Enhancing Clinical Concept Extraction with Contextual Embedding

Yuqi Si  Jingqi Wang  Hua Xu  Kirk Roberts
School of Biomedical Informatics
The University of Texas Health Science Center at Houston
{yuqi.si,kirk.roberts}@uth.tmc.edu

Abstract
Neural network-based representations (“embeddings”) have dramatically advanced natural language processing (NLP) tasks in the past few years. This certainly holds for clinical concept extraction, especially when combined with deep learning-based models. Recently, however, more advanced embedding methods and representations (e.g., ELMo, BERT) have further pushed the state-of-the-art in NLP. While these certainly improve clinical concept extraction as well, there are no commonly agreed upon best practices for how to integrate these representations for extracting concepts. The purpose of this study, then, is to explore the space of possible options in utilizing these new models, including comparing these to more traditional word embedding methods (word2vec, GloVe, fastText). We evaluate a battery of embedding methods on four clinical concept extraction corpora, explore effects of pre-training on extraction performance, and present an intuitive way to understand the semantic information encoded by advanced contextualized representations. Notably, we achieved new state-of-the-art performances across all four corpora.

1 Introduction
Concept extraction is the most common clinical natural language processing (NLP) task (Tang et al., 2013; Kundeti et al., 2016; Unanue et al., 2017; Wang et al., 2018b), and a precursor to downstream tasks such as relations (Rink et al., 2011), frames (Gupta et al., 2018; Si and Roberts, 2018), co-reference (Lee et al., 2011), and phenotyping (Xu et al., 2011; Velupillai et al., 2018). Corpora such as those from i2b2 (Uzuner et al., 2011; Sun et al., 2013; Stubbs et al., 2015), ShARe/CLEF (Suominen et al., 2013; Kelly et al., 2014), and SemEval (Pradhan et al., 2014; Elhadad et al., 2015; Bethard et al., 2016) act as evaluation benchmarks and datasets for training machine learning (ML) models.
Meanwhile, neural network-based representations continue to advance nearly all areas of NLP, from question answering (Shen et al., 2017) to named entity recognition (Chang et al., 2015; Wu et al., 2015; Habibi et al., 2017; Unanue et al., 2017; Florez et al., 2018) (a close analog of concept extraction). Recent advances in contextualized representations, including ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018), have pushed performance even further. These have demonstrated that relatively simple downstream models using contextualized embeddings can outperform complex models (Seo et al., 2016) using embeddings such as word2vec (Mikolov et al., 2013) and GloVe (Pennington et al., 2014).

In this paper we aim to explore the potential impact these representations have on clinical concept extraction. Our contributions include:

1. An evaluation exploring numerous embedding methods: word2vec (Mikolov et al., 2013), GloVe (Pennington et al., 2014), fastText (Bojanowski et al., 2016), ELMo (Peters et al., 2018), and BERT (Devlin et al., 2018).
2. An analysis covering four clinical concept corpora, demonstrating the generalizability of these methods.
3. A performance increase for clinical concept extraction that achieves state-of-the-art results on all four corpora.
4. A demonstration of the effect of pre-training on clinical corpora versus larger open domain corpora, an important trade-off in clinical NLP (Roberts, 2016).
5. A detailed analysis of the effect of pre-training time when starting from pre-built
open domain models, which is important due to the long pre-training times of methods such as ELMo and BERT.

2 Background

This section introduces the theoretical knowledge that supports the shifts from word-level embeddings to contextualized embeddings.

2.1 Word Embedding Models

Word-level vector representation methods learn a real-valued vector to represent a single word. One of the most prominent methods for word-level representation is word2vec (Mikolov et al., 2013). So far, word2vec has widely established its effectiveness for achieving state-of-the-art performances in a variety of clinical NLP tasks (Wang et al., 2018a). GloVe (Pennington et al., 2014) is another unsupervised learning approach to obtain a vector representation for single word. Unlike word2vec, GloVe is a statistical model that aggregates both a global matrix factorization and a local context window. The learning relies on dimensionality reduction on the co-occurrence count matrix based on how frequently a word appears in a context. fastText (Bojanowski et al., 2016) is also an established library for word representations. Unlike word2vec and GloVe, fastText considers individual words as character n-grams. For instance, “cold” is made of the n-grams “c”, “co”, “col”, “cold”, “o”, “ol”, “old”, “l”, “ld”, and “d”. This approach enables handling of infrequent words that are not present in the training vocabulary, alleviating some out-of-vocabulary issues.

