EntityBERT: Entity-centric Masking Strategy for Model Pretraining for the Clinical Domain

Chen Lin[1], Timothy Miller[1], Dmitriy Dligach[2], Steven Bethard[3] and Guergana Savova[1]

[1]Boston Children's Hospital and Harvard Medical School
[2]Loyola University Chicago
[3]University of Arizona
[1]{first.last}@childrens.harvard.edu
[2]ddligach@luc.edu
[3]bethard@email.arizona.edu

Abstract

Transformer-based neural language models have led to breakthroughs for a variety of natural language processing (NLP) tasks. However, most models are pretrained on general domain data. We propose a methodology to produce a model focused on the clinical domain: continued pretraining of a model with a broad representation of biomedical terminology (PubMedBERT) on a clinical corpus along with a novel entity-centric masking strategy to infuse domain knowledge in the learning process. We show that such a model achieves superior results on clinical extraction tasks by comparing our entity-centric masking strategy with classic random masking on three clinical NLP tasks: cross-domain negation detection (Wu et al., 2014), document time relation (DocTimeRel) classification (Lin et al., 2020b), and temporal relation extraction (Wright-Bettner et al., 2020). We also evaluate our models on the PubMedQA(Jin et al., 2019) dataset to measure the models' performance on a non-entity-centric task in the biomedical domain. The language addressed in this work is English.

1 Introduction

Transformer-based neural language models, such as BERT (Devlin et al., 2018), have achieved state-of-the-art performance for a variety of natural language processing (NLP) tasks. Since most are pre-trained on large general domain corpora, many efforts have been made to continue pre-taining general-domain language models on clinical/biomedical corpora to derive domain-specific language models (Lee et al., 2020; Alsentzer et al., 2019; Beltagy et al., 2019).

Yet, as Gu et al. (2020a) pointed out, in specialized domains such as biomedicine, continued pre-training from generic language models is inferior to domain-specific pretraining from scratch. Continued pre-training from a generic model would break down many of the domain specific terms into sub-words through the Byte-Pair Encoding (BPE) (Gage, 1994) or variants like WordPiece tokenization (Wu et al., 2016) because these specific terms are not in the vocabulary of the generic pretrained model. A clinical domain-specific pretraining from scratch would derive an in-domain vocabulary as many of the biomedical terms, such as diseases, signs/symptoms, medications, anatomical sites, procedures, would be represented in their original form. Such an improved word-level representation is expected to bring substantial performance gains in clinical domain tasks because the model would learn the characteristics of the term along with its surrounding context as one unit.

In our preliminary work on a clinical relation extraction task, we observed a performance gain with the PubMedBERT model (Gu et al., 2020a) which outperformed BioBERT (Lee et al., 2020), ClinicalBERT (Alsentzer et al., 2019), and even some larger general domain models like RoBERTa (Liu et al., 2019) and BART-large (Lewis et al., 2019). The performance gain was primarily attributed to PubMedBERT's in-domain vocabulary as we observed that PubMedBERT kept 30% more in-domain words in its vocabulary than BERT. When we swapped PubMedBERT tokenization with BERT or RoBERTa tokenization, the performance of PubMedBERT degraded.

Thus, PubMedBERT appears to provide a vocabulary that is helpful to the clinical domain. However, the language of biomedical literature is different from the language of the clinical documents found in electronic medical records (EMRs). In general, a clinical document is written by physicians who have very limited time to express the numerous details of a patient-physician encounter. Many nonstandard expressions, abbreviations, assumptions and domain knowledge are used in clinical notes which makes the text hard to understand outside of the clinical community and presents

Proceedings of the BioNLP 2021 workshop, pages 191–201
June 11, 2021. ©2021 Association for Computational Linguistics

challenges for automated systems. Pretraining a language model specific to the clinical domain requires large amounts of unlabeled clinical text on par with what the generic models are trained on. Unfortunately, such data are not available to the community. The only available such corpus is MIMIC III used to train ClinicalBERT (Alsentzer et al., 2019) and BlueBERT (Peng et al., 2019), but it is magnitudes smaller and represents one specialty in medicine – intensive care.

Pretraining is agnostic to downstream tasks: it learns representations for all words using a self-supervised data-rich task. Yet, not all words are important for downstream fine-tuning tasks. Numerous pretrained words are not even used in the fine-tuning step, while important words crucial for the downstream task are not well represented due to insufficient amounts of labeled data. Many clinical NLP tasks are centered around entities: clinical named entity recognition aims to detect clinical entities (Wu et al., 2017; Pradhan et al., 2014; Elhadad et al., 2015), clinical negation extraction decides if a certain clinical entity is negated (Chapman et al., 2001; Harkema et al., 2009; Mehrabi et al., 2015), clinical relation discovery extracts relations among clinical entities (Lv et al., 2016; Leeuwenberg and Moens, 2017), etc. Though various masking strategies have been employed during pretraining – masking contiguous spans of text (SpanBERT, Joshi et al., 2020; BART, Lewis et al., 2019), varying masking ratios (Raffel et al., 2019), building additional neural models to predict which words to mask (Gu et al., 2020b), incorporating knowledge graphs (Zhang et al., 2019), masking entities for a named entity recognition task (Ziyadi et al., 2020) – none of the masking techniques so far have investigated and focused on clinical entities.

Besides transformer-based models, there are other efforts (Beam et al., 2019; Chen et al., 2020) to characterize the biomedical/clinical entities at the word embedding level. There are also other statistical methods applied to the downstream tasks. We do not include these efforts in our discussion because the focus of our paper is the investigation of a novel entity-based masking strategy in a transformer-based setting.

In this paper, we propose a methodology to produce a model focused on clinical entities: continued pretraining of a model with a broad representation of biomedical terminology (the PubMedBERT model) on a clinical corpus, along with a novel entity-centric masking strategy to infuse domain knowledge in the learning process[1]. We show that such a model achieves superior results on clinical extraction tasks by comparing our entity-centric masking strategy with classic random masking on three clinical NLP tasks: cross-domain negation detetction (Wu et al., 2014), document time relation (DocTimeRel) classification (Lin et al., 2020b), and temporal relation extraction (Wright-Bettner et al., 2020).

