Article published online: 2022-02-01

Disambiguating Clinical Abbreviations Using a One-Fits-All Classifier Based on Deep Learning Techniques

Areej Jaber[1,2] Paloma Martínez[2]

[1] Applied Computing Department, Palestine Technical University - Kadoorie, Tulkarem, Palestine
[2] Department of Computer Science, Universidad Carlos III de Madrid, Leganés, Spain

Address for correspondence Areej Jaber, MSc, Department of Applied Computing, Palestine Technical University-Kadoorie, Jaffa Street, 7, Tulkarem, Palestine (e-mail: a.jabir@ptuk.edu.ps).

Methods Inf Med 2022;61:e28–e34.

Abstract

Background Abbreviations are considered an essential part of the clinical narrative; they are used not only to save time and space but also to hide serious or incurable illnesses. Misreckoning interpretation of the clinical abbreviations could affect different aspects concerning patients themselves or other services like clinical support systems. There is no consensus in the scientific community to create new abbreviations, making it difficult to understand them. Disambiguate clinical abbreviations aim to predict the exact meaning of the abbreviation based on context, a crucial step in understanding clinical notes.

Objectives Disambiguating clinical abbreviations is an essential task in information extraction from medical texts. Deep contextualized representations models showed promising results in most word sense disambiguation tasks. In this work, we propose a one-fits-all classifier to disambiguate clinical abbreviations with deep contextualized representation from pretrained language models like Bidirectional Encoder Representation from Transformers (BERT).

Methods A set of experiments with different pretrained clinical BERT models were performed to investigate fine-tuning methods on the disambiguation of clinical abbreviations. One-fits-all classifiers were used to improve disambiguating rare clinical abbreviations.

Results One-fits-all classifiers with deep contextualized representations from Bioclinical, BlueBERT, and MS_BERT pretrained models improved the accuracy using the University of Minnesota data set. The model achieved 98.99, 98.75, and 99.13%, respectively. All the models outperform the state-of-the-art in the previous work of around 98.39%, with the best accuracy using the MS_BERT model.

Conclusion Deep contextualized representations via fine-tuning of pretrained language modeling proved its sufficiency on disambiguating clinical abbreviations; it could be robust for rare and unseen abbreviations and has the advantage of avoiding building a separate classifier for each abbreviation. Transfer learning can improve the development of practical abbreviation disambiguation systems.

Keywords
- natural language processing
- clinical abbreviations
- pretrained language model
- word sense disambiguation
- electronic health record

received
August 26, 2021
accepted after revision
October 29, 2021
published online
February 1, 2022

DOI https://doi.org/10.1055/s-0042-1742388.
ISSN 0026-1270.

Georg Thieme Verlag KG, Rüdigerstraße 14, 70469 Stuttgart, Germany

Introduction

Abbreviations are defined as a short form of a word or a phrase. They are extensively used in clinical notes, sometimes for saving time and space and sometimes to hide inconvenient information. Recent studies show that abbreviations constitute 30 to 50% of word count in clinical notes, such as doctor's notes.[1] However, abbreviations are highly ambiguous, which means that one abbreviation has more than one meaning (sense); a recent study shows that nearly one-third of abbreviations are ambiguous.[2] A survey reveals that 43% of a set of abbreviations were identified correctly among more than 200 health care professionals.[3] Their ambiguity is due to several factors; it depends on where they are used (local vs. global scope), and the fact that there are no standard rules for creating them. Usually, abbreviations are created from the first letters of words of their definitions.

Furthermore, it could contain special characters, punctuation marks like 5-FU (5-fluorouracil), or a number like T1 (spin-lattice relaxation time). For example, "AV" could mean "Aortic valve," "Atrioventricular" which is related to the heart, or "Anteverted" related to the uterus. Determining its exact meaning depends on the context in which it is used.

Recognizing and expanding clinical abbreviations is essential for health care systems, which could help clinicians make decisions, predict health outcomes, and improve quality of care.[4] Moreover, disambiguating clinical abbreviations can help physicians, nurses, and patients understand them and prevent medically dangerous misinterpretations.[5]

Disambiguating abbreviations is considered as a type of word sense disambiguation (WSD) task. In natural language processing (NLP), WSD is the task of determining the correct sense of a word based on the surrounding context. WSD is considered a classification problem, in which given an ambiguous word and every possible meaning, the goal is to classify it into one of its sense's classes based on the context.