ELMo (Peters et al., 2018) is one such deep contextualized word representation. Unlike the above traditional word embeddings that constitute a single vector for each word and the vector remains stable in downstream tasks, this contextualized word representation can capture the context information and dynamically alter the representation. At the training time, a language model is created from a large text corpus and learns the context-sensitive embeddings. The training step of learning this context-sensitive embeddings is known as the pre-training. After pre-training, the context-sensitive embedding of each word will be fed into the sentences for downstream tasks. The downstream task learns the shared weights of the inner state of pre-trained language model by optimizing the loss of downstream task.

BERT (Devlin et al., 2018) is also a contextualized word representation model. Compared to ELMo, BERT is deeper in how it handles contextual information due to a deep bidirectional transformer for encoding sentences. Additionally, in terms of the strategy for how to incorporate these models into the downstream task, ELMo is a feature-based language representation while BERT is a fine-tuning approach. The feature-based strategy is similar to traditional word embedding methods that consider the embedding as the input feature of downstream task. The fine-tuning approach, on the other hand, adjusts the entire language model on the downstream task to achieve a task-specific architecture. So while ELMo may be incorporated into a downstream Bi-LSTM, BERT simply replace the downstream model. This fine-tuning strategy is more likely to make use of the encoded information in the pre-trained language models.
2.2 Clinical Concept Extraction

Up to now, the most prominent model for clinical concept extraction is a bidirectional Long Short-Term Memory with Conditional Random Field (Bi-LSTM CRF) architecture (Chalapathy et al., 2016; Habibi et al., 2017; Florez et al., 2018). The bidirectional LSTM-based RNN captures both forward and backward information in the sentence and the CRF layer considers sequential output correlations in the decoding layer using the Viterbi algorithm. All components combine together to predict the entity labels for all words in the sentence.

3 Methods

In this paper, we consider both off-the-shelf embeddings from the open domain as well as pre-training clinical domain embeddings on clinical notes from MIMIC-III (Johnson et al., 2016), which is a public database of Intensive Care Unit (ICU) patients.

For the traditional word embedding experiments, the static embeddings are fed into a Bi-LSTM CRF architecture. All words that occur at least five times in the corpus are included and infrequent words are denoted as UNK. To compensate for the loss due of those unknown words, character embeddings for each word are included.

For ELMo, the context-independent embeddings with trainable weights are used to form context-dependent embeddings, then fed the context-dependent embeddings into downstream task. Specifically, the context-dependent embedding is obtained through a low-dimensional projection and a highway connection after a stacked layer of a character-based Convolutional Neural Network (char-CNN) and a two-layer Bi-LSTM language model (bi-LM). Thus, the contextual word embedding is formed with a trainable aggregation of highly-connected bi-LM. Because the context-independent embeddings already consider representation of characters, it is not necessary to learn a character embedding input for the Bi-LSTM in concept extraction. Finally, the contextualized word embedding for each word is fed into the existing sequence labeling architecture, Bi-LSTM CRF, to predict the label for each token.

For BERT, both the BERT\textsubscript{BASE}(General) and BERT\textsubscript{LARGE}(General) models are fine-tuned on the downstream task. Because BERT integrates sufficient label-correlation information, the CRF layer is abandoned and only a Bi-LSTM architecture is used for sequence labeling. Additionally, two clinical domain embedding models are pre-trained on MIMIC-III, initiated from the BERT\textsubscript{BASE} and BERT\textsubscript{LARGE} checkpoints, which we refer to as BERT\textsubscript{BASE(MIMIC)} and BERT\textsubscript{LARGE(MIMIC)}, respectively.