The contributions of this paper are: (1) a continued pretraining methodology for clinical domain specific neural language models, (2) a novel entity-centric masking strategy to infuse domain specific knowledge, (3) evaluation of the proposed strategies on three clinical tasks: cross-domain negation detection, DocTimeRel classification, and temporal relation extraction, and (4) evaluation of our models on the PubMedQA (Jin et al., 2019) dataset to measure the models' performance on a non-entity-centric task in the biomedical domain.

2 Methods

In this section, we first describe our clinical text datasets and related NLP tasks, the details of our entity-centric masking strategy, and finally the settings we used for both pretraining and fine-tuning.

2.1 Transformer models

Transformer models learn a sequential contextual representation of the input sequence through a multi-layer, multi-head self-attention mechanism, which models long-range dependencies in texts through highly parallel computation. They are usually pretrained through a self-supervised masked language model (MLM) task i.e., predicting the randomly masked subset of the input tokens. Some transformer models also use next sentence prediction (NSP) as a self-supervision task i.e., predicting if two given sentences are adjacent in the original text. A language model can be continuously pretrained on new corpora to further expand its representative power especially for a target domain. For a task-specific application, a pretrained language model's parameters are usually refined through a fine-tuning process on the task-specific training data, and a special [CLS] token is usually used as

[1] Our pretrained models are submitted to PhysioNet(Goldberger et al., 2000). Once approved, they will be publicly available through PhysioNet Credentialed Health Data License 1.5.0.

Dataset	sentence#	word#	entities#/sentence
MIMIC-SMALL	4.6M	125.1M	1+
MIMIC-BIG	15.6M	728.6M	2+

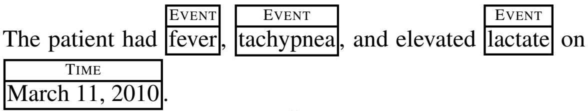
Table 1: Two versions of curated MIMIC data.

the representation of the input instance for text-classification tasks.

2.2 Unlabeled Pre-training Data

MIMIC-III We use the freely-available MIMIC-III (Medical Information Mart for Intensive Care) Clinical dataset (Johnson et al., 2016) (version 1.4) for continued pretraining of the PubMedBERT model. This dataset comprises approximately 2M deidentified notes for over 40,000 patients who stayed in critical care units of the Beth Israel Deaconess Medical Center between 2001 and 2012.

We process the MIMIC-III corpus with the sentence detection, tokenization, and temporal modules of Apache cTAKES (Savova et al., 2010)[2] to identify all entities (events and time expressions) in the corpus. Events are recognized by cTAKES event annotator. Event types include diseases/disorders, signs/symptoms, medications, anatomical sites, and procedures. Time expressions are recognized by cTAKES timex annotator. Time classes includes: date, time, duration, quantifier, prepostesp, and set (Styler IV et al., 2014). Special XML tags (Dligach et al., 2017) are inserted into the text sequence to mark the position of identified entities. Time expressions are replaced by their time class (Lin et al., 2017, 2018) for better generalizability. All special XML-tags and time class tokens are added into the PubMedBERT vocabulary so that they can be recognized. The top line of Figure 1 shows a sample sentence from the MIMIC-III corpus. The entities of this sentence are identified by Apache cTAKES. The bottom line of Figure 1 shows the entities marked by XML tags and the temporal expression replaced by its class. We process the MIMIC corpus sentence by sentence, and discard sentences that have fewer than two entities. The resulting set (MIMIC-BIG) has 15.6 million sentences, 728.6 million words (the bottom row of Table 1). In another setting, from the pool of sentences with at least one entity, we sample a smaller set (MIMIC-SMALL), resulting in 4.6 million sentences and 125.1 million words (the top row of Table 1).


EVENT
EVENT
EVENT
TIME
The patient had fever, tachypnea, and elevated lactate on March 11, 2010.

⇓

The patient had **<e>** fever **</e>**, **<e>** tachypnea **</e>**, and elevated **<e>** lactate **</e>** on **<t>** date **</t>**.

Figure 1: MIMIC-III text with XML-tagged entities: <e> and </e> mark events; <t> and </t> mark time expressions.

#1: she is feeling reasonably well . she has not **<e>** noted **</e>** any new areas of pain and has had no fevers
#2: a **<e>** surgery **</e>** was scheduled on **<t>** date **</t>** .
#3: a **<e1>** surgery **</e1>** was **<e2>** scheduled **</e2>** on march 11th .
#4: she denies any **<e>** fevers **</e>** or chills .
#5: Inpatient versus outpatient management of neutropenic fever in gynecologic oncology patients: is risk stratification useful? **ANSWER:** Based on this pilot data, MASCC score appears promising in determining suitability for outpatient management of NF in gynecologic oncology patients. Prospective study is ongoing to confirm safety and determine impact on cost.

Figure 2: Sample instances for DocTimeRel(1), TLINK:event-time(2), TLINK:event-event(3), Negation (4), and PubMedQA (5).

2.3 Labeled Fine-tuning Data

The following sections describe the labeled datasets that are used as fine-tuning tasks. Figure 2 shows examples of how we format inputs for these tasks (more details below).

THYME The THYME corpus (Styler IV et al., 2014) is widely used (Bethard et al., 2015, 2016, 2017) for clinical temporal relation discovery. There are two types of temporal relations defined in it: (1) The document time relations (DocTimeRel), which link a clinical event (EVENT) to the document creation time (DCT) with possible values of *BEFORE, AFTER, OVERLAP, and BEFORE_OVERLAP*, and (2) pairwise temporal relations (TLINK) between two events (EVENT) or an event and a time expression (TIMEX3) using an extension of TimeML (Pustejovsky et al., 2003; Pustejovsky and Stubbs, 2011). Recently, the TLINK annotations of (2) were refined with values of *BEFORE, BEGINS-ON, CONTAINS, CONSUB, ENDS-ON, NOTED-ON, OVERLAP*, with the revised corpus known as the THYME+ corpus (Wright-Bettner et al., 2020).