Disambiguating clinical abbreviations required annotated data (corpora). The annotation process is considered tedious, expensive, and time consuming.[6] Additionally, due to mainly privacy issues, few annotated data are available. Most previous works[7–9] were based on building a separate classifier to disambiguate each abbreviation but this approach is considered insufficient for rare and unseen abbreviations. Hence, building generalizable methods such as having one classifier of all abbreviations could improve disambiguating unusual and unseen abbreviations.

Three main approaches are used for WSD.[10] First, knowledge-based (KB) approaches are based on existing lexical resources such as dictionaries or semantic networks that usually do not have enough coverage.[11] KB approaches do not require annotated data since they exploit the graph structure of semantic networks to determine the most practical sense. Furthermore, this approach is still not applicable to clinical abbreviations because no standard lexical resources could cover all continuously emerging senses related to clinical abbreviations. Second, unsupervised methods assume that similar senses occur in similar contexts; this approach was applied to generate senses inventory by applying Tight Clustering for Rare Senses.[12]

Third, supervised approaches use a set of training sets containing several examples like (X_i, Y_i), $\ldots(X_n, Y_n)$, where X_i is a vector feature that represents the target abbreviation, and Y_i represents the correct class (expansion or sense) for the target abbreviation. Supervised methods depend heavily on extensive manually annotated data.

Most existing WSD systems to disambiguate clinical abbreviations are supervised methods that build a specific classifier for each abbreviation in the data set.[13,14] They cannot generalize across abbreviations and therefore require sufficient sense-annotated data for every abbreviation to perform an adequate disambiguation.[15]

The surrounding context of an ambiguous word, which is known as a sequence, plays a vital role in providing a clear evidence for this classification.[16] Extracting the features of this context could be performed in different ways. In the traditional feature-based approaches, the features are collected from each word in a sequence individually, like the part-of-speech (POS) tag, the position of the word (left or right), and how far from the ambiguous word.[17] In recent years, low-dimensional word representation vectors[14] are trained on unannotated text data to generate static word embeddings, which are used in WSD tasks. These representations are fed into a neural network to capture the whole sentence representation and each word representation. However, in both approaches, the word representations are independent of the context and it could not be changed based on the context in which it appears.

Pretrained contextualized representations of the sequence, like ELMo[18] and Bidirectional Encoder Representation from Transformers (BERT),[19] are recent approaches that have proved their efficiency on downstream NLP tasks[20,21] and outperform many state-of-the-art architectures in these tasks. BERT representations are obtained by training a huge amount of raw text on neural encoders in two generic tasks (predicting the next word in a sentence and predicting the next sentence in a text).[19] These models can be fine-tuned on new downstream tasks such as named entity recognition[22] or sentence classification[23] by further training steps with small annotated data. The main advantage of these pretrained models is that they provide a different representation for each word, thus determining the different meanings for this word.

In this work, a BERT pretrained language model was fine-tuned using the next sentence prediction (NSP) approach with a one-fits-all classifier to disambiguate clinical abbreviations. Our work improved the accuracy and outperformed the state-of-the-art of this task.

Objectives

Abbreviations are generally used to save time and space while writing in the patients' medical records. Some challenging issues are: (1) mostly, clinical abbreviations are not accompanied by expansions in the text as it happens in biomedical literature, (2) different rules are used to create

Table 1 Some examples of how abbreviations are formed

Rule	Abbreviation	Sense
Truncating the end of long form	DIP	**DIP**ropionate
First letter initialization from each word	VBG	**V**enous **b**lood **g**as
Syllabic initialization	US	**U**ltra**S**ound
Combination if the beginning of some of the words of long form	Ad lib	**Ad lib**itum
Symbols/synonyms substitution or initialization	T3	Triiodothyronine

