

Deep Contextualized Biomedical Abbreviation Expansion

Qiao Jin

University of Pittsburgh
qiao.jin@pitt.edu

Jinling Liu

University of Pittsburgh
jill172@pitt.edu

Xinghua Lu

University of Pittsburgh
xinghua@pitt.edu

Abstract

Automatic identification and expansion of ambiguous abbreviations are essential for biomedical natural language processing applications, such as information retrieval and question answering systems. In this paper, we present DEep Contextualized Biomedical Abbreviation Expansion (**DECBAE**) model. DECBAE automatically collects substantial and relatively clean annotated contexts for 950 ambiguous abbreviations from PubMed abstracts using a simple heuristic. Then it utilizes BioELMo (Jin et al., 2019) to extract the contextualized features of words, and feed those features to abbreviation-specific bidirectional LSTMs, where the hidden states of the ambiguous abbreviations are used to assign the exact definitions. Our DECBAE model outperforms other baselines by large margins, achieving average accuracy of 0.961 and macro-F1 of 0.917 on the dataset. It also surpasses human performance for expanding a sample abbreviation, and remains robust in imbalanced, low-resources and clinical settings.

1 Introduction

Abbreviations are shortened forms of text-strings. They are prevalent in biomedical literature such as scientific articles, clinical notes and user queries in information retrieval systems. Abbreviations can be ambiguous (e.g.: ER can refer to estrogen receptor, endoplasmic reticulum, emergency room etc.), especially when they appear in short or professional texts where the definitions are not given. For instance, about 15% of PubMed queries include abbreviations (Islamaj Dogan et al., 2009), and about 14.8% of all tokens in a clinical note dataset are abbreviations (Xu et al., 2007). In both cases, the definitions of the abbreviations are rarely provided. Thus, automatic expansion of ambiguous abbreviations to their full forms is vital

in biomedical natural language processing (NLP) systems.

In this paper, we focus on the cases where definitions of ambiguous abbreviations are not directly available in the contexts, so reasoning over the contexts is required for disambiguation. Under the conditions where definitions are provided in the contexts, one can easily extract them using rule-based methods.

We present DEep Contextualized Biomedical Abbreviation Expansion (DECBAE) model. DECBAE uses a simple heuristic to automatically construct large supervised disambiguation datasets for 950 abbreviations from PubMed abstracts: In scientific writing, authors define abbreviations the first time they are used, and the same abbreviations in the following sentences have the same definitions as those of the first ones. We extract all the sentences containing the same abbreviations in each PubMed abstract, and use the definition given in the first sentence as the full form label of abbreviations in the following sentences. We group the definitions for each abbreviation and formulate abbreviation expansion as a classification task, where input is an ambiguous abbreviation with its context, and the output is one of its possible definitions.

Recent breakthroughs of language models (LM) pre-trained on large corpora like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) clearly show that unsupervised LM pre-training can vastly improve performance of downstream models. To fully utilize the knowledge encoded in PubMed abstracts, DECBAE uses BioELMo (Jin et al., 2019), a domain adaptation version of ELMo, to embed the words. After the embedding layer, DECBAE applies abbreviation-specific bidirectional LSTM (biLSTM) classifiers to do the abbreviation expansion, where the biLSTM parameters are trained separately for each abbrevi-

ation. We train DECBAE from the automatically collected dataset of 950 ambiguous abbreviations.

At inference time, DECBAE feeds the BioELMo embeddings of the whole sentence and uses the corresponding abbreviation-specific biLSTM classifiers to perform disambiguation of abbreviations in the sentence. We show that DECBAE outperforms other baselines by large margins and even performs better than single human expert. Although training instances of DECBAE are collected from PubMed, it covers 85% of clinically related abbreviations mentioned in a previous work (Xu et al., 2012). Moreover, DECBAE remains robust in low-resource and imbalanced settings.

2 Related Work

Contextualized word embeddings: Recently, contextualized word representations pre-trained by large corpora like ELMo (Peters et al., 2018) and BERT (Devlin et al., 2018) significantly improve the performance of various NLP tasks. ELMo is a pre-trained biLSTM language model. ELMo word embeddings are calculated by a weighted sum of the hidden states of each biLSTM layer. The weights are task-specific learnable parameters while biLSTM layers are fixed. In-domain trained contextual embeddings further improve the performance on domain-specific tasks. In this paper, we use BioELMo, which is a biomedical version of ELMo trained on 10M PubMed abstracts (Jin et al., 2019). BioELMo outperforms general ELMo by large margins on several biomedical NLP tasks.