For the DocTimeRel task, we mark all events in THYME+ corpus with XML tags (“<e>”, “</e>”) and extract 10 tokens from each side of the event as the contextual information. The DocTimeRel

[2] http://ctakes.apache.org

labels are predicted using the special [CLS] embedding and a softmax function.

For the TLINK task, we use the THYME+ annotation and the same window-based processing (Lin et al., 2019; Wright-Bettner et al., 2020) for generating relational candidates. The two entities involved in a relation candidate are marked by XML tags following the style of Dligach et al. (2017). Time expressions are represented by their time classes. The TLINK labels are predicted using the special [CLS] embedding and a softmax function.

Cross-domain Negation We use the same corpora as Miller et al. (2017); Lin et al. (2020a): (1) 2010 i2b2/VA NLP Challenge Corpus (i2b2: Uzuner et al., 2011), (2) the Multi-source Integrated Platform for Answering Clinical Questions Corpus (MiPACQ: Albright et al., 2013), (3) the Strategic Health IT Advanced Research Projects (SHARP) Seed (Seed), and (4) SHARP Stratified (Strat). We use them for fine-tuning the pretrained models for the cross-domain negation task. The same XML tags as described above mark the entities for which the negation status is to be determined. The +1(negated) and -1(not negated) labels are predicted using the special [CLS] embedding and a softmax function.

PubMedQA PubMedQA (Jin et al., 2019) is a biomedical question answering (QA) dataset collected from PubMed abstracts. The task is to answer research questions with yes/no/maybe using the corresponding abstracts or the conclusion sections of the abstracts (i.e., the long answers). For simplicity, we only fine-tune pretrained models on the PubMedQA labeled (PQA-L) data of 1K expert annotations, with the original train/dev/test split with 450, 50, 500 questions, respectively. The unlabeled (PQA-U) and artificially generated QA instances (PQA-A) are not used. Pretrained models are fine-tuned on the PQA-L data in the reasoning-free setting (without reasoning the full abstracts as contexts) by concatenating the questions and related long answers. The question and the answer is separated by "ANSWER:" (as shown in the bottom case of fig. 2) instead of the special [SEP] token in order not to involve the Next Sentence Prediction (NSP). The yes/no/maybe labels are predicted using the special [CLS] embedding and a softmax function.

2.4 Entity-centric Masking

Conventional BERT-style Masked Language Model (MLM) randomly chooses 15% of the input tokens for corruption, among which 80% are replaced by a special token "[MASK]", 10% are left unchanged, and 10% are randomly replaced by a token from the vocabulary. The language model is trained to reconstruct the masked tokens.

We propose an entity-centric masking strategy (as shown in Figure 3). All entities in the input sequence are marked with XML tags, which are added into the vocabulary and mapped to unique IDs. Then 40% of entities and 12% of random words are chosen respectively within each sequence block for corruption, following the same 80%-10%-10% ratio for [MASK], unchanged, and random replacement. We refer to this masking strategy as entity-centric masking.

We did not use the Next Sentence Prediction (NSP) task in our pretraining experiments based on Liu et al. (2019).

The PubMedBERT base uncased version was pretrained from scratch using abstracts from PubMed and full-text articles from PubMedCentral. We applied continued pretraining on it with MIMIC-BIG and MIMIC-SMALL with entity-centric masking and random masking. We denote the model pretrained with entity-centric masking **EntityBERT**, and model pretrained with random masking **RandMask**. For both masking strategies, we use different random seeds.

The pretrained models are then fine-tuned for the three clinical tasks (TLINK temporal relation extraction, DocTimeRel classification, and negation detection) and one biomedical task (PubMedQA). Since the TLINK task has the most relation types and is the most complicated task among the three, we use it as the primary testing task. The best models derived on the TLINK task are then tested on the other tasks.

2.5 Settings

We used an NVIDIA Titan RTX GPU cluster of 7 nodes for pre-training and fine-tuning experiments through HuggingFace's Transformer API (Wolf et al., 2019) version 2.10.0.

For pretraining, we set the max steps to 200k to allow full model convergence, and set the block size to 100. For fine-tuning, the batch size is selected from (16, 32), the learning rate is selected from (1e-5, 2e-5, 3e-5, 4e-5, 5e-5).

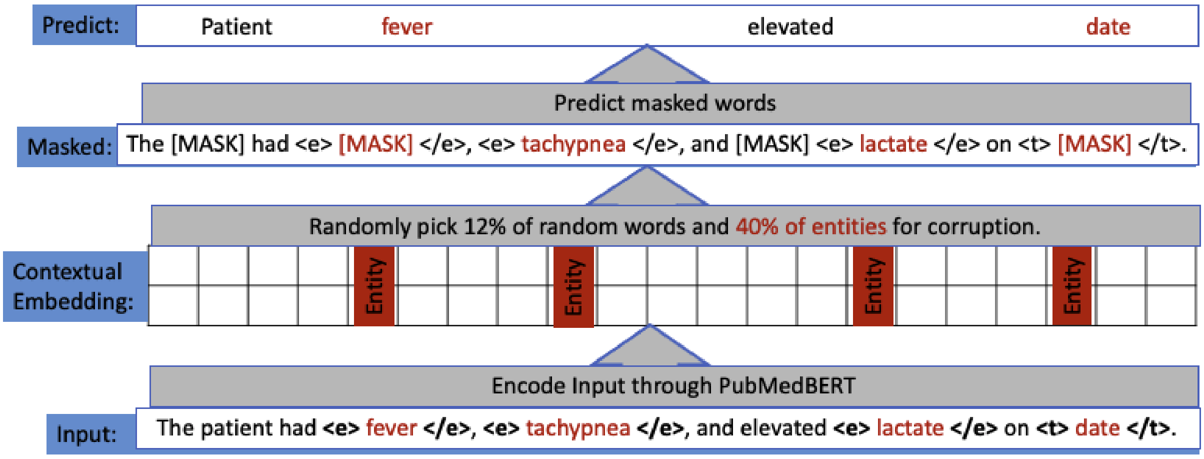

Predict:
Patient
fever
elevated
date
Predict masked words
Masked:
The [MASK] had <e> [MASK] </e>, <e> tachypnea </e>, and [MASK] <e> lactate </e> on <t> [MASK] </t>.
Randomly pick 12% of random words and 40% of entities for corruption.
Contextual Embedding:
Entity
Entity
Entity
Entity
Encode Input through PubMedBERT
Input:
The patient had <e> fever </e>, <e> tachypnea </e>, and elevated <e> lactate </e> on <t> date </t>.