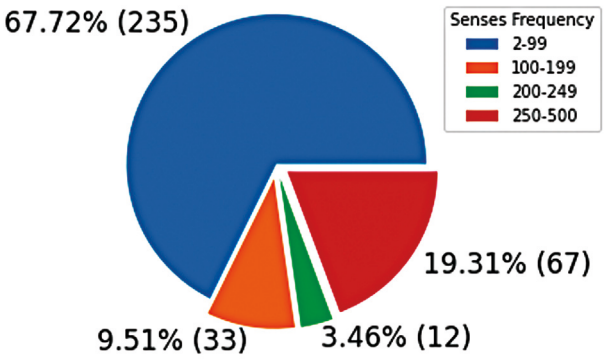

Fig. 1 Pie chart of senses and number of examples relation, showing the frequency and the percentage of each sense in the University of Minnesota (UMN) data set.

them; ►**Table 1** shows some examples of rules used, (3) highly ambiguous terms (few abbreviations with only one expansion), (4) the scope of abbreviations concerning the community using them; there are international standardized resources or national ones, like the list of medical abbreviations proposed by the Spanish Ministry of Health,[24] but others have a more local scope that could be abbreviations used by just one hospital or even just one clinician, and (5) bilingual <abbreviation, expansion> pairs, for instance, PSA (prostatic-specific antigen) English abbreviation is used in Spanish clinical narrative instead of APE (*antígeno prostático específico*).

Research in developing automatic expansion processes is needed to minimize the side effect of misunderstanding abbreviations in clinical documentation and avoid patient safety issues. This work aims to develop a single classifier model that disambiguates clinical abbreviations using the recent existing language models that deals with rare senses and proves its efficiency in many WSD tasks, and outperforms the state-of-the-art tasks.

Methods

Clinical abbreviation disambiguation can be solved as a multiclassification task. It takes a text with previously recognized abbreviations as input and outputs the correct expansion (sense) of each abbreviation. In our approach, we propose different pretrained Bidirectional Encoder Representations from Transformers (BERT) models to represent the input texts. The method has three stages that are described below: (1) prepare a data set to fit input requirements for a pretrained model, (2) adjust the pretrained model's weights and do further training processes on the data set, and (3) apply a neural classifier.

Dataset Description

This study uses one of the few publicly annotated clinical notes data sets created by the University of Minnesota-affiliated (UMN) Fairview Health Services in the Twin Cities.[25] It was collected from admission notes, inpatient consult notes, and discharge summaries between 2004 and 2008. The data set contains 75 abbreviations of the most frequent acronyms and abbreviations from this clinical repository. Each abbreviation has 500 sentences that are annotated with different senses. There are 351 senses with an average of 4.7 senses per abbreviation (highly ambiguous abbreviations). ►**Table 2** illustrates the senses distributions of three abbreviations ("AMA," "BAL," and "OTC") taken from the UMN data set.

As shown in ►**Fig. 1**, 235 senses have between 2 and 99 examples in the data set, while 67 senses have between 200 and 249 examples. Consequently, implementing a separate classifier for each abbreviation requires more examples for each sense. It is noticeable how the senses are strongly unbalanced such as most abbreviations in the data set.

Table 2 Sample of University of Minnesota (UMN) data set abbreviations with its number of senses and distributions

Abbreviations	Sentences	Tokens	Senses	Senses no.	Senses (%)
AMA	2,881	37,887	Against medical advice	444	88.8
			Advanced maternal age	31	6.2
			Antimitochondrial antibody	25	5.0
BAL	3,267	38,483	Bronchoalveolar lavage	457	91.4
			Blood alcohol level	43	8.6
OTC	6,173	37,356	Over the counter	469	93.8
			Ornithine transcarbamoylase	31	6.2

Table 3 The pretrained models architecture is used in this study

Characteristic	No.
Layers	12
Hidden units	768
Self-attention heads	12
Total trainable parameters	110M

Approach

BERT[19] is an embedding layer representation or language model that is obtained by making use of Transformer architecture,[26] an attention mechanism that learns contextual relations between words (or subwords) in an annotated text. The transformer has two components, but only the encoder is used in BERT. On the contrary to directional models, which read the input sequence either left-to-right or right-to-left, BERT reads the whole sequence of words in both directions at once. This is the main feature for BERT that allows the model to deeply represent words based on all the surrounding contexts.