We don't use BERT for contextualized embeddings due to its fine-tuning nature: users just need to download 1 BioELMo and N abbreviation-specific biLSTM weights to run DECBAE locally, which takes significantly less disk size than N fine-tuned BERTs for each abbreviation. N is the number of abbreviations.

Word sense disambiguation (WSD): The goal of WSD is to determine the correct sense of words in different contexts. Abbreviation expansion is a specific case of WSD where the ambiguous words are abbreviations. In this paper, we use abbreviation expansion and abbreviation disambiguation interchangeably. Several human-annotated datasets are available for supervised WSD (Navigli et al., 2013; Camacho-Collados et al., 2016; Raganato et al., 2017b). However, human anno-

tations could be expensive, especially in domain specific settings. To address this problem, some automatic dataset collection methods have been proposed (Yu et al., 2007; Ciosici et al., 2019), where abbreviations are automatically labeled if they are defined previously in the same documents. We use a similar approach in this work.

Peters et al. (2018) report that just matching the ELMo embedding of the target words with the nearest sense representations, calculated by averaging their ELMo embeddings, leads to comparable WSD performance with state-of-the-art models using hand crafted features (Iacobacci et al., 2016) or task-specific biLSTM trained with multiple tasks (Raganato et al., 2017a). Instead of searching the nearest contextualized embeddings neighbors of the abbreviation and definitions, we model abbreviation expansion as classification.

Biomedical abbreviation expansion: Various methods have been introduced for automatically expanding biomedical abbreviations. Yu et al. (2007) train naive Bayes and SVM classifiers with bag-of-words features on an automatically collected dataset from PubMed. Some works disambiguate abbreviations to their senses in controlled vocabularies like Medical Subject Headings¹ (MeSH) and Unified Medical Language System² (UMLS). Xu et al. (2015) use pooled neighbor word embeddings of the abbreviations as features to train SVM classifiers for clinical abbreviation disambiguation. Jimeno-Yepes et al. (2011) introduced MSH WSD dataset to test the performance of supervised biomedical WSD systems and several supervised models have been proposed on it (Antunes and Matos; Yepes, 2017). Recently Pesaranghader et al. (2019) presented deep-BioWSD which sets new state-of-the-art performance on it. DeepBioWSD uses a single biLSTM encoder for disambiguation of all abbreviations by calculating the pairwise similarity between context representations and sense representations.

To the best of our knowledge, DECBAE is the first model that uses deep contextualized word embeddings for biomedical abbreviation expansion.

3 Methods

Figure 1 shows the architecture of DECBAE. During training, we first construct abbreviation ex-