Figure 3: The architecture for continued pretraining of PubMedBERT with the entity-centric masking strategy.

For the TLINK task, the maximal sequence length is set to 100. The models are fine-tuned on the THYME colon cancer training set, parameters are optimized through the THYME colon development set, and tested on the THYME colon cancer test set. The performance is evaluated by the Clinical TempEval evaluation script (Bethard et al., 2017) modified to accommodate the refined temporal relations (Wright-Bettner et al., 2020).

For the DocTimeRel task, the maximal sequence length is set to 30. The models are fine-tuned on the THYME colon cancer training set, parameters are optimized on the THYME colon cancer development set, and tested on both the THYME colon cancer test set and the THYME brain cancer test set for portability evaluation.

For the negation task, the maximal sequence length is set to 64. We follow the same source-target setting as (Lin et al., 2020a) to carry out the cross-domain negation experiments.

For PubMedQA, the maximal sequence length is set to 100 to accommodate both the question and the long answer. The average PubMedQA question length is 14.4 tokens, while the average long answer length is 43.2 tokens (Jin et al., 2019).

Following (Reimers and Gurevych, 2017) in addition to reporting the best scores, we executed multiple runs with varied settings (e.g. random seeds, learning rates, etc.). We compared the distributions with two-sample t-test and report related p-values.

3 Results

Table 2 shows that on the test set of the TLINK task, the best rates for randomly masking entities and words are 40% and 12%, respectively. The table shows only the most successful rates; we considered more entity rates (20%, 40%, 60%, 80%, 100%) and word rates (0%, 8%, 10%, 12%, 14%, 16%). We found that (1) masking non-entity words in addition to masking entities is important as non-entity words capture semantic/syntactic information, and (2) masking too many tokens may make the reconstruction task too hard.

Entity-rate	Word-rate	Overall TLINK F1
30%	10%	0.631
30%	12%	0.644
30%	14%	0.642
40%	10%	0.640
40%	12%	**0.651**
40%	14%	0.642
40%	16%	0.639
50%	12%	0.643
50%	14%	0.641
60%	8%	0.638
60%	10%	0.634
60%	12%	0.631

Table 2: Effect of masking rates for entities (entity-rate) and random words (word-rate) when pretraining PubMedBERT on MIMIC-SMALL for temporal relation extraction. Performance is in terms of overall F1.

Table 3 shows that continuously pretraining PubMedBert (PM) with entity-centric masking (Entity) outperforms random masking (Rand) on both MIMIC-SMALL (p=0.034 with a two-sample t-test) and MIMIC-BIG (p=0.046). The best scores are marked in bold. MIMIC-BIG models have a lower inter-seed variance and slightly better average performance than MIMIC-SMALL. We also combined entity-centric masking with SpanBERT (Joshi et al., 2020) and continuously pre-

			Random Seed					
Mask	BERT	MIMIC	3	4	12	13	42	avg
Rand	PM	Small	0.628	0.641	0.632	0.628	0.641	0.634
Entity	PM	Small	0.643	0.641	0.640	0.634	**0.651**	0.642
Rand	PM	Big	0.634	0.637	0.641	0.634	0.635	0.636
Entity	PM	Big	0.641	0.642	0.640	0.643	**0.648**	0.643
Rand	Span	Small	0.632	0.630	0.632	0.641	0.636	0.634
Entity	Span	Small	0.638	0.635	0.637	**0.643**	**0.643**	0.639

Table 3: Effect of masking strategy (random or entity-centric) on continuously pretraining models (PubMedBERT (PM) or SpanBERT) on MIMIC (BIG or SMALL) for the TLINK task, across different random seeds. Performance is in terms of overall F1.

trained the model on MIMIC-SMALL with different random seeds. The last two rows of Table 3 show that entity-centric masking also helps SpanBERT on the TLINK task (p=0.004).

For our experiments on the downstream tasks, we choose the EntityBERT model continuously pretrained on MIMIC-SMALL with random seed 42 (0.651 F1) and the RandMask model continuously pretrained on MIMIC-BIG with random seed 12 (0.641 F1) because of their best performance. For RandMask models that all get 0.641 F1, we pick the one continuously pretrained on MIMIC-BIG. We fine-tuned them for the specific tasks. The detailed model performance on all TLINK categories is in the bottom two rows in Table 4. The top three rows of Table 4 show the previous best TLINK scores on the same THYME+ corpus by BioBERT and BART-large (Wright-Bettner et al., 2020) and the original PubMedBERT performance.

Table 5 shows that for cross-domain negation detection, out of 12 cross-domain pairs, the entity-centric masking is helpful for 9 pairs. EntityBERT's cross-domain negation average F1 is 0.781 while RandMask's average F1 is 0.773.

Table 6 shows that for DocTimeRel classification, EntityBERT improves over RandMask in the cross-domain setting. When trained and tested in the same colon cancer domain, EntityBERT gets the same overall F1 score as RandMask (0.92 F1). But when trained on colon cancer and tested on brain cancer, EntityBERT significantly improves the overall F1 from 0.69 F1 to 0.72 F1 (p=0.0007).

Table 7 shows PubMedBERT, RandMask and EntityBERT fine-tuning results on the PQA-L test set in the reasoning-free final-phase only setting. It is an extremely low resource setting where there are only 450 training instances used for fine-tuning the models. Results are reported in accuracy using the provided evaluation script. EntityBERT is on par with RandMask (p=0.307) even though these clinical-domain models are both out-of-domain for this biomedical-domain task.