BERT uses two strategies during the training process: Masked Language Model (MLM) and Next Sentence Prediction (NSP). In the MLM strategy, before training, 15% of the words in the input sequence are replaced with [MASK] tokens; during the training, the model goal is to predict the correct words based on the context surrounding of [MASK] tokens. On the other hand, in NSP, the input sample is composed of a pair of sentences, and the goal of the classification is to predict whether the pair of sentences is correlated or not. Thanks to these strategies, the BERT encoder is able to obtain a vector not only for each word, but also for each sense of a word.

Furthermore, BERT can be used as part of a feature-based approach or fine-tuning approach. For feature-based WSD models, deep contextualized word embedding representations from BERT are used as input features for task-specific architectures.[27] Fine-tuning WSD models use the weights of the pretrained models and perform an additional training using annotated corpora.[28]

In this work, a fine-tuning approach was applied to the UMN data set, implementing NSP strategies with a one-fits-all classifier. Three clinical pretrained BERT models were fine-tuned to disambiguate clinical abbreviations. These models were built from BERT-base types, which mean that they have the same architecture (►**Table 3**). The differences among them are the type and size of the resources that were used to pretrain the model, as described below:

- Clinical BioBert:[29] Two million clinical notes from the MIMIC-III[30] v1.4 database were used to train the BioBERT[31] model to generate a pretrained word embedding.
- BlueBERT:[32] This BERT model was trained with more than 4,000 million words from PubMed abstracts (biomedical resources) combined with more than 500 million words from MIMIC-III as clinical resources.
- MS-BERT:[33] This model was developed in the University of Toronto and the Data Science and Advanced Analytics department at St. Michael's Hospital. A BlueBERT model is fine-tuned on another 35 million words from Multiple Sclerosis (MS) examinations clinical notes.

Preprocessing

As described below, different preprocessing steps were performed on the text before it was fed into the model.

- Cleaning: First, punctuation and special characters were removed. Then, all the characters were converted to lowercase.
- Tokenization: Contexts and the expansions are tokenized utilizing each model tokenizer applying Word-Piece tokenizing techniques. If the entire word does not exist on the model vocabulary, it will be broken down into substrings to handle unseen words.
- Special token addition: Since our objective is to represent the problem as a NSP BERT approach, a special token [CLS] was added at the beginning of the first segment, which is the context in our data set; then two special tokens [SEP] were added, the first one was to separate between the two segments (context and expansion) and the second was added at the end of the second segment.
- Truncate and padding, attention mask: All input sequences should have the same length, which is 512 tokens, to fit the BERT input format; so that if the sequence length is more than 512, it was truncated; however, [PAD] tokens were added at the end of sequences which is less than the required length. Attention mask layers were added to indicate to the model which tokens should be attended to, and which should not. For sequence classification with BERT, token type IDs (also called segment IDs) are deployed. They were represented as a binary mask identifying the two types of sequence in the model (context and expansion in this work).
- Data set splitting: 80% of the data set was used as a training data set with 29,560 sentences and 20% from

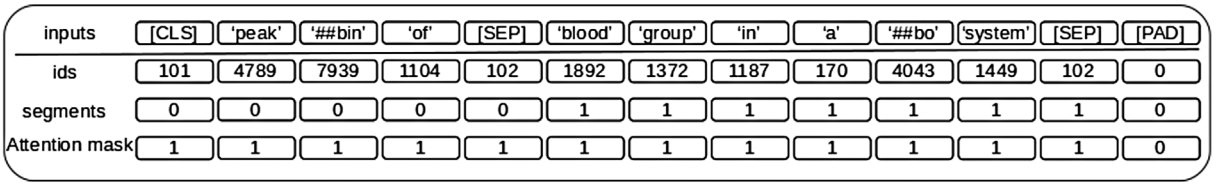

Fig. 2 An example of input representation for one sequence including [CLS], [SEP], and [PAD] tokens, in addition to added segments, attention mask, and embedding layers.