¹<https://www.nlm.nih.gov/mesh>

²<https://www.nlm.nih.gov/research/umls/>

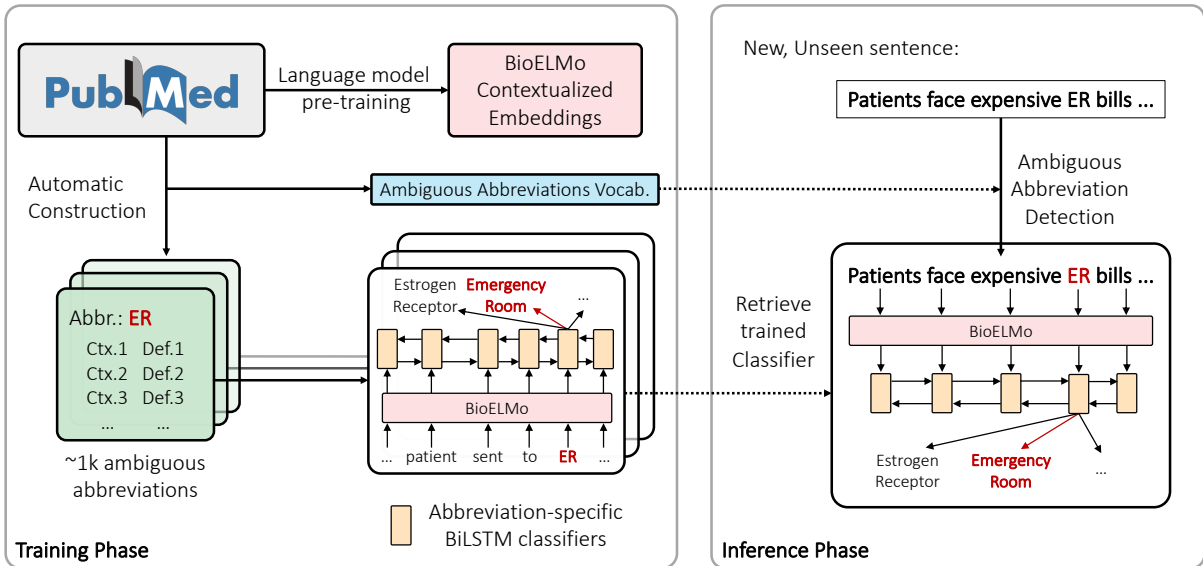


Figure 1: Architecture of DECBAE. Training and inference phases are illustrated in the left and right boxes, respectively. The PubMed corpus is used in training BioELMo (Jin et al., 2019) and collecting the disambiguation dataset. We train a separate biLSTM classifier for each abbreviation, and the specific pre-trained classifier is retrieved in inference phase.

pansion datasets from PubMed (§3.1). We use BioELMo (§3.2) to get the contextualized representations of words, and train a specific biLSTM classifier (§3.3) for each abbreviation. During inference (§3.5), we first detect whether there are ambiguous abbreviations in input sentences by the expert-curated ambiguous abbreviation vocabulary. If so, we use BioELMo and the corresponding abbreviation-specific biLSTM classifiers to do the disambiguation.

3.1 Dataset Collection

Figure 2 shows our approach of automatically collecting disambiguation dataset. For each abstract, we first detect and extract the pattern of “*Definition (Abbreviation)*”, e.g.: “endoplasmic reticulum (ER)”. Then we collect all the following sentences that contain the abbreviation, and label them with the definition.

This would generate a noisy label set due to the variations of writing the same definition (e.g.: emergency department and emergency departments). To group the same definitions together, we use MetaMap-derived MeSH terms (Demner-Fushman et al., 2017) as features of definitions and define the MeSH similarity between definition a and definition b as:

$$s = \frac{|\mathcal{M}_a \cap \mathcal{M}_b|}{\sqrt{|\mathcal{M}_a| |\mathcal{M}_b|}}$$

where \mathcal{M}_a and \mathcal{M}_b are the MeSH term sets of definition a and b , respectively. We group those definitions with high MeSH similarity and close edit distance by heuristic thresholds.

We collected 1970 abbreviations. However, due to the unsupervised nature of the collection process, some abbreviations are invalid or not ambiguous. For this, one biomedical expert³ filtered the abbreviations we found, based on 1) **Validity**: abbreviations should be biomedically meaningful; 2) **Ambiguity**: abbreviations should have multiple possible definitions, and prevalence of the dominant one should be $< 99\%$. After the filtering, there are 950 valid ambiguous abbreviations. Their statistics are shown in Table 1. We split the instances of each abbreviation into training, development and test sets: If there is more than 10k instances, we randomly select 1k for both development and test sets. Otherwise, we randomly select 10% of all instances for both development and test sets.

3.2 BioELMo

BioELMo is a biomedical version of ELMo pre-trained on 10 millions of PubMed abstracts (Jin et al., 2019). It serves as a contextualized feature extractor in DECBAE: given an input sentence of

³A post-doctoral fellow with a Ph.D. degree in biology.

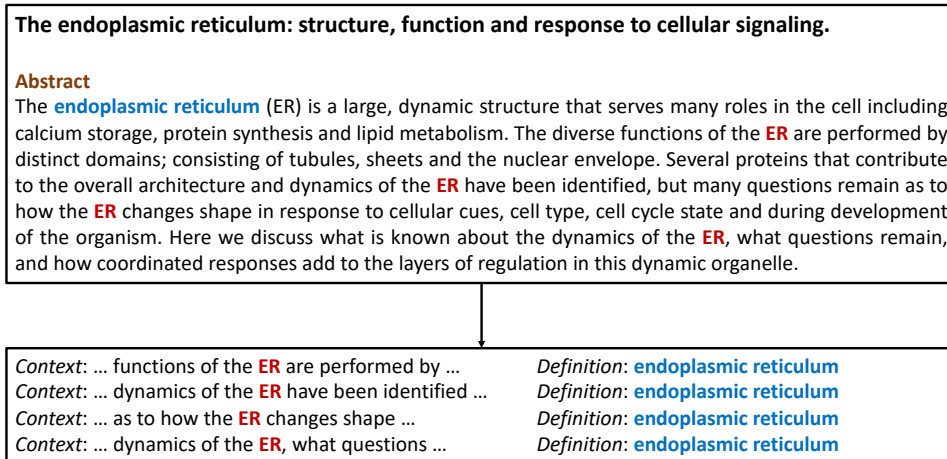


Figure 2: An example of automatically generated training instances for disambiguation from the abstract of Schwarz and Blower (2016). In this case, we extract “endoplasmic reticulum” as the definition for all ER mentions in the abstract, and store those instances to the dataset.

