Clinical Concept Linking with Contextualized Neural Representations

Elliot Schumacher[1] **Andriy Mulyar**[2]* **Mark Dredze**[1]
[1]Johns Hopkins University [2]Virginia Commonwealth University
{eschumac,mdredze}@cs.jhu.edu aymulyar@vcu.edu

Abstract

In traditional approaches to entity linking, linking decisions are based on three sources of information – the similarity of the mention string to an entity's name, the similarity of the context of the document to the entity, and broader information about the knowledge base (KB). In some domains, there is little contextual information present in the KB and thus we rely more heavily on mention string similarity. We consider one example of this, concept linking, which seeks to link mentions of medical concepts to a medical concept ontology. We propose an approach to concept linking that leverages recent work in contextualized neural models, such as ELMo (Peters et al., 2018), which create a token representation that integrates the surrounding context of the mention and concept name. We find a neural ranking approach paired with contextualized embeddings provides gains over a competitive baseline (Leaman et al., 2013). Additionally, we find that a pre-training step using synonyms from the ontology offers a useful initialization for the ranker.

1 Introduction

Medical concept linking produces structured topical content from clinical free text (Aronson and Lang, 2010). Healthcare providers often refer to medical concepts in clinical text notes that are absent from associated health record metadata despite their importance to understanding a patient's medical status. For example, in *The patient reports a history of **seizure disorder**...*, the phrase **seizure disorder** refers to the concept *epilepsy* contained within the Unified Medical Language System (UMLS) ontology (Bodenreider, 2004). However, this may be absent from metadata as it is not part of the current diagnosis. Concept mentions can use non-standard terms (e.g. *epilepsy*), thus concept linking requires non-lexical methods. Additionally, some terms (**cancer**) are ambiguous and could refer to multiple concepts (*breast cancer*, *colon cancer*, etc.)

The related task of Entity Linking – linking named entities (people, places, and organizations) to a knowledge base – has been explored in non-medical domains (Dredze et al., 2010; Durrett and Klein, 2014; Gupta et al., 2017). Entity linking systems consider three sources of information: 1) similarity between mention strings and names for the KB entity; 2) comparison of the document context to information about the KB entity (e.g. entity description); 3) information contained in the KB, such as entity popularity or inter-entity relations.

In contrast to the dense KBs in entity linking, concept linking uses sparse ontologies, which contain a unique identifier (CUI), title, and links to synonyms and related concepts, but rarely long-form text. For example, while the concept **epilepsy** has many synonyms in UMLS, it has no definition or other long description. Furthermore, UMLS concept names are more formal than clinical notes, making mention matching challenging. Therefore, we need an approach that can use local context from the mention (surrounding sentence), and whatever information may be present in the ontology to build a contextualized non-lexical representation for matching.

Additionally, Entity Linking systems are often able to leverage greater amounts of annotated data, which are not available in the clinical space. Text that does not have restrictive privacy protections can be annotated more easily through crowdsourcing, or other sources of non-gold standard data collected (e.g., Wikipedia cross-links). As the annotation of clinical notes is expensive due to the knowledge required of annotators and the protected status of clinical records, any effort in clinical concept linking must focus on leveraging a small amount

* Contribution performed during an internship at Johns Hopkins University.

Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 8585–8592

of annotations, and using larger amounts of related or unannotated data when possible.

We propose learning contextualized representations that leverage both free text and information from knowledge bases. We train a contextualized language model (Peters et al., 2018) on unannotated clinical text, leveraging sentence context to construct a mention. We explore several methods of building representations of the mention span and concept, including pooling and attention, and pre-training our linker with additional data from the ontology to augment the small amount of annotated data present. The resulting ranker outperforms a non-contextualized version of our model, and beats the previous best performing system (Leaman et al., 2013) in most metrics.