4 Datasets and Experiments

4.1 Datasets

| Dataset     | Subset | #. Notes | #. Entities |
|-------------|--------|----------|-------------|
| i2b2 2010   | Train  | 349      | 27,837      |
|             | Test   | 477      | 45,009      |
| i2b2 2012   | Train  | 190      | 16,468      |
|             | Test   | 120      | 13,594      |
| SemEval 2014| Train  | 199      | 5,816       |
| SemEval 2014| Task 7 | 99       | 5,351       |
| SemEval 2015| Train  | 298      | 11,167      |
| SemEval 2015| Task 14| 133      | 7,998       |

Table 1: Descriptive statistics for concept extraction datasets

Our experiments are performed on four open challenges, the 2010 i2b2/VA challenge (Uzuner et al., 2011), the 2012 i2b2 challenge (Sun et al., 2013), the SemEval 2014 Task 7 (Pradhan et al., 2014) and the SemEval 2015 Task 14 (Elhadad et al., 2015). The descriptive statistics for the datasets are shown in Table 1. The 2010 i2b2/VA challenge data contains a total of 349 training and 477 testing reports, where clinical entities of PROBLEM, TEST and TREATMENT were manually annotated. The 2012 i2b2 challenge data contains a total of 190 training and 120 testing discharge summaries, and 6 clinical entities consisting of PROBLEM, TEST, TREATMENT, CLINICAL DEPARTMENT, EVIDENTIAL, and OCCURRENCE. The SemEval 2014 Task 7 data contains 199 training and 99 testing reports with gold standard annotations of DISEASE DISORDER. The SemEval 2015 Task 14 data consists of a total of 298 training and 133 testing reports with gold standard annotations of DISEASE DISORDER. For our experiments, we only focus on the concept extraction in four tasks. For the SemEval tasks, the disjoint entities are handled with “BIOHD” tagging schema (Tang et al., 2015).
The clinical embeddings are trained on MIMIC III (Johnson et al., 2016), which consists of almost 2 million clinical notes. Notes that have an ERROR tag are first removed, ending up with 1,908,359 notes and 786,414,528 tokens and a vocabulary of size 712,286. All the unstructured clinical notes are preprocessing in the same way to split sentences and tokenize words. For pre-training traditional word embeddings, words are lowercased, as is standard practice. For pre-training ELMo and BERT, casing is preserved.

### 4.2 Experimental Setting

#### 4.2.1 Concept Extraction

Concept extraction is based on the model proposed in Lample et al., (2016), a Bi-LSTM CRF architecture. For all embedding methods, we use the same hyperparameters setting: hidden unit dimension at 512, dropout probability at 0.5, learning rate at 0.001, learning rate decay at 0.9, and Adam as the optimization algorithm. Early stopping of training is set to 5 epochs without improvement to prevent overfitting.

#### 4.2.2 Pre-training of Clinical Embeddings

Across embedding methods, two different scenarios of pre-training are investigated and compared:

1. Off-the-shelf embeddings from the official release, referred to as the General model.

2. Pre-train on MIMIC-III to obtain clinical domain specific embedding, referred to as the MIMIC model.

In the first scenario, more details related to the embedding models are shown in Table 2. We also apply BioBERT (Lee et al., 2019), which is the most recent pre-trained model on biomedical literature initiated from BERTBASE. In the second scenario, we follow the instructions provided to preprocess MIMIC-III notes into required formats.

For word2vec, we adopt the pre-trained word2vec embedding model of clinical domain from our previous study (Roberts, 2016), which concatenates 300 dimension embeddings pre-trained from MIMIC-III, MEDLINE and WebMD. For GloVe, we apply most of the default settings: minimum count of word is 5, window size of 15, and iterations of 15, and increase the embedding size to 300 to match with the off-the-shelf GloVe embedding model. For fastText, the following parameters are set in pre-training stage: embedding size of 300, 15 epochs, learning rate of 0.05, window size of 15 and minimum count of word is 5.

For ELMo, the hyperparameter setting for pre-training follows the default in Peters et al. (2018). Specifically, a char-CNN embedding layer is applied with 16 dimensional character embeddings, filter widths of [1, 2, 3, 4, 5, 6, 7] with respective [32, 32, 64, 128, 256, 512, 512] number of filters. After that, a two-layer Bi-LSTM with 4,096 hidden units in each layer is added. The output of the final bi-LM language model is projected to 512 dimensions with a high-way connection. The total number of tokens in pre-training on MIMIC-III was 786,414,528. MIMIC-III was split into a training corpus (80%) for pre-training and a held-out testing corpus (20%) for evaluating perplexity. The pre-training step is performed on the training corpus for 10 epochs. The average perplexity on the testing corpus is 9.929.

