Nurses and Midwives in the Digital Age
M. Honey et al. (Eds.)

doi:10.3233/SHTI210717

Comparison of Word and Character Level Information for Medical Term Identification Using Convolutional Neural Networks and Transformers

Sandaru SENEVIRATNE[a,1], Artem LENSKIY[a], Christopher NOLAN[b], Eleni DASKALAKI[a] and Hanna SUOMINEN[a,c,d]
[a] *School of Computing, The Australian National University (ANU), Australia*
[b] *ANU Medical School and John Curtin School of Medical Research, ANU, Australia*
[c] *Data61, Commonwealth Scientific and Industrial Research Organisation, Australia*
[d] *Department of Computing, University of Turku, Finland*

Abstract. Complexity and domain-specificity make medical text hard to understand for patients and their next of kin. To simplify such text, this paper explored how word and character level information can be leveraged to identify medical terms when training data is limited. We created a dataset of medical and general terms using the Human Disease Ontology from BioPortal and Wikipedia pages. Our results from 10-fold cross validation indicated that convolutional neural networks (CNNs) and transformers perform competitively. The best F score of 93.9% was achieved by a CNN trained on both word and character level embeddings. Statistical significance tests demonstrated that general word embeddings provide rich word representations for medical term identification. Consequently, focusing on words is favorable for medical term identification if using deep learning architectures.

Keywords. Terminology, text simplification, deep learning, word embedding

1. Introduction

Technological advancements have resulted in easy access to vast amounts of information for patients, their next of kin, and carers to understand a disease and its implications for health which has a direct impact on their ability to engage in optimal management of the disease. However, given that medical text contains domain-specific terms and shorthand that are difficult for a person with limited medical background to understand which can result in misunderstandings, distress, and incorrect care decisions [9]. To bridge the gap of knowledge and avoid misinterpretation of text, it is vital to develop new methods to improve the understandability through simplification of text.

Text simplification can be lexical and/or syntactic [8]. Lexical simplification of medical text could aim to modify the content to become easier for laypeople to understand. Its first step is complex word identification (CWI) in text [2]. Given a sentence 'Early indications are related to hyperglycemia and include polydipsia,

[1] Corresponding Author, Sandaru Seneviratne, ANU School of Computing, 145 Science Road, Canberra, ACT 2600, Australia; E-mail: sandaru.seneviratne @anu.edu.au.

polyphagia, polyuria, and blurred vision', such CWI system should be able to identify medical terms of 'hyperglycemia', 'polydipsia', 'polyphagia', and 'polyuria' as complex for laypeople (Figure 1). This will then initiate a search for definitions or alternatives for these identified complex terms.

Existing CWI systems use thresholds, lexica, and machine learning models with heavy feature engineering [3]. To make feature engineering lighter, advanced deep learning models like Long Short Term Memory (LSTM) and Convolutional Neural Networks (CNN) have also been used for CWI where the task is treated as a sequence modelling problem, using sentences with labelled words [1; 3]. Transformers [10], with the self-attention mechanism, are also heavily used in sequence modelling tasks. For natural language processing (NLP), these deep learning models use word embeddings trained on large datasets, which capture semantic similarities among words. Word embeddings are commonly used to map raw text data into numerical representations (i.e., features) deviating from more classical approaches of heavy feature engineering for text. A number of pretrained embedding models, such as word2vec trained on Google News and Glove [6], are available for general NLP tasks. For tasks specific to medicine, medical word embeddings created using resources such as PubMed [11] are available.

In this study, we explore, using deep learning models, the effectiveness of word and character embeddings for medical term identification that is modelled as a CWI task. Given the limited medical data for training, manual annotation of words in sentences is expensive considering the large scale required by these deep learning models [4]. Therefore, we address the CWI problem as a simplified problem of word classification through exploring the impact of word representations trained on general and medical text corpora, along with character embeddings, on identifying complex medical terms.

Our main outcomes are giving evidence of the following for medical term identification: (i) both CNNs and transformers perform competitively and (ii) both word and medical embeddings provide rich representations of words, as opposed to one-hot encoded character embeddings.

2. Methods

The dataset, collected from the Human Disease Ontology from BioPortal [7] and Wikipedia pages, contained 16,580 terms and consisted of 6,932 unique medical terms and 9,648 unique non-medical terms. Medical terms included symptoms, diseases, and medical drug names. A wide range of topics from Wikipedia were used to obtain non-medical terms. All the extracted terms were preprocessed. The terms obtained from Human Disease Ontology were identified as medical terms and terms obtained from Wikipedia pages were identified as non-medical terms.

The medical CWI problem was defined as a binary classification problem where the target is to predict if an input term $\mathrm{x}_i \in \mathbb{R}^n$ should be classified as 0 or 1 with n being the number of features for the input. Mapping the problem to the CWI task, features of an input term could be its frequency, length, number of synonyms, hyponyms, hypernyms, word embedding, or character embedding. The focus of this paper was on word and character embeddings as features for CWI using CNNs and transformers (Figure 1). Two different word embedding models and one character embedding model were used for the experiments.