Statistic	Whole	Random	Imbalanced	Low-resources	Clinical	Human
# of all abbreviations	950	100	42	28	11	1
Average # of instances	8790.0	6564.3	19493.1	958.8	28642.8	8312.0
Average # of possible definitions	4.1	3.7	2.3	2.2	8.5	4.0
Average % of dominant definition	64.1	63.5	96.7	66.7	53.3	63.8

Table 1: Statistics of the automatically generated abbreviation disambiguation dataset and its subsets.

L tokens:

$$\text{input} = [t_1; t_2; \dots; t_L]$$

We use BioELMo to embed it to

$$\mathbf{E} = [\mathbf{e}_1; \mathbf{e}_2; \dots; \mathbf{e}_L] \in \mathbb{R}^{L \times D}$$

where $\mathbf{e} \in \mathbb{R}^D$ is the token embedding and D is the embedding dimension⁴.

3.3 Abbreviation-specific biLSTM Classifiers

For each abbreviation, we train a specific biLSTM classifier, denoted as biLSTM_i for abbreviation i . We feed the BioELMo representations of sentences containing abbreviation i to biLSTM_i :

$$\text{biLSTM}_i(\mathbf{E}) = [\mathbf{h}_1; \mathbf{h}_2; \dots; \mathbf{h}_L] \in \mathbb{R}^{L \times 2H}$$

where $\mathbf{h} \in \mathbb{R}^{2H}$ is the concatenation of forward and backward hidden states of the biLSTM. We take as input the concatenated hidden states of the abbreviation i (i.e. the ambiguous token) \mathbf{h}_a and use several feed-forward neural network (FFN)

⁴Note that it’s after scaling and averaging the 3 BioELMo layers using task-specific weights.

layers with softmax output unit to predict its definition:

$$p(\text{def}_k | \text{input}) \propto \exp(\mathbf{w}_k^T \text{FFN}_i(\mathbf{h}_a))$$

where \mathbf{w}_k is the learnt weight vector corresponding to definition k , and def_k is the k -th definition of abbreviation i in our dataset. Similarly, we train FFN separately for different abbreviations.

3.4 Training

The weights of BioELMo are pre-trained and fixed, while the averaging weights and scaling factor of BioELMo embeddings are trained separately for each abbreviation along with the abbreviation-specific biLSTM classifiers. We use Adam (Kingma and Ba, 2014) to optimize the cross-entropy loss of the predicted label and ground-truth label.

3.5 Inference

At inference time, we denote the tokenized input sentence as $[t_1; t_2; \dots; t_L]$ and our ambiguous abbreviation set as \mathcal{A} . If $\exists t_j \in \mathcal{A}$, we run DECBAE to expand the t_j : First, we use BioELMo to compute the representations of all the input tokens to

$\mathbf{E} = [\mathbf{e}_1; \mathbf{e}_2; \dots; \mathbf{e}_L]$. The trained biLSTM for abbreviation t_j , denoted as biLSTM_{t_j} , is retrieved and used to calculate the hidden states given the BioELMo embeddings of the input sentence:

$$\text{biLSTM}_{t_j}(\mathbf{E}) = [\mathbf{h}_1; \mathbf{h}_2; \dots; \mathbf{h}_L] \in \mathbb{R}^{L \times 2H}$$

Then \mathbf{h}_{t_j} , which is the concatenated hidden states of the ambiguous abbreviation t_j , is used for disambiguation through the trained abbreviation-specific FFN:

$$\text{Definition}(t_j) = \text{def}_{\text{argmax}_k \mathbf{w}_k^T \text{FFN}_{t_j}(\mathbf{h}_{t_j})}$$

4 Experiments

4.1 Baseline Settings

A trivial baseline is to predict the majority of definition for all cases, which could still lead to high accuracy in severely imbalanced datasets. We denote this method as **Majority**. We also test other baseline settings of different feature learning schemes. They are all followed by several FFN layers and a softmax output unit.