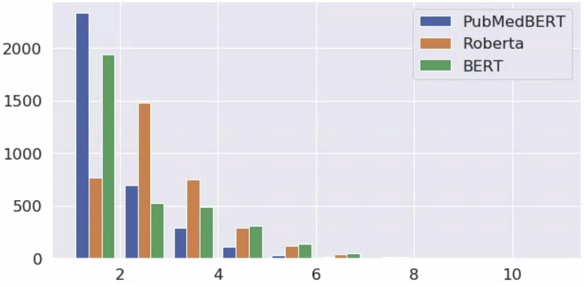


Figure 4: Histogram of token numbers after using different tokenization methods to process all single-token events in THYME Colon training set.

4 Discussion

The benefit of an in-domain vocabulary. To study the in-domain vocabulary's contribution to a clinical task, we extract all 3,471 gold standard events in the THYME colon cancer training set and feed them into the PubMedBERT, RoBERTa, and BERT tokenizers. These events are all single-token events. Figure 4 shows the histogram of tokens per event after tokenization (x-axis shows the number of tokens each event is represented by). We see that PubMedBERT keeps the majority of the events (2,330) as one unit instead of breaking them into multiple sub-words. The BERT tokenizer keeps 1,729 events as one unit. The 601 events that PubMedBERT recognizes but BERT breaks into word pieces are of importance for the TLINK task. If we remove these 601 events from the PubMedBERT vocabulary – forcing them to be broken down into word pieces – the model performance on the TLINK task drops from 0.638 F1 (Table 4 row three) to 0.541 F1, which is the same performance we get if we replace PubMedBERT's tokenizer entirely with BERT's.

What makes a difference? By comparing the TLINK predictions (without applying temporal closure) produced by the best EntityBERT (0.651 F1) and by the best RandMask (0.641 F1), we found that EntityBERT predicted 4,924 correct TLINKs, while RandMask predicted 4,778 correct TLINKs (Table 8). By comparing the entities involved in those correct TLINKs, we found that EntityBERT recognized 131 more entities than Rand-

	BEFORE			BEGINS-ON			CONTAINS			ENDS-ON		
Model	P	R	F1	P	R	F1	P	R	F1	P	R	F1
BioBERT	0.278	0.458	0.346	0.423	0.175	0.248	0.793	0.708	0.748	0.112	0.210	0.146
BART-large	0.300	0.422	0.351	0.378	0.175	0.239	0.796	0.710	0.750	0.124	0.210	0.156
PubMedBERT	0.302	0.493	0.375	0.368	0.172	0.234	0.786	0.734	0.759	0.131	0.227	0.166
RandMask	0.309	0.460	0.370	0.376	0.165	0.229	0.804	0.726	0.763	0.131	0.160	0.144
EntityBERT	0.308	0.467	0.371	0.398	0.179	0.247	0.802	0.739	0.769	0.149	0.185	0.165

	NOTED-ON			OVERLAP			**OVERALL**		
Model	P	R	F1	P	R	F1	P	R	F1
BioBERT	0.786	0.706	0.744	0.353	0.508	0.416	0.696	0.568	0.625
BART-large	0.786	0.707	0.744	0.404	0.470	0.435	0.718	0.558	0.628
PubMedBERT	0.791	0.728	0.758	0.403	0.489	0.442	0.704	0.583	0.638
RandMask	0.767	0.742	0.754	0.404	0.519	0.455	0.717	0.580	0.641
EntityBERT	0.783	0.758	0.770	0.408	0.534	0.462	**0.723**	**0.592**	**0.651**

Table 4: Performance of previous state-of-the-art and the proposed model (EntityBERT) on the TLINK task.

Source	Target	RandMask	EntityBERT
Seed	Strat	0.830	**0.834**
Seed	Mipacq	0.759	**0.761**
Seed	i2b2	0.827	**0.828**
Strat	Seed	0.722	**0.772**
Strat	Mipacq	0.758	0.754
Strat	i2b2	0.881	**0.886**
Mipacq	Seed	0.780	0.772
Mipacq	Strat	0.756	**0.785**
Mipacq	i2b2	0.878	0.871
i2b2	Seed	0.730	**0.732**
i2b2	Strat	0.662	**0.664**
i2b2	Mipacq	0.693	**0.713**
Overall		0.773	**0.781**

Table 5: Effect of masking strategy (Rand or Entity) on cross-domain negation detection. Performance is in terms of F1.

Model	Domain	after	before	bfr/ovlp	overlap	**overall**
RandMask	same	0.88	0.92	0.78	0.94	0.92
EntityBERT	same	0.88	0.92	0.79	0.94	0.92
RandMask	cross	0.65	0.65	0.34	0.74	0.69
EntityBERT	cross	0.64	0.66	0.40	0.77	**0.72**

Table 6: Effect of masking strategy (Rand or Entity) for in-domain (same) and cross-domain settings of the DocTimeRel task. Performance is in terms of F1.

	PubMedBERT	RandMask	EntityBERT
Accuracy	**0.760**	0.738	0.750

Table 7: Performance of models on PubMedQA.

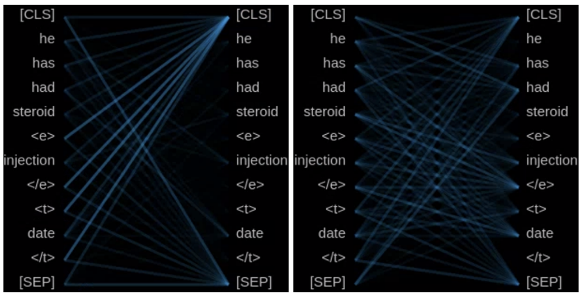
Model	within-sentence	cross-sentence	total
RandMask	4,021	757	4,778
EntityBERT	4,156	768	4,924

Table 8: Correctly predicted TLINK counts by EntityBERT and RandMask before temporal closure.