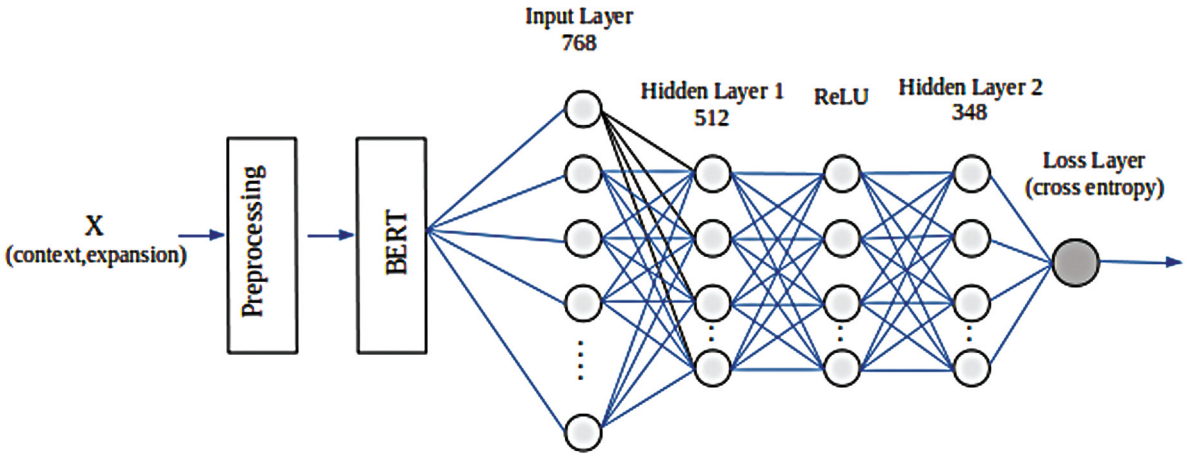

Fig. 3 The architecture of the proposed model.

the data set was separated in half between development and test data sets with 3,695 samples for each one.

►**Fig.2** shows an example of the input sequence for the abbreviation "AB." Here, the context is "... received phototherapy for a peak bilirubin level of 11.5 mg%. Her blood type was **AB** positive...," and the correct expansion is "blood group in ABO system." [CLS] token was added at the beginning of the context, two [SEP] tokens were added to indicate two sequence classifications. Finally, the [PAD] tokens unify the length of all sequences as mentioned above.

- Labeling: A multiclass model for developing one classifier for the entire data set is the contribution of this research. The data set contains 348 unique senses, labeled from 0 to 347, so the model has 347 different classes.

Proposed Architecture

This work aims to develop a one-fits-all classifier instead of one classifier for each abbreviation in the data set. To achieve it, a simple architecture was added on the top of the pretrained BERT layer to prove that BERT can provide a great contextualized presentation that could perform well without further sophisticated architectures.

►**Fig. 3** shows the three main stages of the system. The first is the preprocessing step, and then the sequence embedding was computed using a BERT encoder. Since the hidden state of [CLS] token represents the whole sequence, it was fed into a feedforward layer,[34] activation ReLU,[35] and another feedforward layer as the following equation:

$$P = L_2\left(ReLU\left(L_1(f)\right)\right)$$

where are fully connected linear layers, Notice that the softmax layer was not explicity implemented because it is within the cross-entropy loss function.

The hyperparameters for all experiments were identically adjusted in the three pretrained models. The batch size was 8, and the epochs for the models were 5. The learning rate was 1×10^{-5}. Adam optimizer[36] was used, in addition to the cross-entropy loss function.

Evaluation Metric

The accuracy measure was used to evaluate the performance of the three models and compare it with other similar models. The accuracy is defined as follows:

$$Accuracy = \frac{No.ofcorrectpredictions}{No.oftotalpredictions}$$

Results

►**Table 4** gives overall results on the UMN data set. The first observation is that the three models achieved a high accuracy with a bit of difference. A multiclass pretrained model based on MS_BERT achieved the best performance among the three, with a slight difference.