2 Concept Linking

Concept linking (alternatively: named entity recognition, entity normalization), has a long history (Pradhan et al., 2013; Luo et al., 2019) in the clinical NLP community, with common approaches including generating lexical variations to increase matches (Metamap) (Aronson, 2001; Aronson and Lang, 2010), dictionary matching algorithms (Kipper-Schuler et al., 2008; Savova et al., 2010), rule based systems (D'Souza and Ng, 2015), and mention/ontology context overlap (Aggarwal and Barker, 2015). Learned ensembles can also be effective (Rajani et al., 2017). Concept linking has also been applied to bio-medical literature (Doğan et al., 2014; Zheng et al., 2015; Tsai and Roth, 2016; Zhao et al., 2019) and is most similar to the task of entity linking (Dredze et al., 2010; Durrett and Klein, 2014; Gupta et al., 2017; Mueller and Durrett, 2018). Similar to our approach, Choi et al. (2016) learn representations of concepts in UMLS. While we cannot make a direct comparison since they do not cover all of our KB (SNOMED-CT), initial experiments with their embeddings performed worse than our method.

While some jointly consider the task of mention finding and linking (Durrett and Klein, 2014), we follow the more common convention of separating the two and assuming gold mention spans (Leaman et al., 2013; D'Souza and Ng, 2015). Formally, we are given a mention m in a document and must select the best CUI (concept) c from an ontology/KB, or *CUI-less* if no relevant concept exists.

Many systems utilize a rule-based approach – often as a pre-processing step – that uses the training data to augment a dictionary (D'Souza and Ng, 2015; Luo et al., 2019). While this approach does quite well, it poorly generalizes to unseen mentions or new domains.[1] Therefore, our work will focus on a learned system and compare it to similar baselines.

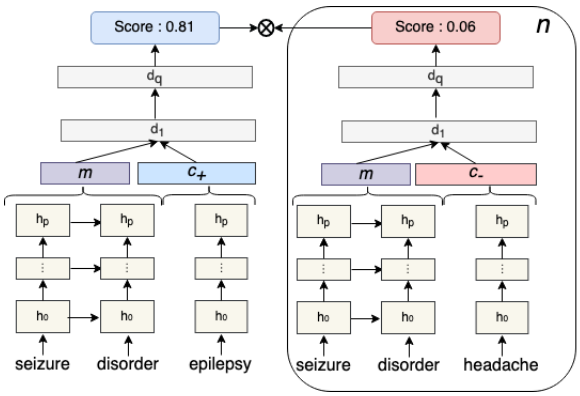

Figure 1: Architecture for our neural ranker. The input consists of gold standard mention string representation m (purple), gold standard concept representation c_+ (blue), and n randomly selected negative concept representation c_- pairings (red). The ELMO hidden states are noted as h, and the hidden states of our feed forward neural network are noted as d. To build our ELMO representations for m, c_+ and c_-, we select the representation from the lowest layer of the model.

While related to concept linking, entity linking requires a different solution due to several factors. Many entity linking systems (Upadhyay et al., 2018; Kolitsas et al., 2018) leverage context from a large document, such as Wikipedia, to make linking decisions, while a similar source is not present in UMLS. Further, earlier work (Zheng et al., 2014) showed that standard Entity Linking systems don't work well on the related domain of biomedical journal literature, which suggests that separate solutions are required.

3 Methods

Our concept linking system is based on a pairwise neural network ranker (§3.1) using contextualized representations (§3.2) for both the mention and concept. We leverage the context present in clinical notes for our representations and synonyms present within the UMLS to train our linker.

3.1 Neural Ranker

For a given mention string m and document, the system ranks all possible candidates c in the KB. Figure 1 shows our ranking system, based on the *Rank model* of Dehghani et al. (2017). We learn the parameters θ of a scoring function $S(m, c; \theta)$, which consists of a feed-forward neural network with hidden layers d that takes input representations of m and c in addition to pairwise features. We train using pairwise loss, in which we have two point-wise networks – one which takes the mention m and correct concept c_+ as input, the other which takes the mention m and incorrect concept c_- – with shared parameters that are updated to minimize the loss function. Using a pairwise model allows us to learn a scoring function that does not rely on annotated scores.