Mask. Some entities only appear in EntityBERT-identified relations, e.g. *staging*, *hemoglobin*, *finding*, *consideration*, *consider*, *develops*, *request*, *treatment*, *neuropathy*, *carcinoma*, *metastasis*, *injection*, *resected*, and *staged* are involved in multiple relations. Entity-centric masking masks more entities than random masking so that those clinical entities can be better represented by the language model in terms of their semantic and syntactic usage. When the model is fine-tuned for an entity-centric task like the TLINK extraction task, these entities can be better utilized for reasoning relations which they are part of.

In Figure 5 we visualize with BertViz (Vig, 2019) the attention weights of head zero from the last layer of the fine-tuned RandMask and EntityBERT models on the TLINK task for a relation that EntityBERT correctly predicted but RandMask missed. The context is *he has had steroid <e> injection </e> <t> date </t>*. A plausible explanation is that because the key entities, *injection* and *date*, are not well represented in RandMask model, the [CLS] token of RandMask model (Figure 5 (a)) focuses on entity markers, *<e>*, *</e>*, *<t>*, and *</t>*. It may figure out this is an event-time relation but incorrectly infers its type. The [CLS] token of EntityBERT (Figure 5 (b)) bakes in representations of all tokens with knowledge that *injection* is related to *steroid* and *date* is related to *<e> injection </e>*, which shows the key entities are well represented.

Table 8 also shows that the EntityBERT model is most helpful for within-sentence relations (135 more correct within-sentence predictions vs. 11



(a) RandMask (b) EntityBERT

Figure 5: Attention visualization of the last layer of RandMask (a) and EntityBERT (b).

more correct cross-sentence predictions). It could mean the better-learned entity representation is most helpful within a relatively close distance for the current model architecture. To help inferring longer-distanced relations, we may need enhanced model architectures, e.g., DeBERTa (He et al., 2020), that can represent the relative distance between two entities in a disentangled fashion.

Combining Entity-centric Masking with Another Masking Strategy. Some other pre-trained language models like BART (Lewis et al., 2019) and SpanBERT (Joshi et al., 2020) are not pretrained on clinical/biomedical corpora. Yet, they are suitable for clinical tasks in that they mask contiguous random spans instead of individual tokens/word pieces during pretraining – in the clinical domain there are a lot of events and entities that span multiple tokens (e.g., *ascending colon cancer*, *March 11, 2011*). Even without any continued pretraining on a clinical corpus, BART-large achieves 0.628 F1 on the TLINK task (Table 4, row 2), and with continued pretraining on MIMIC-SMALL, SpanBERT-base achieves 0.641 F1 (Table 3, row 5, seed 13). Interestingly, entity-centric masking can further increase SpanBERT performance in the continued pretraining setting (Table 3, last two rows, p=0.004). The reason could be that even though clinical entities could span multiple tokens, a contiguous random span may not be a clinical entity. So, specifically masking clinical entities still has its advantage during continued pretraining a contiguous-span-based language model. We may even see further improved performance if BART or SpanBERT can be pretained from scratch on large clinical/biomedical corpora (however, such a corpus is not available currently!) and then combined with entity-centric masking.

The Strength and Limitations of EntityBERT: EntityBERT assumes that clinical entities are important words, thus if a clinical language model can represent clinical entities better, it will benefit downstream clinical entity-centric tasks. Therefore, such a masking strategy increases the entity concentrations in the masked words during the model pretraining, but does not increase the overall computational loads either for pre-training or for fine-tuning since the overall total number of masked items is similar to random word masking. This is unlike building an additional neural network for selective masking Gu et al. (2020b) or incorporating knowledge graphs Zhang et al. (2019).

The better representation of clinical entities is not only beneficial in an in-domain setting, e.g., the TLINK task, but also effective in a cross-domain setting, e.g., the negation and DocTimeRel tasks. For the DocTimeRel task, both EntityBERT and RandMask achieve very good in-domain performance of 0.92 F1 (see Table 6). In its cross-domain setting, EntityBERT has a clear edge of 0.71 F1 over RandMask 0.69 F1 (see Table 6). Even though some of the improvements may not seem big, they are statistically significant.

We acknowledge some limitations of the current EntityBERT model. First, it is pretrained with a relatively small block size (100 tokens) which is sufficient for a sentence- or a short-paragraph-level reasoning tasks but may be not sufficient for document-level tasks. Second, EntityBERT aims to improve the performance of entity-centric clinical tasks. For tasks that may not directly leverage entities, such as question answering or document classification, entities may still play a supporting role but may not prove as effective. However, we hypothesize that even in those cases its performance would be on-par with RandMask because of its in-domain vocabulary and continued training on a clinical corpus.

Based on the results of Table 7 on PubMedQA, we can see that even though RandMask and EntityBERT models are continuously pretrained from the PubMedBERT model, the continued pretraining on a clinical corpus has made them diverge from its biomedical domain. For the PubMedQA biomedical domain task, the original PubMedBERT model was pretrained from scratch in this target domain, thus performs the best in this task. Yet, even for this non-entity-centric task, EntityBERT performs slightly (but not significantly) better than the RandMask model (0.750 vs. 0.738 in accuracy).

MIMIC-BIG vs. MIMIC-SMALL: RandMask and EntityBERT models pretrained on MIMIC-SMALL perform almost on par with models pretrained on the much bigger corpus, MIMIC-BIG (Table 3) for the TLINK task. The reason could be that even though clinical language varies, the crucial clinical entities are not that many. For example, for the TLINK task, there are only 3,471 unique gold standard events in the training set. Thus, although the size of the corpus is smaller, it could be sufficient for the model to learn representations of the important unique entities.

MIMIC-BIG was created by filtering sentences with fewer than two entities with the goal of capturing pair-wise interactions between events in the language model. One of the limitations of our architecture is its block size. Perhaps with a model that can effectively represent the relative distances, the interactions among entities can be represented better. In addition, by eliminating sentences that only have one or no entity, MIMIC-BIG misses some language phenomena. MIMIC-SMALL, despite its smaller size, thus may encounter more diverse language. This could be the explanation of why an EntityBERT model pretrained on MIMIC-SMALL gets the best TLINK performance (0.651 F1; Table 3 row 2 and Table 4 bottom row).