Bag-of-words: Following most of the previous works, we use bag-of-words features to represent the context by $\mathbf{c} \in \mathbb{R}^{|\mathcal{V}|}$, where $|\mathcal{V}|$ is the vocabulary size.

BioELMo: We take the BioELMo embeddings of the ambiguous abbreviations as input features.

biLSTM: We use biomedical w2v (Moén and Ananiadou) as word embeddings and train task-specific biLSTMs and use the hidden states of the ambiguous abbreviations as input features.

We also measure the **human performance**: due to limitation of resources, we just study single-expert performance on one sampled abbreviation. For this, the expert is shown with the test sentences, and asked to classify the ambiguous abbreviation to its possible definitions. An ensemble of experts will obviously generate better results, so our single-human results just represent the lower bound of human performance.

4.2 Subset Settings

We report the model performance on different subsets of our dataset. Statistics of those datasets are shown in Table 1.

Random samples: It’s computationally expensive⁵ and unnecessary to test the models on all 950

⁵ The rate-determining step is BioELMo due to its large size and recurrent nature.

abbreviations. Instead, we use randomly sampled 100 abbreviations to represent the whole set.

Imbalanced samples: We define abbreviations whose dominant definitions have over 95% frequency as imbalanced samples. Multi-label classification with imbalanced classes is considered as a hard machine learning task.

Low-resources samples: We define abbreviations that have less than 1k training instances as low-resources samples. It’s motivated by the fact that most biomedical datasets are typically limited by scale, so models that can still perform well under low-resources settings have the potential to be applied in real world settings.

Clinical samples: Though our abbreviations are collected from PubMed abstracts, we have included 11 out of 13 of clinical ambiguous abbreviations mentioned in a previous work of clinical abbreviation disambiguation (Xu et al., 2012). We also test our models on the subset of these 11 clinically related abbreviations.

Testing sample for human expert: We test human performance on one abbreviation (DAT), due to limited resources. The statistics of DAT abbreviation expansion dataset are close to the averages of the whole dataset, as shown in Table 1. Possible definitions of DAT include: 1) Dopamine transporter (63.9%); 2) Direct antiglobulin test (5.8%); 3) Direct agglutination test (5.8%); 4) Dementia of the Alzheimer type (24.5%).

4.3 Evaluation Metrics

We model abbreviation expansion as a multi-label classification task, and use the following metrics to measure the performance of different models:

Accuracy: Accuracy is defined as the proportion of right predictions in all predictions. Most of the definition labels are imbalanced, so accuracy could be misleadingly high for a trivial majority solution in these cases, thus may not reflect the real capability of models.

Macro-F1: In multi-label classification, macro-F1 is calculated as an unweighted average of F1 score for each class. Class-wise F1 score is defined as follows:

$$F1 = 2 \cdot \frac{\text{precision} \cdot \text{recall}}{\text{precision} + \text{recall}}$$

where precision and recall are calculated for each class.

Kappa Statistic: Cohen’s kappa was originally introduced as a metric to measure inter-rater