For BERT, two clinical-domain models initial-

| Method | Resource (#. Tokens/ #. Vocab) | Size | Language model |
|--------|---------------------------------|------|----------------|
| GloVe  | Gigaword5 + Wikipedia2014 (6B/0.4M) | 300  | NA             |
| fastText | Wikipedia 2017+ UMBC webbase corpus and statmt.org news (16B/1M) | 300  | NA             |
| ELMo   | WMT 2008-2012 + Wikipedia (5.5B/0.7M) | 512  | 2-layer, 4096-hidden, 93.6M parameters |
| BERTBASE | BooksCorpus+ English Wikipedia (3.3B/0.03M*) | 768  | 12-layer, 768-hidden, 12-heads, 110M parameters |
| BERTLARGE | BooksCorpus + English Wikipedia (3.3B/0.03M*) | 1024 | 24-layer, 1024-hidden, 16-heads, 340M parameters |

Table 2: Resources of off-the-shelf embeddings from open domain. (*Vocabulary size calculated after word-piece tokenization)
Table 3: Test set comparison in exact F-measure of embedding methods across tasks.

| Method       | i2b2 2010 | i2b2 2012 | Semeval 2014 Task 7 | Semeval 2015 Task 14 |
|--------------|-----------|-----------|---------------------|----------------------|
|              | General   | MIMIC     | General             | MIMIC                |
| w2v          | -         | 82.67     | -                   | 73.77                |
| GloVe        | 84.08     | 85.07     | 74.95               | 75.27                |
| fastText     | 83.46     | 84.19     | 73.24               | 74.83                |
| ELMo         | 83.83     | 87.80     | 76.61               | 80.5                 |
| BERT_{BASE}  | 84.33     | 89.55     | 76.62               | 80.34                |
| BERT_{LARGE} | 85.48     | 90.25     | 78.14               | 80.91                |
| BioBERT      | 84.76     | -         | 77.77               | -                    |

ized from BERT_{BASE} and BERT_{LARGE} are pre-trained. Unless specified, we follow the detailed instruction provided to set up the pre-training parameters, as other options are experimented and it has been concluded that this is a useful recipe when pre-training from their released model (e.g., poor model convergence). The vocabulary list consisting of 28,996 word-pieced tokens applied in BERT_{BASE} and BERT_{LARGE} are adopted. According to their paper, the performance on the downstream tasks decrease as the training steps increase; thus we decide to save the intermediate checkpoint (every 20,000 steps) and report the performance of intermediate models on the downstream task. The stopping checkpoints for BERT_{BASE}(MIMIC) and BERT_{LARGE}(MIMIC) are set at 0.6 million steps.

4.2.3 Fine-tuning BERT

Fine-tuning the BERT clinical model on the downstream task needs some detailed adjustments. First, the model is more likely to fall into a local minimum if the Bi-LSTM is randomly initialized. To mitigate this problem, Xavier Initialization is performed on the weights of the Bi-LSTM output layer, namely set the weights initialized with a variance of the inverse of the hidden unit size. The early stopping technique is applied to prevent overfitting. Finally, the post-processing steps are conducted to align the BERT output with the concept gold standard, including handling truncated sentences and word-pieced tokenizations.

4.2.4 Evaluation

10% of the official training set are used to constitute the development set and the official test set are used to report the performances. The performances are calculated by precision, recall and F-measure for strict/exact matching.

The pre-training BERT experiments are implemented in TensorFlow (Abadi et al., 2016) on a NVidia Tesla V100 GPU (32G), other experiments are performed in TensorFlow on a NVIDIA Quadro M5000 (8G).

5 Results

5.1 Embedding Comparison

The performances on the test set between different embedding methods on four clinical concept extraction tasks are reported in Table 3. The performance is evaluated in exact matching F-measure. In general, as for the same embedding method, embeddings pre-trained on clinical corpus performed better than those pre-trained on open domain corpus.

For i2b2 2010, the best performance is achieved by BERT_{LARGE}(MIMIC) with an F-measure of 90.25. It improves the performance by 5.18 over the best performance of the traditional embeddings achieved by GloVe (MIMIC) with F measure of 85.07. As expected, both ELMo and BERT clinical embeddings outperform the off-the-shelf embeddings with relative increase up to 10%.

The best performance on the i2b2 2012 task is achieved by BERT_{LARGE}(MIMIC) with an F-measure of 80.91 across all the methods. It increases F-measure by 5.64 over GloVe(MIMIC), which obtains the best score (75.27) among the traditional embedding methods. Not surprisingly, ELMo and BERT with pre-trained clinical corpus exceed the off-the-shelf open domain models.

