157–167 (2019)

[22] Peng, Y., Yan, S., Lu, Z.: Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets. BioNLP 2019, 58 (2019)

[23] Harnoune, A., Rhanoui, M., Mikram, M., Yousfi, S., Elkaimbillah, Z.,

El Asri, B.: Bert based clinical knowledge extraction for biomedical knowledge graph construction and analysis. Computer Methods and Programs in Biomedicine Update **1**, 100042 (2021)

[24] Zhai, Z., Nguyen, D.Q., Akhondi, S.A., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M., Verspoor, K.: Improving chemical named entity recognition in patents with contextualized word embeddings. arXiv preprint arXiv:1907.02679 (2019)

[25] Zhang, T., Cai, Z., Wang, C., Qiu, M., Yang, B., He, X.: SMedBERT: A knowledge-enhanced pre-trained language model with structured semantics for medical text mining. In: Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers), pp. 5882–5893. Association for Computational Linguistics, Online (2021). https://doi.org/10.18653/v1/2021.acl-long.457. https://aclanthology.org/2021.acl-long.457

[26] Liu, F., Shareghi, E., Meng, Z., Basaldella, M., Collier, N.: Self-alignment pretraining for biomedical entity representations. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, pp. 4228–4238 (2021)

[27] Nadeau, D., Sekine, S.: A survey of named entity recognition and classification. Lingvisticae Investigationes **30**(1), 3–26 (2007)

[28] Doddington, G.R., Mitchell, A., Przybocki, M.A., Ramshaw, L.A., Strassel, S.M., Weischedel, R.M.: The automatic content extraction (ACE) program - tasks, data, and evaluation. In: Proceedings of the Fourth International Conference on Language Resources and Evaluation, LREC 2004, May 26-28, 2004, Lisbon, Portugal. European Language Resources Association, ??? (2004). http://www.lrec-conf.org/proceedings/lrec2004/summaries/5.htm

[29] Cohen, J.: The gale project: A description and an update. In: 2007 IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU), pp. 237–237 (2007). IEEE

[30] Todorovic, B.T., Rancic, S.R., Markovic, I.M., Mulalic, E.H., Ilic, V.M.: Named entity recognition and classification using context hidden markov model. In: 2008 9th Symposium on Neural Network Applications in Electrical Engineering, pp. 43–46 (2008). IEEE

[31] Cucchiarelli, A., Velardi, P.: Unsupervised named entity recognition using syntactic and semantic contextual evidence. Computational Linguistics **27**(1), 123–131 (2001)

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

[33] Li, J., Sun, A., Han, J., Li, C.: A survey on deep learning for named entity recognition. IEEE Transactions on Knowledge and Data Engineering **34**(1), 50–70 (2020)

[34] Schmidhuber, J., Hochreiter, S., *et al.*: Long short-term memory. Neural Comput **9**(8), 1735–1780 (1997)

[35] Straková, J., Straka, M., Hajic, J.: Neural architectures for nested ner through linearization. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, pp. 5326–5331 (2019)

[36] Kotitsas, S., Pappas, D., Androutsopoulos, I., McDonald, R., Apidianaki, M.: Embedding biomedical ontologies by jointly encoding network structure and textual node descriptors. In: Proceedings of the 18th BioNLP Workshop and Shared Task, pp. 298–308 (2019)

[37] Grover, A., Leskovec, J.: node2vec: Scalable feature learning for networks. In: Proceedings of the 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, pp. 855–864 (2016)

[38] Park, J., Kim, K., Hwang, W., Lee, D.: Concept embedding to measure semantic relatedness for biomedical information ontologies. Journal of biomedical informatics **94**, 103182 (2019)

[39] Lamurias, A., Sousa, D., Clarke, L.A., Couto, F.M.: Bo-lstm: classifying relations via long short-term memory networks along biomedical ontologies. BMC bioinformatics **20**(1), 1–12 (2019)

[40] Beam, A.L., Kompa, B., Schmaltz, A., Fried, I., Weber, G., Palmer, N., Shi, X., Cai, T., Kohane, I.S.: Clinical concept embeddings learned from massive sources of multimodal medical data. In: PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020, pp. 295–306 (2019). World Scientific

[41] Mao, Y., Fung, K.W.: Use of word and graph embedding to measure semantic relatedness between unified medical language system concepts. Journal of the American Medical Informatics Association **27**(10), 1538–1546 (2020)

[42] Liu, Y., Ott, M., Goyal, N., Du, J., Joshi, M., Chen, D., Levy, O., Lewis, M., Zettlemoyer, L., Stoyanov, V.: Roberta: A robustly optimized bert pretraining approach. arXiv preprint arXiv:1907.11692 (2019)

[43] Lan, Z., Chen, M., Goodman, S., Gimpel, K., Sharma, P., Soricut, R.:

ALBERT: A lite BERT for self-supervised learning of language representations. In: 8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020. OpenReview.net, ??? (2020). https://openreview.net/forum?id=H1eA7AEtvS

[44] Fiorini, N., Leaman, R., Lipman, D.J., Lu, Z.: How user intelligence is improving pubmed. Nature biotechnology **36**(10), 937–945 (2018)

[45] Johnson, A.E., Pollard, T.J., Shen, L., Lehman, L.-w.H., Feng, M., Ghassemi, M., Moody, B., Szolovits, P., Anthony Celi, L., Mark, R.G.: Mimic-iii, a freely accessible critical care database. Scientific data **3**(1), 1–9 (2016)

[46] Ramshaw, L., Marcus, M.: Text chunking using transformation-based learning. In: Third Workshop on Very Large Corpora (1995). https://aclanthology.org/W95-0107

[47] Dogan, R.I., Leaman, R., Lu, Z.: NCBI disease corpus: A resource for disease name recognition and concept normalization. J. Biomed. Informatics **47**, 1–10 (2014). https://doi.org/10.1016/j.jbi.2013.12.006

[48] Li, J., Sun, Y., Johnson, R.J., Sciaky, D., Wei, C., Leaman, R., Davis, A.P., Mattingly, C.J., Wiegers, T.C., Lu, Z.: Biocreative V CDR task corpus: a resource for chemical disease relation extraction. Database J. Biol. Databases Curation **2016** (2016). https://doi.org/10.1093/database/baw068

[49] Vashishth, S., Newman-Griffis, D., Joshi, R., Dutt, R., Rosé, C.P.: Improving broad-coverage medical entity linking with semantic type prediction and large-scale datasets. Journal of Biomedical Informatics **121**, 103880 (2021). https://doi.org/10.1016/j.jbi.2021.103880

[50] Fei, H., Ji, D., Li, B., Liu, Y., Ren, Y., Li, F.: Rethinking boundaries: End-to-end recognition of discontinuous mentions with pointer networks. In: Proceedings of the AAAI Conference on Artificial Intelligence, pp. 12785–12793 (2021)

[51] Muis, A.O., Lu, W.: Labeling gaps between words: Recognizing overlapping mentions with mention separators. In: Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, pp. 2608–2618 (2017)

[52] Li, F., Lin, Z., Zhang, M., Ji, D.: A span-based model for joint overlapped and discontinuous named entity recognition. In: Proceedings of the ACL (2021)