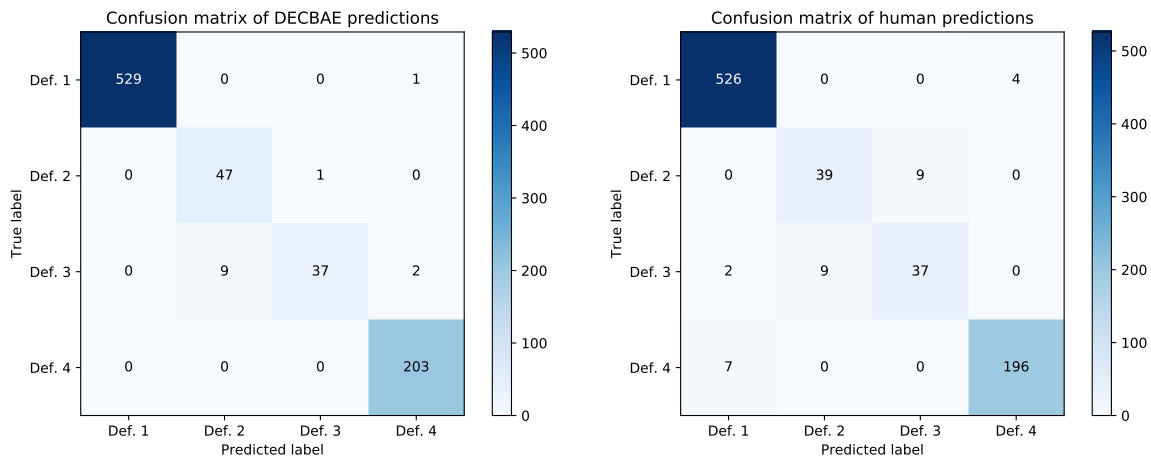


Figure 3: Confusion matrix for the predictions of DECBAE (left) and the human expert (right). Def. 1: dopamine transporter; Def. 2: direct antiglobulin test; Def. 3: direct agglutination test; Def. 4: dementia of the Alzheimer type.

Model	Random Set	Imbalanced Set	Low-resources Set	Clinical Set	Human Set
Majority					
Accuracy	63.6 ± 21.0 [†]	96.7 ± 1.0 [†]	67.0 ± 15.6 [†]	53.3 ± 25.7 [†]	63.9
Macro-F1	28.3 ± 14.9 [†]	45.4 ± 8.8 [†]	37.2 ± 8.8 [†]	12.0 ± 10.6 [†]	19.5
Kappa Statistic	0.0 ± 0.0 [†]	0.0 ± 0.0 [†]	0.0 ± 0.0 [†]	0.0 ± 0.0 [†]	0.0
BoW-FFN					
Accuracy	84.4 ± 11.2 [†]	97.5 ± 1.7 [†]	89.6 ± 7.5 [†]	76.1 ± 12.5 [†]	84.3
Macro-F1	73.1 ± 17.1 [†]	71.5 ± 19.9 [†]	83.4 ± 14.6 [†]	57.9 ± 14.2 [†]	71.9
Kappa Statistic	63.8 ± 25.3 [†]	50.4 ± 33.7 [†]	71.1 ± 24.8 [†]	60.6 ± 8.9 [†]	69.6
BioELMo					
Accuracy	94.1 ± 7.2 [†]	96.3 ± 15.3	98.1 ± 2.7	91.1 ± 8.4	97.1
Macro-F1	86.0 ± 17.4 [†]	81.3 ± 23.5 [†]	95.4 ± 9.3	75.5 ± 21.7	92.6
Kappa Statistic	86.1 ± 19.8 [†]	73.2 ± 34.2 [†]	93.2 ± 10.8 [†]	86.6 ± 9.3	94.6
biLSTM					
Accuracy	88.0 ± 16.8 [†]	98.0 ± 1.9 [†]	92.7 ± 10.5 [†]	88.2 ± 8.2 [†]	97.3
Macro-F1	77.1 ± 26.0 [†]	70.2 ± 27.0 [†]	82.9 ± 24.5 [†]	68.8 ± 26.1	93.2
Kappa Statistic	69.3 ± 37.2 [†]	49.1 ± 45.7 [†]	70.4 ± 41.5 [†]	70.5 ± 35.3	94.9
DECBAE					
Accuracy	96.1 ± 5.5	98.9 ± 1.4	98.7 ± 2.2	95.1 ± 3.3	98.4
Macro-F1	91.7 ± 13.2	87.2 ± 17.8	98.3 ± 3.5	83.0 ± 21.9	93.9
Kappa Statistic	90.9 ± 15.5	79.6 ± 30.2	96.8 ± 6.8	91.7 ± 5.5	97.0
Human Expert					
Accuracy	–	–	–	–	96.3
Macro-F1	–	–	–	–	89.0
Kappa Statistic	–	–	–	–	92.8

Table 2: Mean and standard deviation of model performance on different subsets. [†]Significantly lower than the corresponding metric of DECBAE. Significance is defined by $p < 0.05$ in paired t-test. All numbers are in percentages. High deviations are expected due to the variety of abbreviations in each subset.

agreement (Cohen, 1960). It can also be used to evaluate predictions of multi-label classification:

$$\kappa = \frac{p_o - p_e}{1 - p_e}$$

where p_o is the observed agreement and in the case of classification $p_o = \text{accuracy}$, p_e is the expected agreement which can be achieved by pure chance:

$$p_e = \sum_c p_c \hat{p}_c$$

p_c and \hat{p}_c refer to the proportion of class c in ground truth labels and predictions, respectively. Empirical results in Table 2 show that Kappa statistics are often lower than accuracy and macro-F1, and thus serving as a more distinctive metric for our task.