In the future, we will investigate combining entity-centric masking with DeBERTa (He et al., 2020) with the goal of developing a strategy for a deep neural model that combines entities and their relative position in an input sequence. We will experiment with more flavors of EntityBERT with different block sizes for a wider range of clinical applications. Further testing EntityBERT on a wider range of clinical and biomedical tasks would be helpful for understanding its capabilities.

Acknowledgments

The study was funded by R01LM10090, U24CA248010 and UG3CA243120 from the Unites States National Institutes of Health. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health.

The authors would like to thank the anonymous reviewers for their valuable suggestions and criticism.

The authors would also like to acknowledge Boston Children's Hospital's High-Performance Computing Resources BCH HPC Cluster Enkefalos 2 (E2) made available for conducting the research reported in this publication. Software used in the project was installed and configured by BioGrids (Morin et al., 2013).

References

Daniel Albright, Arrick Lanfranchi, Anwen Fredriksen, William F Styler IV, Colin Warner, Jena D Hwang, Jinho D Choi, Dmitriy Dligach, Rodney D Nielsen, James Martin, et al. 2013. Towards comprehensive syntactic and semantic annotations of the clinical narrative. *Journal of the American Medical Informatics Association*, 20(5):922–930.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Andrew L Beam, Benjamin Kompa, Allen Schmaltz, Inbar Fried, Griffin Weber, Nathan Palmer, Xu Shi, Tianxi Cai, and Isaac S Kohane. 2019. Clinical concept embeddings learned from massive sources of multimodal medical data. In *PACIFIC SYMPOSIUM ON BIOCOMPUTING 2020*, pages 295–306. World Scientific.

Iz Beltagy, Kyle Lo, and Arman Cohan. 2019. Scibert: A pretrained language model for scientific text. *arXiv preprint arXiv:1903.10676*.

Steven Bethard, Leon Derczynski, Guergana Savova, Guergana Savova, James Pustejovsky, and Marc Verhagen. 2015. Semeval-2015 task 6: Clinical tempeval. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 806–814.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical tempeval. *Proceedings of SemEval*, pages 1052–1062.

Steven Bethard, Guergana Savova, Martha Palmer, James Pustejovsky, and Marc Verhagen. 2017. Semeval-2017 task 12: Clinical tempeval. *Proceedings of the 11th International Workshop on Semantic Evaluation (SemEval-2017)*, pages 563–570.

Wendy W Chapman, Will Bridewell, Paul Hanbury, Gregory F Cooper, and Bruce G Buchanan. 2001. A simple algorithm for identifying negated findings and diseases in discharge summaries. *Journal of biomedical informatics*, 34(5):301–310.

Qingyu Chen, Kyubum Lee, Shankai Yan, Sun Kim, Chih-Hsuan Wei, and Zhiyong Lu. 2020. Bioconceptvec: creating and evaluating literature-based biomedical concept embeddings on a large scale. *PLoS computational biology*, 16(4):e1007617.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Dmitriy Dligach, Timothy Miller, Chen Lin, Steven Bethard, and Guergana Savova. 2017. Neural temporal relation extraction. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 2, Short Papers*, pages 746–751.

Noémie Elhadad, Sameer Pradhan, Sharon Gorman, Suresh Manandhar, Wendy Chapman, and Guergana Savova. 2015. SemEval-2015 task 14: Analysis of clinical text. In *Proceedings of the 9th International Workshop on Semantic Evaluation (SemEval 2015)*, pages 303–310, Denver, Colorado. Association for Computational Linguistics.

Philip Gage. 1994. A new algorithm for data compression. *C Users Journal*, 12(2):23–38.

Ary L Goldberger, Luis AN Amaral, Leon Glass, Jeffrey M Hausdorff, Plamen Ch Ivanov, Roger G Mark, Joseph E Mietus, George B Moody, Chung-Kang Peng, and H Eugene Stanley. 2000. Physiobank, physiotoolkit, and physionet: components of a new research resource for complex physiologic signals. *circulation*, 101(23):e215–e220.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2020a. Domain-specific language model pretraining for biomedical natural language processing. *arXiv preprint arXiv:2007.15779*.

Yuxian Gu, Zhengyan Zhang, Xiaozhi Wang, Zhiyuan Liu, and Maosong Sun. 2020b. Train no evil: Selective masking for task-guided pre-training. *arXiv preprint arXiv:2004.09733*.

Henk Harkema, John N Dowling, Tyler Thornblade, and Wendy W Chapman. 2009. Context: an algorithm for determining negation, experiencer, and temporal status from clinical reports. *Journal of biomedical informatics*, 42(5):839–851.

Pengcheng He, Xiaodong Liu, Jianfeng Gao, and Weizhu Chen. 2020. Deberta: Decoding-enhanced bert with disentangled attention. *arXiv preprint arXiv:2006.03654*.

Qiao Jin, Bhuwan Dhingra, Zhengping Liu, William W Cohen, and Xinghua Lu. 2019. Pubmedqa: A dataset for biomedical research question answering. *arXiv preprint arXiv:1909.06146*.

Alistair EW Johnson, Tom J Pollard, Lu Shen, H Lehman Li-Wei, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Artuur Leeuwenberg and Marie Francine Moens. 2017. Structured learning for temporal relation extraction from clinical records. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 1150–1158.

Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. Bart: Denoising sequence-to-sequence pre-training for natural language generation, translation, and comprehension. *arXiv preprint arXiv:1910.13461*.

Chen Lin, Steven Bethard, Dmitriy Dligach, Farig Sadeque, Guergana Savova, and Timothy A Miller. 2020a. Does bert need domain adaptation for clinical negation detection? *Journal of the American Medical Informatics Association*, 27(4):584–591.