Between all-models comparisons, since the UMN data set was released, many researchers tried to work on it from different aspects, like increasing the number of samples for each sense in the data set or implementing a separated classifier for each abbreviation in the data set. The result of our models will be compared to the most similar work, which focuses on building one classifier for all the abbreviations in the data set. ►**Table 5** shows the results of the most related previous work concerning this work.

The first two approaches Convolutional Neural Network (CNN)[8] and CLASS-GATOR[37] focus on autogenerated training data to use in their models in addition to the inputs'

Table 4 Accuracy results for the University of Minnesota (UMN) data set

Pretrained language model (LM)	Accuracy (%)		
	Training	Validation	Test
Bio_Clinical	98.85	98.97	98.99
BlueBERT	98.46	98.73	98.75
MS_BERT	98.98	99.11	99.13

Note: Slightly different between the three pretrained models.

Table 5 Accuracy results for the University of Minnesota (UMN) data set compared with several previous works

Methods	Accuracy (%)
Multiclassifier	
Convolutional Neural Network (CNN)[8]	77.83
CLASS-GATOR[37]	76.92
One-fits-all classifier	
ELMo + Topic[9]	70.41
Latent Meaning Cells (LMC)[38]	71.00
Candidate Classification[39]	98.39
MS_BERT (our approach)	99.13

Note: Our model achieves the state-of-the-art with MS_BERT pretrained model.

representation. Pretrained static embeddings from PubMed were combined with POS tags, and clinical note features were used to represent the input in the UMN data set, then fed into CNN models. The reverse substitution was used to generate more training data by searching exact matching strings for the senses.[8] CLASS-GATOR[37] trained a logistic regression model on a biomedical data set (PubMed), then tested the trained model on an unannotated clinical data set from MIMIC III. This approach was applied to 52 abbreviations from the UMN data set with BioBERT, which achieved better performance than their logistic regression model.

On the one-fits-all classifier side, the first two approaches ELMo + Topic[9] and Latent Meaning Cells (LMC)[38] combined the content with external knowledgeable parts, like headers, to get the contextualized representation for the data set; the first one used ELMo. This approach was applied to 30 abbreviations from the same data set and tested on 15 abbreviations from the MIMIC-III data set. The second used LMC.[38] However, in the third approach (Candidate Classification),[39] the classifier was fed with embeddings of candidate expansions contextualized by the context as input.

Error Analysis

In this section, we analyze the error of the MS_BERT predictions, which had the best performance. The test data set contains 38 unseen senses, and the model correctly predicted 29 of them. After analyzing the results, the errors are mainly due to the right and predicted senses belonging to the same medical problem, and the classifiers cannot distinguish them. For example, "AVR" has 6 senses, two of which belong to heart problems; "aortic valve resistance" and "aortic valve replacement." However, "aortic valve replacement" has 318 examples in the whole data set, and "aortic valve resistance" has 4 examples. The exact situation happened with "LE" and "RA" abbreviations, both having senses related to heart problems, "left ventricle:LV" and "right atrium" with 5 and 345 examples, respectively. "GT" has 16 senses, "guttae" has one example and "gutta" has 16 examples. "guttae" is the plural of "gutta," another case that the model was not able to distinguish.

Discussion

We propose in this research a solution to disambiguate clinical abbreviations to advance in methods to solve the issue of rare senses (senses with few examples in training data). Three pretrained BERT models have been tested to analyze the impact of previously generated language models in the task of WSD. ►Table 4 shows that the results are closely related, with MS BERT achieving the best results by a slight margin. Therefore, models that have been pretrained with extra clinical data perform better than models that have just been pretrained with biomedical scientific literature.

The results are also compared with the most recent previous work in WSD for abbreviations. From the results shown in ►Table 5, it is noticeable that our WSD models outperform the state-of-the-art of the previous work using the same data set.

Conclusion

Currently, more than 80% of data in electronic health records is unstructured data (images, text, etc.); it is essential to foster research in information extraction from clinical text to obtain structured data that could be used in decision-making processes. Abbreviations are very common in the clinical narrative. The aim of this research is to explore methods to disambiguate senses of abbreviations in the clinical text, which has been less explored than biomedical literature. In this article, publicly published pretrained language models such as BioBERT, BlueBert, and MS_Bert were fine-tuned to disambiguate clinical abbreviations. In contrast to the well-known approaches of building a specific classifier for each abbreviation using supervised disambiguating methods, we explore one-fits-all classifiers, which proved their ability to achieve the best performance. This one-fits-all classifier approach could be applied not only to restricted resources domains (such as those in the clinical domain) but also to process text in languages other than English.