Adapting the approach of Dehghani et al. (2017), we use adaptive hinge loss, which considers n negative concepts and selects the highest scoring concept as the negative sample. For mention m, correct concept c_+, and n negative samples c_{0-} to c_{n-}, our loss function is:

$$L(\theta) = \mathbf{max}\{0, \epsilon - (S(\{m, c_+\}; \theta) - \mathbf{max}\{S(\{m, c_{0-}\}; \theta) \dots S(\{m, c_{n-}\}; \theta)\}\} \quad (1)$$

3.2 Contextualized Representations

Recent work (Devlin et al., 2019) proposed representations of words that integrate the context of the surrounding sentence. We use ELMo (Peters et al., 2018), a bi-directional recurrent neural network (RNN), to build representations for each token in a sentence trained using language model objectives. For each direction, the model first builds a context-independent token representation using a convolutional neural network over the characters. Then the representation is passed through $L = 2$ layers of long-short term memory (LSTM) RNN. The final layer is used to predict the next token. These models are robust to out-of-vocabulary types, so they provide broad coverage to the diverse types present in clinical text. We train ELMo on clinical notes and create mention representations m by running the entire sentence through the model and selecting the resulting word representations for the mention (the lowest token representation) from the LSTM.[2].

The concept representations c are created in the same manner as m except that only the name of the concept, as there is often no available context[3].

For multi-word mentions and concept names, we explore two methods of creating a single embedding. First, we use max-pooling over the set of token embeddings (reported as **Max** in Table 1). Second, we run self-attention (Vaswani et al., 2017)[4] over the set of token embeddings, with a single head to attend over the tokens (noted as **Attention**).

3.3 Pre-training with Structured Data

Pre-training a model using an alternative data source has been frequently used in the field of machine learning (Erhan et al., 2010; Sharif Razavian et al., 2014), and presented (Tsujimura et al., 2019) at a recent shared task (Luo et al., 2019). A model is pre-trained on a large amount of a related dataset and then is trained on the target task, which allows a model to see more examples to achieve a better initialization for training on the final task.

As creation is expensive, most annotated clinical datasets are small, such as for our task. Therefore, we look to alternative data sources for pre-training our model. For a given concept (e.g. **epilepsy**), the UMLS includes synonyms (e.g. **seizure disorder**, **epileptic fits**), which can be used to pre-train our linker. Unlike in the annotated clinical data, there is no surrounding context, and terms in the UMLS are more likely to be formal. However, training on synonyms will allow for a greater variety of terms to be seen by our model than otherwise possible.

Therefore, using all synonyms taken from the annotated subset of the UMLS, we pre-train our linker before training on the annotated clinical notes. We follow the previous training procedure by replacing the mention representation m with the synonym string representation only (without surrounding sentence), thus training the linker to assign a higher score to the synonym paired with the corresponding concept representation c_+ against negatively sampled concepts c_-. We use this pre-training initialization with the Attention model discussed in

[1] An extension of this approach could use unsupervised methods to discover synonyms in a new dataset (Schumacher and Dredze, 2019)

[2] While there are now a multitude of deep transformer-based LMs (Devlin et al., 2019), the principle of contextualized representations are the same. Additionally, others have found ELMo trained on MIMIC does better than a similarly trained BERT model (Schumacher and Dredze, 2019)

[3] We ran experiments that padded the names with synonyms or other forms of available text within the knowledge base. However, we did not see consistent improvements.

[4] We use the implementation provided by https://github.com/kaushalshetty/Structured-Self-Attention.