For SemEval 2014 task, since the experiments are performed only in concept extraction, the mod-
els are evaluated using the official evaluation script from SemEval 2014 task. Both tasks share the same proportion of training set, and SemEval 2015 task uses the 99 testing notes from SemEval 2014 task as the development set. The best performance on 2015 task is achieved by BERT\textsubscript{LARGE}(MIMIC) with an F-measure of 81.65 over the rest of the embedding models.

The detailed performance for each entity category including \textsc{Problem}, \textsc{Test} and \textsc{Treatment} on 2010 task is shown in Table 4. At first glance, the three events have similar performance, yet \textsc{Treatment} gets slightly poorer results. This is reasonable considering different amount of samples in each category (# \textsc{Problem} > # \textsc{Treatment} > # \textsc{Test}).

|               | GloVe   | FastText | ELMo     | BERT\textsubscript{BASE} | BERT\textsubscript{LARGE} |
|---------------|---------|----------|----------|--------------------------|---------------------------|
| \textsc{Problem} | 85.08   | 84.32    | 88.76    | 89.61                    | 89.26                     |
| \textsc{Test}   | 84.96   | 84.01    | 87.39    | 88.09                    | 88.8                      |
| \textsc{Treatment} | 84.73   | 83.89    | 86.98    | 88.3                     | 89.14                     |

Table 4: Performance of each label category with pre-trained MIMIC models on 2010 task.

Table 5 shows the results for each event type on the 2012 task with embeddings pre-trained from MIMIC-III. Generally, the most improvement by contextualized-level representation over traditional embedding is achieved on the \textsc{Problem} type (BERT\textsubscript{LARGE}: 86.1, GloVe: 77.83). This is quite reasonable because in the actual clinical notes, certain types of diseases or conditions normally appear in certain types of surrounding context with similar grammar structures. Thus it is necessarily important to take advantage of contextualized representations to capture contextual information for that particular entity. \textsc{Occurrence} has the lowest performance across all methods. It is assumed that \textsc{Occurrence} is describing patients status in clinical notes and there is no obvious trend for this type of term.

|               | GloVe   | FastText | ELMo     | BERT\textsubscript{BASE} | BERT\textsubscript{LARGE} |
|---------------|---------|----------|----------|--------------------------|---------------------------|
| \textsc{Problem} | 77.83   | 75.35    | 84.1     | 85.91                    | 86.1                      |
| \textsc{Test}   | 81.26   | 76.94    | 84.76    | 86.88                    | 86.56                     |
| \textsc{Treatment} | 78.52   | 76.88    | 83.9     | 84.27                    | 85.09                     |
| \textsc{Clinical Dept} | 77.92   | 77.27    | 83.71    | 77.92                    | 78.23                     |
| \textsc{Evidential} | 74.26   | 72.94    | 72.95    | 74.21                    | 74.96                     |
| \textsc{Occurrence} | 64.19   | 61.02    | 66.27    | 62.36                    | 65.65                     |

Table 5: Performance of each label category with pre-trained MIMIC models on 2012 task.

5.2 Efficiency of pre-trained clinical model

The efficiency of pre-trained ELMo and BERT models are investigated by reporting the loss during pre-training steps and by evaluating the intermediate checkpoints on downstream tasks. It is observed for both ELMo and BERT at their pre-training stages, the train perplexity or loss decreases as the steps increase, indicating that the language model is actually adapting to the clinical corpus. If there is no intervention to stop the pre-training process, it will lead to a very small number of loss value (more details in Supplement). However, this will bring to another common issue that the model might be overfitting on the training set.

![Figure 2: Performances on the i2b2 2010 task governed by the steps of pre-training epochs on ELMo (MIMIC) and BERT\textsubscript{BASE} (MIMIC).](image)

Using i2b2 2010 as the downstream task, the final performance at each intermediate checkpoint of the pre-trained model is shown in Figure 2. Apparent trends appear in both pre-trained models. As the pre-training proceeds, the performance of the downstream task decreases after a certain number of iterations. ELMo(MIMIC) model has more obvious such trend and the performance remains stable despite of increasing pre-training steps (the maximum F-measure reaches 87.80 at step 280K).