In future works, we plan to extend the work to auto disambiguate more clinical abbreviations, not restricted to the definite number used in this work and explore different neural network architectures for disambiguating clinical abbreviations that make clinical notes more readable for the patients.

Ethical Approval

No human subjects were involved in this project, and institutional review board approval was not required.

Funding

This work has been supported by the Madrid Government (Comunidad de Madrid-Spain) under the Multiannual Agreement with UC3M in the line of Excellence of University Professors (EPUC3M17), and in the context of the V PRICIT (Regional Programme of Research and Technological Innovation) and Palestine Technical University - Kadoorie (Palestine). The work was also supported by the PID2020-116527RB-I00 project.

Conflict of Interest
None declared.

References

1 Grossman LV, Mitchell EG, Hripcsak G, Weng C, Vawdrey DK. A method for harmonization of clinical abbreviation and acronym sense inventories. J Biomed Inform 2018;88:62–69
2 Holper S, Barmanray R, Colman B, Yates CJ, Liew D, Smallwood D. Ambiguous medical abbreviation study: challenges and opportunities. Intern Med J 2020;50(09):1073–1078
3 Sinha S, McDermott F, Srinivas G, Houghton PWJ. Use of abbreviations by healthcare professionals: what is the way forward? Postgrad Med J 2011;87(1029):450–452
4 Yim WW, Yetisgen M, Harris WPKS, Kwan SW. Natural language processing in oncology: a review. JAMA Oncol 2016;2(06): 797–804
5 Murff HJ, FitzHenry F, Matheny ME, et al. Automated identification of postoperative complications within an electronic medical record using natural language processing. JAMA 2011;306(08): 848–855
6 Hanauer D, Aberdeen J, Bayer S, et al. Bootstrapping a de-identification system for narrative patient records: cost-performance tradeoffs. Int J Med Inform 2013;82(09):821–831
7 Jaber A, Martínez P. Disambiguating Clinical Abbreviations using Pre-trained Word Embeddings. In: Proceedings of the 14th International Joint Conference on Biomedical Engineering Systems and Technologies. Vol. 5. SCITEPRESS - Science and Technology Publications; 2021:501–508
8 Joopudi V, Dandala B, Devarakonda M. A convolutional route to abbreviation disambiguation in clinical text. J Biomed Inform 2018;86:71–78
9 Li I, Yasunaga M, Nuzumlalı MY, Caraballo C, Mahajan S, Krumholz H, Radev D. A neural topic-attention model for medical term abbreviation disambiguation 2019;arXiv preprint arXiv:1910.14076
10 Navigli R. Word sense disambiguation: a survey. ACM Comput Surv 2009;41(02):1–69
11 Mihalcea R. Knowledge-Based Methods for WSD. In: Agirre E, Edmonds P, eds. Word Sense Disambiguation: Algorithms and Applications. Dordrecht: Springer Netherlands;2006:107–131
12 Xu H, Wu Y, Elhadad N, Stetson PD, Friedman C. A new clustering method for detecting rare senses of abbreviations in clinical notes. J Biomed Inform 2012;45(06):1075–1083
13 Finley GP, Pakhomov SVS, McEwan R, Melton GB. Towards comprehensive clinical abbreviation disambiguation using machine-labeled training data. AMIA Annu Symp Proc 2017;2016:560–569
14 Wu Y, Xu J, Zhang Y, Xu H. Clinical abbreviation disambiguation using neural word embeddings. In Proceedings of BioNLP 15; 2015:171–176
15 Màrquez L, Escudero G, Martínez D, Rigau G. Supervised corpus-based methods for WSD. In: Agirre E, Edmonds P, eds. Word Sense Disambiguation: Algorithms and Applications. Dordrecht-Springer Netherlands;2006:167–216