	CUI		All	
	Acc	MRR	Acc	MRR
DNorm	**0.73**	0.75	0.55	0.57
Word2vec	0.26	0.33	0.21	0.30
Max	0.66	0.70	0.58	0.67
Attention	0.70	0.75	**0.62**	**0.71**
Att. + Pre.	0.70	**0.78**	0.59	**0.71**

Table 1: Accuracy (top-1) and MRR (mean reciprocal rank) for the test sets, for mentions with linked concepts (CUI) and all mentions (All). For each metric, we compare the best score (in bold) to the baseline using a two-tailed z-score test (for CUI ACC, we compare to the next best score). We find that for all CUI models, the difference is not significant, while for All models, $p < 0.05$.

the previous section and note this as **Att. + Pre.** in Table 1.

4 Experimental Setup

We train and evaluate our system on the ShARe/CLEF eHealth Evaluation Lab 2013 Task 1b dataset (Pradhan et al., 2013), which consists of span-level annotations for disorder concepts taken from the MIMIC 2.5 clinical note dataset (Saeed et al., 2011). The publicly available training set includes 200 clinical notes, which we split into a 100 note training set, and development and testing sets of 50 documents each - the shared task test set was not available. The data is annotated against SNOMED-CT (Spackman et al., 1997), one of the ontologies within UMLS. We choose to focus on this smaller dataset as leveraging small amounts of annotated data is critical to building useful tools in the clinical domain.

We only included mention annotations for concepts that occur in the selected subset of the ontology noted in the annotation guidelines for the respective datasets or are marked as *CUI-less* [5]. In Table 1, we report results on only mentions with links to the ontology (**CUI**) and mentions with links to the ontology and *CUI-less* mentions (**All**). We train ELMo on 199,987 clinical notes from MIMIC III (Johnson et al., 2016) as the source of our clinical text, pre-processing the data using the NLTK toolkit (Řehůřek and Sojka, 2010). For the Pre-training model, we augment the clinical text training data with synonyms, definitions, and names of related concepts from the selected subset of UMLS. All together, this resulted in 645,863 additional sentences of training data.

We compare our system to DNorm (Leaman et al., 2013) for the SHARE/Clef 2013 dataset, the best performing system in the SHARE/Clef 2013 shared task.[6] Unlike many other concept linking systems, DNorm scores each mention against all concepts and does not use a triage system, allowing a fair comparison to our system. DNorm builds term frequency-inverse document frequency (TF-IDF) representations of both the mention and concept and learns a weighted similarity to rank concepts for each mention. It is unable to return concept candidates for mentions that are out-of-vocabulary as it uses a word-level measure. The authors add a specific `CUI-less` representation, which is made of entries occurring more than four times in training. We report results on our recreated test set, as the evaluation set provided for the shared task was not available to us. We also compare with using Word2vec (Mikolov et al., 2013) representations instead of ELMo representations in the same linking architecture to test the effect of contextualized embeddings. We trained the Word2vec model on the MIMIC dataset. We created single embeddings ($d = 600$) for mentions and concepts by max pooling over all embeddings for words in the corresponding text, ignoring all out-of-vocabulary words.

We explored several parameter configurations for our model suggested in Dehghani et al. (2017), reporting the best performing models on development. These include hidden layers of size [256, 512, 1024] and number of layers in [1,2,3], with a Tanh activation function for final layer and ReLu (Glorot et al., 2011) for all others. We optimize using the ADAM optimizer (Kingma and Ba, 2014), and a dropout rate of 0.2. Parameter values and development metrics are available in Appendix A. For the ELMo models, we trained for 10 epochs

[5] We included all concepts in the SNOMED-CT Disorder Semantic group or in the *Finding*, *Body Substance*, and *Mental Process* semantic types. We include all preferred entries, with the default settings of UMLS 2011AA, in the SNOMED-CT Disorder Semantic group (116,436 unique concepts), but also include the first non-preferred entries that do not have a preferred entry (8,926 unique concepts.), and annotations marked *CUI-less*. Mentions that do not have a corresponding concept in the ontology (e.g. *calcifications*) were classified as CUI-less (or `NIL`) entries by annotators. Some annotations consist of concepts outside of the subsets described in the shared task paper, and we exclude those exceptions.