4.4 Results

In Table 2, we report means and standard deviations of each model’s performance on different subsets evaluated by the three metrics. In all subsets, DECBAE performs significantly better than most other models by large margins. A general trend of DECBAE > BioELMo > biLSTM > BoW-FFN > Majority conserves across subsets.

In the **Random** subset which represents the whole dataset, all metrics of DECBAE exceed 0.90, setting very promising state-of-the-art performance despite the potential noise of the dataset.

In the **Imbalanced** subset where the most frequent definitions consist of over 95% of all the labels, a trivial Majority solution gets over 95% accuracy. However, for macro-F1 and kappa statistic, performance of the baselines drop dramatically while DECBAE can still generate decent results.

DECBAE and BioELMo alone remain robust in **Low-resources** setting. This is due to the transfer learning nature of BioELMo, which utilizes the knowledge encoded in the PubMed abstracts.

Our abbreviation expansion dataset covers roughly 85% of clinical abbreviations mentioned in Xu et al. (2012). On this **Clinical** subset, DECBAE gets pretty good results and vastly outperform other baselines despite its variety in possible definitions (8.5 possible definitions per abbreviation, as shown in Table 1).

On the testset for human performance (i.e.: abbreviation expansion for DAT), DECBAE and even some neural baselines outperform single human expert.

5 Analysis

In Fig. 3, we use confusion matrices to visualize the differences between DECBAE or the human expert and the ground truth labels, for disambiguation of abbreviation “DAT”. The high agreement level between human expert predictions and the automatically assigned labels indicates that our pipeline of collecting the abbreviation disambiguation dataset is valid.

In general, both DECBAE and the human expert perform well in the task, with only few misclassifications. Specifically, DECBAE, and even other neural baselines like biLSTM and BioELMo, outperform the human expert in all metrics. Compared to DECBAE, the human expert is more likely to misclassify direct agglutination test with direct antiglobulin test (9 v.s. 1), and misclassify dementia of the Alzheimer type with dopamine transporter (7 v.s. 0). We show several instances of human and DECBAE’s errors in Table 3.

One limitation of this work is that we just test DECBAE on our automatically collected dataset. Since the proposed model can also be used on other biomedical abbreviation expansion datasets as well, evaluating on other datasets like MSH WSD is a clear future work to do.

Another potential direction for improvement is to accelerate the inference speed. Currently DECBAE uses BioELMo for embedding and abbreviation-specific biLSTM for classification, resulting in two recurrent models in total. Our results show that just BioELMo with several FFN layers also generates decent results, so in some cases we might use only BioELMo as a compromise for faster inference.

6 Conclusion

We present DECBAE, a state-of-the-art biomedical abbreviation expansion model on the automatically collected dataset from PubMed. The results show that, with only minimum expert involvement, we can still perform well in such a domain-specific task by automatically collecting training data from a large corpus and utilize embeddings from pre-trained biomedical language models.

7 Acknowledgement

We are grateful for the anonymous reviewers of BioNLP 2019 who gave us very insightful comments and suggestions. J.L. is supported by NLM training grant 5T15LM007059-32.

Test sentence	Label	Human	DECBAE
The reduction of the number of different segments in DAT compared to controls and patients suffering from depression may be helpful for differential diagnosis.	Def. 4	Def. 1	Def. 4
Reliance on objective brain phenotype measures, for example, those afforded by brain imaging, might critically improve detection of DAT genotype-phenotype association.	Def. 1	Def. 1	Def. 4
DAT was more commonly positive among BO incompatible (21.5% in BO vs. 14.8% in AO , P=0.001) and black (18.8% in blacks vs. 10.8% in nonblacks , P=0.003) infants.	Def. 2	Def. 3	Def. 2
NPY-LI showed a significant reduction in DAT but not in FTD.	Def. 4	Def. 1	Def. 4
The study included 122 healthy subjects, aged 18-83 years, recruited in the multicentre ‘ENC- DAT ’ study (promoted by the European Association of Nuclear Medicine).	Def. 1	Def. 4	Def. 1

Table 3: Some samples of errors made by the human expert and DECBAE. Def. 1: dopamine transporter; Def. 2: direct antiglobulin test; Def. 3: direct agglutination test; Def. 4: dementia of the Alzheimer type.