16 Wang Y, Hou Y, Che W, Liu T. From static to dynamic word representations: a survey. Int J Mach Learn Cybern 2020;11:1611–1630
17 Moon S, Pakhomov S, Melton GB. Automated disambiguation of acronyms and abbreviations in clinical texts: window and training size considerations. AMIA Annu Symp Proc 2012; 2012:1310–1319
18 Peters M, Neumann M, Iyyer M, et al. Deep Contextualized Word Representations. arXiv preprint 2018;arXiv:1802.05365
19 Devlin J, Chang MW, Lee K, Toutanova K. BERT: pre-training of deep bidirectional transformers for language understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, (Long and Short Papers) Vol. 1. 2019: 4171–4186
20 Liu Y, Lapata M. Text summarization with pretrained encoders. CoRR 2019:abs/1908.0
21 Chalkidis I, Fergadiotis M, Malakasiotis P, Androutsopoulos I. Large-scale multi-label text classification on {EU} Legislation. CoRR 2019:abs/1906.0
22 Hakala K, Pyysalo S. Biomedical Named Entity Recognition with Multilingual {BERT}. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. Hong Kong, China: Association for Computational Linguistics; 2019:56–61
23 Gao Z, Feng A, Song X, Wu X. Target-dependent sentiment classification with BERT. IEEE Access 2019;7:154290–154299
24 Laguna JY, Alberola V. Dictionary of medical acronyms, abbreviations and hospital discharge codification related terms. Ministry of Health Publications Center 2003
25 Moon S, Pakhomov S, Liu N, Ryan JO, Melton GB. A sense inventory for clinical abbreviations and acronyms created using clinical notes and medical dictionary resources. J Am Med Inform Assoc 2014;21(02):299–307
26 Vaswani A, Shazeer N, Parmar N, et al. Attention is all you need. CoRR 2017:abs/1706.0
27 Jin Q, Liu J, Lu X. Deep contextualized biomedical abbreviation expansion 2019; arXiv preprint arXiv:1906.03360
28 Du, J, Qi, F, Sun M. Using BERT for word sense disambiguation. arXiv preprint 2019;arXiv:1909.08358
29 Alsentzer E, Murphy JR, Boag W, Weng WH, Jin D, Naumann T, McDermott M. Publicly available clinical BERT embeddings. arXiv preprint 2019;arXiv:1904.03323
30 Johnson AEW, Pollard TJ, Shen L, et al. MIMIC-III, a freely accessible critical care database. Sci Data 2016;3:160035
31 Lee J, Yoon W, Kim S, et al. BioBERT: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics 2020;36(04):1234–1240
32 Peng Y, Yan S, Lu Z. Transfer learning in biomedical natural language processing: an evaluation of BERT and ELMo on ten benchmarking datasets. arXiv preprint 2019;arXiv:1906.05474
33 MS-BERT. Accessed December 22, 2021: https://huggingface.co/NLP4H/ms_bert
34 Schmidhuber J. Deep learning in neural networks: an overview. Neural Netw 2015;61:85–117
35 Agarap AF. Deep Learning using Rectified Linear Units (ReLU). arXiv preprint 2019;arXiv:1803.08375v2 [cs.NE]
36 Kingma DP, Ba JL. Adam: a method for stochastic optimization. In: 3rd International Conference on Learning Representations, ICLR 2015 - Conference Track Proceedings. 2015:1–15
37 Kashyap A, Burris H, Callison-Burch C, Boland MR. The CLASSE GATOR (CLinical Acronym SenSE disambiGuATOR): a method for predicting acronym sense from neonatal clinical notes. Int J Med Inform 2020;137:104101
38 Adams G, Ketenci M, Bhave S, Perotte A, Elhadad N. Zero-shot clinical acronym expansion via Latent Meaning Cells. CoRR 2020: abs/2010.0:12–40
39 Kim Juyongand Gong Linyuanand Khim Justinand Weiss Jeremy C. and Ravikumar P. Improved Clinical Abbreviation Expansion via Non-Sense-Based Approaches. 2020. Available at: http://proceedings.mlr.press/v136/kim20a.html