[6] As of this writing, there are no papers describing the 2019 N2C2 methods. Additionally, since we are interested in non-training data-based dictionaries, a direct comparison to shared task submissions wasn't possible.

using the default configuration. For `CUI-less` mentions, we select a threshold score based on the development set, equal to the mean score of all `CUI-less` entries. If an entry does not have a scored concept above that threshold, we consider it `CUI-less`, adding `CUI-less` at that position in the list for MRR. We use the Pytorch framework and code from the Spotlight library (Kula, 2017).

5 Results

Table 1 reports accuracy and mean reciprocal rank (MRR) for all models. We compare our models (**Word2Vec**, **Max**, **Attention**, and **Att. + Pre.**) to DNorm for all mentions (All) and only those with links to concepts in the KB (CUI). While DNorm has higher accuracy on entries with CUIs, our models have higher MRR on entities with CUIs (**Att. + Pre.**) and perform best on all entities in both accuracy and MRR (**Attention** and **Att. + Pre.**).

6 Discussion

Our neural ranking models with attention outperform all other models, except for CUI-only accuracy. In the case of entities with CUIs, we find that pre-training the model does provide a gain in ranking accuracy (MRR). In the case of all entities, we find that the attention models provide a sizable gain in both accuracy and MRR.

We conducted an error analysis of the best performing MRR model (**Att. + Pre.**) on the development data, looking at errors where the gold standard concept was not highly ranked (assigned a rank of 10 or above). Of those errors ($n = 110$), we find that 26% are mentions that contain only acronyms (e.g. *LBP* for *lower back pain*), and 14% are mentions containing some other abbreviation (a shorted word, e.g. *post nasal drip* for *Posterior rhinorrhoea*, or a partial acronym, *Seizure d / o* for *Epilepsy*). Comparing to similar errors from **Attention** model ($n = 161$), we find that the number of acronym errors is nearly the same (24) as the better performing model (26). In contrast, the number of non-abbreviation errors drops significantly.

This suggests that pre-training provides useful signal for mentions that consist of variations appearing in the ontology. However, it does not help with acronyms or other abbreviations that are less likely to appear in the ontology or are shorter and more ambiguous (e.g., 'R' for Rhonchus).

While the linker often predicted unrelated concepts (40% of errors) for concepts where the correct concept was ranked above 10, many incorrect concept predictions were somewhat related to the gold concept (e.g., for mention *atherosclerotic plaque* with gold concept *Atherosclerotic fibrous plaque* our model predicted the concept *Atherosclerosis*). We further noticed that in 21% of cases the linker predicted a relevant concept (e.g., mention *thrombosed* and *Thrombosis*), but is not counted as correct due to annotation decisions. This could be due to multiple possible concepts in the ontology or the presence of closely-related concepts.

Deploying our system in a large-volume clinical setting would likely require several alterations. The main computational barrier to labeling a large amount of data, the speed of prediction, can be addressed by using an accurate candidate selection system to prune the number of concepts considered. Considering a smaller subset (e.g., 20) of concepts instead of all would significantly improve the speed. Further, if using a consistent portion of the ontology, caching the concept embeddings c as opposed to building them in-model also enhances efficiency. Depending on the application, a less accurate but faster linker might be a better choice (e.g. for all clinical notes at a medical institution). In contrast, a more complex linker, such as ours, maybe a better option for specific subsets of notes that require better accuracy (e.g., the results of specific clinical studies).

Our results demonstrate the advantages of using contextualized embeddings for ranking tasks, and that using information from the knowledge base for training is an essential direction for learning concept representations for sparse KB domains. Future work will consider additional methods for integrating ontology structure into representation learning.