BERT\textsubscript{BASE}(MIMIC) also has similar issues and the maximum F-measure is 89.55 at step 340K. In addition, BERT\textsubscript{BASE}(MIMIC) tends to be un-smooth at the beginning and the similar trend appears to be near the end. This overfitting problem is feasible to the actual situation, since the more steps pre-training on MIMIC-III, the more the pre-trained model would lose its original information. We hope this is a takeaway lesson for the clinical NLP community when they intend to generate their own pre-trained model from a clinical corpus.
6 Discussion

This paper explores the effects of numerous embedding methods in four shared clinical concept extraction tasks. It is demonstrated the higher efficacy of domain-specific embedding models over open-domain embedding models. All types of embeddings enable consistent gains in concept extraction tasks when pre-trained on clinical domain corpus. Further, the contextualized embeddings outperform traditional embeddings in performance. Specifically, large improvements can be achieved by pre-training a deep language model from a large corpus, followed by a task-specific fine-tuning.

6.1 State-of-the-art Comparison

Among the four shared challenges of clinical concept extraction, the i2b2 2012 task report the partial matching F-measure as the organizers reported in Sun et al., (2013) and other three tasks report the exact matching F-measure. Currently, the state-of-the-art models in the i2b2 2010 challenges, the 2012 challenges, the SemEval 2014 task 7, and the Semeval 2015 Task 14 are reported with F-measure of 88.60 (Zhu et al., 2018), 92.29 (Liu et al., 2017), 80.3 (Tang et al., 2015) and 81.3 (Zhang et al., 2014), respectively. With the most advanced language model representation method pre-trained on a large clinical corpus, namely BERT\text{LARGE}(MIMIC), we achieved new state-of-the-art performances across all tasks. BERT\text{LARGE}(MIMIC) outperform the state-of-the-art models on all four tasks with respective F-measures of 90.25, 93.18 (partial F-measure), 80.74, and 81.65.

6.2 Semantic Information from Contextual Embeddings

In addition, we attempt to explore the semantic information captured by contextualized representation and infer that the contextualized embedding can encode information that a single word vector fails to. We select 10 sentences from both web texts and clinical notes in which the word cold appears (The actual sentences can be found in Supplement). The embedding vectors of cold in 20 sentences from two embedding models, ELMo(General) and ELMo(MIMIC), were derived. That results in 20 vectors for the same word across two embeddings. For each embedding method, PCA clustering is performed to reduce dimension of the large embedding size into 2 dimensions.

PCA clustering is shown in Figure 3. As expected, the vectors of cold generated by ELMo(MIMIC) are split into two groups, while the vectors from ELMo(General) are mixed. ELMo (MIMIC) is able to discriminate the different meanings of the word cold, specifically between temperature and symptom. The embedding vectors of a word in different sentences are not exactly the same, whereas depending in part on the surrounding context.

7 Conclusion

In this paper, we present an analysis of different word embedding methods and investigate their effectiveness on four clinical concept extraction tasks. We compare between traditional word representation methods as well as the advanced contextualized representation methods. We also build pre-trained contextualized embeddings using large clinical corpus and compare the performance with off-the-shelf pre-trained models on open domain data. Primarily, the efficacy of contextualized embeddings over traditional word vector representations are highlighted by comparing the performances on clinical concept extractions. Contextualized embeddings also provide interesting semantic information that is not accounted for in traditional word representations. Further, our results highlight the benefits of embeddings through unsupervised learning on clinical text corpora, which achieve higher performance than off-the-shell embedding models across all tasks.
Acknowledgments

This work was supported by the U.S. National Library of Medicine, National Institutes of Health (NIH), under award R00LM012104, as well as the Cancer Prevention Research Institute of Texas (CPRIT), under awards RP170668 and RR180012.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: a system for large-scale machine learning. In OSDI, volume 16, pages 265–283.

Steven Bethard, Guergana Savova, Wei-Te Chen, Leon Derczynski, James Pustejovsky, and Marc Verhagen. 2016. Semeval-2016 task 12: Clinical temporal. In Proceedings of the 10th International Workshop on Semantic Evaluation (SemEval-2016), pages 1052–1062.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. arXiv preprint arXiv:1607.04606.

Raghavendra Chalapathy, Ehsan Zare Borzeshi, and Massimo Piccardi. 2016. Bidirectional lstm-crf for clinical concept extraction. arXiv preprint arXiv:1611.08373.

FX Chang, J Guo, WR Xu, and S Rely Chung. 2015. Application of word embeddings in biomedical named entity recognition tasks. Journal of Digital Information Management, 13(5).