References

- Rui Antunes and Sérgio Matos. Supervised learning and knowledge-based approaches applied to biomedical word sense disambiguation. *Journal of integrative bioinformatics*, 14(4).
- José Camacho-Collados, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Nasari: Integrating explicit knowledge and corpus statistics for a multilingual representation of concepts and entities. *Artificial Intelligence*, 240:36–64.
- Manuel R Ciosici, Tobias Sommer, and Ira Assent. 2019. Unsupervised abbreviation disambiguation. *arXiv preprint arXiv:1904.00929*.
- Jacob Cohen. 1960. A coefficient of agreement for nominal scales. *Educational and psychological measurement*, 20(1):37–46.
- Dina Demner-Fushman, Willie J Rogers, and Alan R Aronson. 2017. Metamap lite: an evaluation of a new java implementation of metamap. *Journal of the American Medical Informatics Association*, 24(4):841–844.
- 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*.
- Ignacio Iacobacci, Mohammad Taher Pilehvar, and Roberto Navigli. 2016. Embeddings for word sense disambiguation: An evaluation study. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 897–907.
- Rezarta Islamaj Dogan, G Craig Murray, Aurélie Névéol, and Zhiyong Lu. 2009. Understanding pubmed® user search behavior through log analysis. *Database*, 2009.
- Antonio J Jimeno-Yepes, Bridget T McInnes, and Alan R Aronson. 2011. Exploiting mesh indexing in medline to generate a data set for word sense disambiguation. *BMC bioinformatics*, 12(1):223.
- Qiao Jin, Bhuwan Dhingra, William W Cohen, and Xinghua Lu. 2019. Probing biomedical embeddings from language models. *arXiv preprint arXiv:1904.02181*.
- Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.
- SPFGH Moen and Tapio Salakoski2 Sophia Ananiadou. Distributional semantics resources for biomedical text processing.
- Roberto Navigli, David Jurgens, and Daniele Vannella. 2013. Semeval-2013 task 12: Multilingual word sense disambiguation. In *Second Joint Conference on Lexical and Computational Semantics (*SEM), Volume 2: Proceedings of the Seventh International Workshop on Semantic Evaluation (SemEval 2013)*, volume 2, pages 222–231.
- Ahmad Pesaranhader, Stan Matwin, Marina Sokolova, and Ali Pesaranhader. 2019. deepbioword: effective deep neural word sense disambiguation of biomedical text data. *Journal of the American Medical Informatics Association*, 26(5):438–446.
- Matthew E Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. *arXiv preprint arXiv:1802.05365*.
- Alessandro Raganato, Claudio Delli Bovi, and Roberto Navigli. 2017a. Neural sequence learning models for word sense disambiguation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 1156–1167.
- Alessandro Raganato, Jose Camacho-Collados, and Roberto Navigli. 2017b. Word sense disambiguation: A unified evaluation framework and empirical comparison. In *Proceedings of the 15th Conference of the European Chapter of the Association for Computational Linguistics: Volume 1, Long Papers*, pages 99–110.

- Dianne S Schwarz and Michael D Blower. 2016. The endoplasmic reticulum: structure, function and response to cellular signaling. *Cellular and Molecular Life Sciences*, 73(1):79–94.
- Hua Xu, Peter D Stetson, and Carol Friedman. 2007. A study of abbreviations in clinical notes. In *AMIA annual symposium proceedings*, volume 2007, page 821. American Medical Informatics Association.
- Hua Xu, Peter D Stetson, and Carol Friedman. 2012. Combining corpus-derived sense profiles with estimated frequency information to disambiguate clinical abbreviations. In *AMIA Annual Symposium Proceedings*, volume 2012, page 1004. American Medical Informatics Association.
- Jun Xu, Yaoyun Zhang, Hua Xu, et al. 2015. Clinical abbreviation disambiguation using neural word embeddings. *Proceedings of BioNLP 15*, pages 171–176.
- Antonio Jimeno Yepes. 2017. Word embeddings and recurrent neural networks based on long-short term memory nodes in supervised biomedical word sense disambiguation. *Journal of biomedical informatics*, 73:137–147.
- Hong Yu, Won Kim, Vasileios Hatzivassiloglou, and W John Wilbur. 2007. Using medline as a knowledge source for disambiguating abbreviations and acronyms in full-text biomedical journal articles. *Journal of biomedical informatics*, 40(2):150–159.