References

Nitish Aggarwal and Ken Barker. 2015. Medical concept resolution. In *International Semantic Web Conference (Posters & Demos)*.

Alan R Aronson. 2001. Effective mapping of biomedical text to the umls metathesaurus: the metamap program. In *Proceedings of the AMIA Symposium*, page 17. American Medical Informatics Association.

Alan R Aronson and François-Michel Lang. 2010. An overview of metamap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229–236.

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

Youngduck Choi, Chill Yi-I Chiu, and David Sontag. 2016. Learning low-dimensional representations of medical concepts. *AMIA Summits on Translational Science Proceedings*, 2016:41.

Mostafa Dehghani, Hamed Zamani, Aliaksei Severyn, Jaap Kamps, and W Bruce Croft. 2017. Neural ranking models with weak supervision. In *Proceedings of the 40th International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 65–74. ACM.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Rezarta Islamaj Doğan, Robert Leaman, and Zhiyong Lu. 2014. Ncbi disease corpus: a resource for disease name recognition and concept normalization. *Journal of biomedical informatics*, 47:1–10.

Mark Dredze, Paul McNamee, Delip Rao, Adam Gerber, and Tim Finin. 2010. Entity disambiguation for knowledge base population. In *Conference on Computational Linguistics (COLING)*, pages 277–285. Association for Computational Linguistics.

Jennifer D'Souza and Vincent Ng. 2015. Sieve-based entity linking for the biomedical domain. In *Association for Computational Linguistics (ACL)*, pages 297–302.

Greg Durrett and Dan Klein. 2014. A joint model for entity analysis: Coreference, typing, and linking. *Transactions of the Association for Computational Linguistics*, 2:477–490.

Dumitru Erhan, Yoshua Bengio, Aaron Courville, Pierre-Antoine Manzagol, Pascal Vincent, and Samy Bengio. 2010. Why does unsupervised pre-training help deep learning? *Journal of Machine Learning Research*, 11(Feb):625–660.

Xavier Glorot, Antoine Bordes, and Yoshua Bengio. 2011. Deep sparse rectifier neural networks. In *Proceedings of the fourteenth international conference on artificial intelligence and statistics*, pages 315–323.

Nitish Gupta, Sameer Singh, and Dan Roth. 2017. Entity linking via joint encoding of types, descriptions, and context. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2681–2690, Copenhagen, Denmark. Association for Computational Linguistics.

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.

Diederik Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *International Conference on Learning Representations*.

Karin Kipper-Schuler, Vinod Kaggal, James Masanz, Philip Ogren, and Guergana Savova. 2008. System evaluation on a named entity corpus from clinical notes. In *Language resources and evaluation conference, LREC 2008*.

Nikolaos Kolitsas, Octavian-Eugen Ganea, and Thomas Hofmann. 2018. End-to-end neural entity linking. In *Proceedings of the 22nd Conference on Computational Natural Language Learning*, pages 519–529, Brussels, Belgium. Association for Computational Linguistics.

Maciej Kula. 2017. Spotlight. `https://github.com/maciejkula/spotlight`.

Robert Leaman, Rezarta Islamaj Doğan, and Zhiyong Lu. 2013. Dnorm: disease name normalization with pairwise learning to rank. *Bioinformatics*, 29(22):2909–2917.

Yen-Fu Luo, Weiyi Sun, and Anna Rumshisky. 2019. Mcn: A comprehensive corpus for medical concept normalization. *Journal of biomedical informatics*, page 103132.

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.

David Mueller and Greg Durrett. 2018. Effective use of context in noisy entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1024–1029, Brussels, Belgium. Association for Computational Linguistics.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, Louisiana. Association for Computational Linguistics.

Sameer Pradhan, Noemie Elhadad, Brett R South, David Martinez, Lee M Christensen, Amy Vogel, Hanna Suominen, Wendy W Chapman, and Guergana K Savova. 2013. Task 1: Share/clef ehealth evaluation lab 2013. In *CLEF (Working Notes)*.