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.

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.

Edson Florez, Frédéric Precioso, Michel Riveill, and Romaric Pighetti. 2018. Named entity recognition using neural networks for clinical notes. In International Workshop on Medication and Adverse Drug Event Detection, pages 7–15.

Anupama Gupta, Imon Banerjee, and Daniel L Rubin. 2018. Automatic information extraction from unstructured mammography reports using distributed semantics. Journal of biomedical informatics, 78:78–86.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. Bioinformatics, 33(14):i37–i48.

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, 3:160035.

Liadh Kelly, Lorraine Goeuriot, Hanna Suominen, Tobias Schreck, Gondy Leroy, Daniille L Mowery, Sumithra Velupillai, Wendy W Chapman, David Martinez, Guido Zuccon, et al. 2014. Overview of the share/clef ehealth evaluation lab 2014. In International Conference of the Cross-Language Evaluation Forum for European Languages, pages 172–191.

Sriniwas Rao Kundeti, J Vijayanananda, Srikanth Mujiga, and M Kalyan. 2016. Clinical named entity recognition: Challenges and opportunities. In Big Data (Big Data), 2016 IEEE International Conference on, pages 1937–1945.

Guillaume Lample, Miguel Ballesteros, Sandeep Subramanian, Kazuya Kawakami, and Chris Dyer. 2016. Neural architectures for named entity recognition. arXiv preprint arXiv:1603.01360.

Heeyoung Lee, Yves Peirsman, Angel Chang, Nathanael Chambers, Mihai Surdeanu, and Dan Jurafsky. 2011. Stanford’s multi-pass sieve coreference resolution system at the conll-2011 shared task. In Proceedings of the fifteenth conference on computational natural language learning: Shared task, pages 28–34.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2019. Biobert: pre-trained biomedical language representation model for biomedical text mining. arXiv preprint arXiv:1901.08746.

Zengjian Liu, Ming Yang, Xiaolong Wang, Qingcai Chen, Buzhou Tang, Zhe Wang, and Hua Xu. 2017. Entity recognition from clinical texts via recurrent neural network. BMC medical informatics and decision making, 17(2):67.

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

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), pages 1532–1543.
A Supplemental Material

Figure A.1: Training loss of language model over iteration steps

(a) ELMo (MIMIC)

(b) BERT\textsubscript{BASE} (MIMIC)
| No. | Web corpus                                                                 | Clinical corpus                                                                                                                                 |
|-----|---------------------------------------------------------------------------|-----------------------------------------------------------------------------------------------------------------------------------------------|
| 1   | Cold is the presence of low temperature, especially in the atmosphere.    | Cough that gets worse while other cold symptoms improve.  
|     |                                                                           | This patient’s sudden onset sore throat, tender cervical lymphadenopathy, fever, and lack of cough or other symptoms of cold.         |
| 2   | Its getting cold and colder because of the six horsemen of climate change | Denies any specific symptoms of illness - no cough or cold symptoms, no abd pain, diarrhea, vomiting.                                     |
| 3   | What could be better on a cold winter’s day than coming home to a fully cooked meal? |                                                                                                                                             |
| 4   | This time of year, when people are traveling and bouncing between hot and cold weather, | She also noted that she felt cold with shaking chills at the onset of the headache.                                                           |
| 5   | I don’t know how they get through these cold winters.                     | She reports that he stated that he was cold over the last few days with a mild cough.                                                        |
| 6   | Washing in cold water might actually be just as effective and will save you some money. | Guaifenesin 100mg/5ml Liquid Sig: 01-28 teaspoons P0 every six hours as needed for cold symptoms.                                        |
| 7   | The cold weather exhilarated the walkers.                                | he gets cold easily but denies fevers.                                                                                                        |
| 8   | Try to choose plants that live best with cold weather, and planting areas that face west. | He does complain of cold symptoms times one week.                                                                                           |
| 9   | In cold weather, many of these, volunteer firefighters wore coats made of the skin of buffalo to keep them warm and dry. | he said he had cold symptoms. vomited several times. the throat may be affected                                                              |
| 10  | The unusually long period of cold weather has shown how even warm climates can sometimes freeze over. | the patient stated that he had cold for 2 days prior to evaluation also nasal drainage, coughing.                                          |

Table A.1: Sentences for PCA clustering