Chen Lin, Timothy Miller, Dmitriy Dligach, Hadi Amiri, Steven Bethard, and Guergana Savova. 2018. Self-training improves recurrent neural networks performance for temporal relation extraction. In *Proceedings of the Ninth International Workshop on Health Text Mining and Information Analysis*, pages 165–176.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2017. Representations of time expressions for temporal relation extraction with convolutional neural networks. In *BioNLP 2017*, pages 322–327.

Chen Lin, Timothy Miller, Dmitriy Dligach, Steven Bethard, and Guergana Savova. 2019. A bert-based universal model for both within-and cross-sentence clinical temporal relation extraction. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 65–71.

Chen Lin, Timothy Miller, Dmitriy Dligach, Farig Sadeque, Steven Bethard, and Guergana Savova. 2020b. A bert-based one-pass multi-task model for clinical temporal relation extraction. In *Proceedings of the 19th SIGBioMed Workshop on Biomedical Language Processing*, pages 70–75.

Yinhan Liu, Myle Ott, Naman Goyal, Jingfei Du, Mandar Joshi, Danqi Chen, Omer Levy, Mike Lewis, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Roberta: A robustly optimized bert pretraining approach. *arXiv preprint arXiv:1907.11692*.

Xinbo Lv, Yi Guan, Jinfeng Yang, and Jiawei Wu. 2016. Clinical relation extraction with deep learning. *International Journal of Hybrid Information Technology*, 9(7):237–248.

Saeed Mehrabi, Anand Krishnan, Sunghwan Sohn, Alexandra M Roch, Heidi Schmidt, Joe Kesterson, Chris Beesley, Paul Dexter, C Max Schmidt, Hongfang Liu, et al. 2015. Deepen: A negation detection system for clinical text incorporating dependency relation into negex. *Journal of biomedical informatics*, 54:213–219.

Timothy Miller, Steven Bethard, Hadi Amiri, and Guergana Savova. 2017. Unsupervised domain adaptation for clinical negation detection. In *BioNLP 2017*, pages 165–170.

Andrew Morin, Ben Eisenbraun, Jason Key, Paul C Sanschagrin, Michael A Timony, Michelle Ottaviano, and Piotr Sliz. 2013. Cutting edge: Collaboration gets the most out of software. *elife*, 2:e01456.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of bert and elmo on ten benchmarking datasets.

Sameer Pradhan, Noémie Elhadad, Wendy Chapman, Suresh Manandhar, and Guergana Savova. 2014. SemEval-2014 task 7: Analysis of clinical text. In *Proceedings of the 8th International Workshop on Semantic Evaluation (SemEval 2014)*, pages 54–62, Dublin, Ireland. Association for Computational Linguistics.

James Pustejovsky, José M Castano, Robert Ingria, Roser Sauri, Robert J Gaizauskas, Andrea Setzer, Graham Katz, and Dragomir R Radev. 2003. Timeml: Robust specification of event and temporal expressions in text. *New directions in question answering*, 3:28–34.

James Pustejovsky and Amber Stubbs. 2011. Increasing informativeness in temporal annotation. In *Proceedings of the 5th Linguistic Annotation Workshop*, pages 152–160. Association for Computational Linguistics.

Colin Raffel, Noam Shazeer, Adam Roberts, Katherine Lee, Sharan Narang, Michael Matena, Yanqi Zhou, Wei Li, and Peter J Liu. 2019. Exploring the limits of transfer learning with a unified text-to-text transformer. *arXiv preprint arXiv:1910.10683*.

Nils Reimers and Iryna Gurevych. 2017. Reporting score distributions makes a difference: Performance study of lstm-networks for sequence tagging. *arXiv preprint arXiv:1707.09861*.

Guergana K Savova, James J Masanz, Philip V Ogren, Jiaping Zheng, Sunghwan Sohn, Karin C Kipper-Schuler, and Christopher G Chute. 2010. Mayo clinical text analysis and knowledge extraction system (ctakes): architecture, component evaluation and applications. *Journal of the American Medical Informatics Association*, 17(5):507–513.

William F Styler IV, Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C De Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. 2014. Temporal annotation in the clinical domain. *Transactions of the association for computational linguistics*, 2:143–154.

Özlem Uzuner, Brett R South, Shuying Shen, and Scott L DuVall. 2011. 2010 i2b2/va challenge on concepts, assertions, and relations in clinical text. *Journal of the American Medical Informatics Association*, 18(5):552–556.

Jesse Vig. 2019. A multiscale visualization of attention in the transformer model. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics: System Demonstrations*, pages 37–42, Florence, Italy. Association for Computational Linguistics.

Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, et al. 2019. Huggingface's transformers: State-of-the-art natural language processing. *arXiv preprint arXiv:1910.03771*.

Kristin Wright-Bettner, Chen Lin, Timothy Miller, Steven Bethard, Dmitriy Dligach, Martha Palmer, James H Martin, and Guergana Savova. 2020. Defining and learning refined temporal relations in the clinical narrative. In *Proceedings of the 11th International Workshop on Health Text Mining and Information Analysis*, pages 104–114.

Stephen Wu, Timothy Miller, James Masanz, Matt Coarr, Scott Halgrim, David Carrell, and Cheryl Clark. 2014. Negation's not solved: generalizability versus optimizability in clinical natural language processing. *PloS one*, 9(11):e112774.

Yonghui Wu, Min Jiang, Jun Xu, Degui Zhi, and Hua Xu. 2017. Clinical named entity recognition using deep learning models. In *AMIA Annual Symposium Proceedings*, volume 2017, page 1812. American Medical Informatics Association.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. 2016. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*.

Zhengyan Zhang, Xu Han, Zhiyuan Liu, Xin Jiang, Maosong Sun, and Qun Liu. 2019. ERNIE: enhanced language representation with informative entities. *CoRR*, abs/1905.07129.

Morteza Ziyadi, Yuting Sun, Abhishek Goswami, Jade Huang, and Weizhu Chen. 2020. Example-based named entity recognition. *arXiv preprint arXiv:2008.10570*.