Nazneen Fatema Rajani, Mihaela Bornea, and Ken Barker. 2017. Stacking with auxiliary features for entity linking in the medical domain. *BioNLP 2017*, pages 39–47.

Radim Řehůřek and Petr Sojka. 2010. Software Framework for Topic Modelling with Large Corpora. In *Proceedings of the LREC 2010 Workshop on New Challenges for NLP Frameworks*, pages 45–50, Valletta, Malta. ELRA. http://is.muni.cz/publication/884893/en.

Mohammed Saeed, Mauricio Villarroel, Andrew T Reisner, Gari Clifford, Li-Wei Lehman, George Moody, Thomas Heldt, Tin H Kyaw, Benjamin Moody, and Roger G Mark. 2011. Multiparameter intelligent monitoring in intensive care ii (mimic-ii): a public-access intensive care unit database. *Critical care medicine*, 39(5):952.

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.

Elliot Schumacher and Mark Dredze. 2019. Learning unsupervised contextual representations for medical synonym discovery. *JAMIA Open*. Ooz057.

Ali Sharif Razavian, Hossein Azizpour, Josephine Sullivan, and Stefan Carlsson. 2014. Cnn features off-the-shelf: an astounding baseline for recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition workshops*, pages 806–813.

Kent A Spackman, Keith E Campbell, and Roger A Côté. 1997. Snomed rt: a reference terminology for health care. In *Proceedings of the AMIA annual fall symposium*, page 640. American Medical Informatics Association.

Chen-Tse Tsai and Dan Roth. 2016. Concept grounding to multiple knowledge bases via indirect supervision. In *Transactions of the Association of Computational Linguistics*, volume 4, pages 141–154.

Tomoki Tsujimura, Noriyuki Mori, Masaki Asada, Makoto Miwa, and Yutaka Sasaki. 2019. Neural medical concept normalization with two-step training.

Shyam Upadhyay, Nitish Gupta, and Dan Roth. 2018. Joint multilingual supervision for cross-lingual entity linking. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2486–2495.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Sendong Zhao, Ting Liu, Sicheng Zhao, and Fei Wang. 2019. A neural multi-task learning framework to jointly model medical named entity recognition and normalization. In *Proceedings of the AAAI Conference on Artificial Intelligence*, volume 33, pages 817–824.

Jin G Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2015. Entity linking for biomedical literature. *BMC medical informatics and decision making*, 15:S4.

Jin Guang Zheng, Daniel Howsmon, Boliang Zhang, Juergen Hahn, Deborah McGuinness, James Hendler, and Heng Ji. 2014. Entity linking for biomedical literature. In *Proceedings of the ACM 8th International Workshop on Data and Text Mining in Bioinformatics*, pages 3–4.

A Replication Information

	Max	Attention	Pretraining	Pre + Att
Dev Acc (CUI)	0.685	0.730	-	0.704
Dev MRR (CUI)	0.719	0.766	-	0.776
Reported Epoch	2499	4000	1	750
Random Seed	3011457727	3027767026	589590319	3635932273
Learning Rate	1e-5	1e-5	1e-5	1e-5
Hidden Layers	[1024, 512]	[1024, 512]	[1024, 512]	[1024, 512]
Batch Size	12	12	32	16
Num. Negative Samples	10	10	10	10
Est. Training Time per epoch (minutes)	7.2	3.4	1860	4.6
GPU Type	Tesla K80	GTX 1080ti	Tesla K80	Tesla K80

Table 2: The above table contains replication information for the models trained on SHaRE data. Note the pre-training model contains parameters for the pre-training stage only (and thus we do not note accuracy or mean reciprocal rank), while Pre + Att contains parameters for the final trained model. All GPU types have 12 GB of memory.