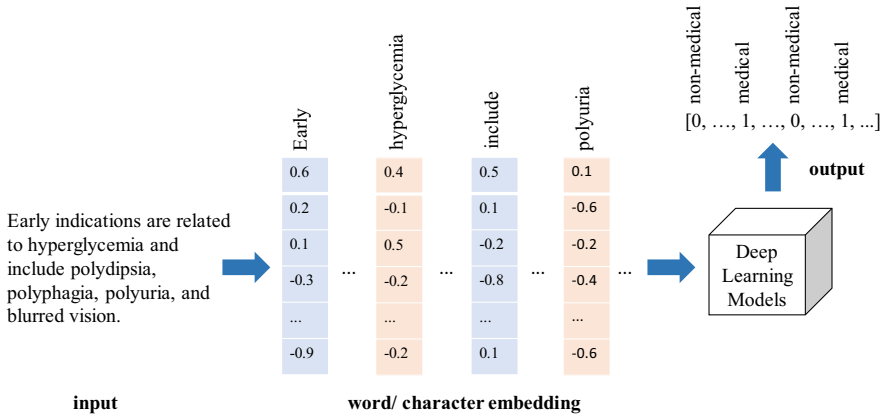

Figure 1. Proposed methodology for medical CWI using a complex medical sentence input.

For each term in the dataset, a 300-dimensional word embedding from the word2vec model trained on Google News and a 200-dimensional medical word embedding from the BioWordVec [12] model trained on PubMed text data were obtained. Out-of-the-vocabulary words were assigned a zero vector. For character embedding, a set of 30 characters (incl. lower-case English alphabet letters, a 'UNK' character, and common punctuation mark characters) were defined. For each character, a 30-dimensional one-hot encoding was created. These character encodings were used to obtain a word vector for each word in the dataset.

CNN models with word, medical, and character embeddings used one convolutional layer followed by a rectified linear unit, max pooling, and a linear layer to obtain the final output using a SoftMax layer. CNN models with word and character embeddings had two convolutional layers with character embeddings as the input to the first convolutional layer and a concatenation of the output from the first layer and the word embedding as the input to the second convolutional layer. These models were trained with a learning rate of 1e-3.

Transformer models with the word, medical, and character embeddings used one transformer encoder layer [10] followed by a linear and a SoftMax layer to obtain the final output. Transformer models with word and character embeddings had two transformer encoder layers with character embedding as the input to the first layer and a concatenation of output from the first layer and word embedding as input to the second transformer encoder layer. These models were trained with a learning rate of 1e-5.

Each model was implemented using PyTorch [5]. Adam optimizer was used with categorical cross entropy as the loss function, with a batch size of 256 and 60 epochs.

As evaluation metrics, macro-averaged precision, recall, and F score were used. To compare the models, a 10-fold cross validation approach that produced ten F scores was employed. Values from each model pair were compared to check if they originated from the same distribution; if not, the model with the higher mean F score was assumed to be better.

3. Results

We observed that CNNs and transformers trained on word and medical embeddings performed well in the medical CWI task (Table 1). CNNs trained on character and word embeddings gave the best F score (93.9%). In comparison, models trained only on character embeddings gave poor results. However, both medical and general word embeddings performed well. All the models except those based on character embeddings gave high precision, recall, and F score values.
To evaluate the statistical significance between the F scores of the model pairs, we ran pairwise tests. Statistically significant differences were observed across all the model pairs except for four denoted by the letters a, b, c, d in Table 1.

Table 1. Macro-averaged precision, recall and F score values for the CNN and Transformer models based on character embeddings and word embeddings for the test dataset.

Architecture\Embeddings	Medical Word	Character	Word	Character & Word
CNNs:				
Precision	0.9279	0.7550	0.9412	0.9414
Recall	0.9283	0.7537	0.9378	0.9383
F score	0.9281	0.7532	0.9393[a,c]	0.9396[a,b]
Transformers:				
Precision	0.9025	0.7358	0.9432	0.9165
Recall	0.9216	0.7353	0.9354	0.9169
F score	0.9214[d]	0.7347	0.9387[b,c]	0.9165[d]

4. Discussion

From the results obtained, we can conclude word and medical embedding-based CNNs and transformers perform competitively in identifying the medical terms. However, models based only on character embeddings showed poor results even though they make it possible to compute vectors for misspelled and rare words. Statistical significance tests demonstrated that general word embeddings perform well in medical CWI. Based on these results, we can consider that the word embeddings from Google News word2vec model and BioWordVec model provide rich representations of words compared to the one-hot encoded character embeddings.

We modeled our data using word and character level embeddings in such a way that the models could focus on specific data points in the embeddings which were the most useful in the classification task. Both CNNs and transformers showed promising results for medical CWI in the proposed approach.

Some limitations in the study are use of one-hot encoding for character embedding. The creation of character embeddings should be further investigated to identify the full potential of the character level features. For future work, it would be interesting to explore sub word representations and different types of medical embeddings. To address the issue of lack of training data in medical NLP, we modeled the problem as a word classification task and simplified the task in such a way that required data can be easily collected. The experiments were performed on the created dataset. However, it would be interesting to use related existing baseline datasets to validate the proposed approach [9].

Our study explored different word representations for medical CWI using deep learning models. However, it is crucial to investigate the applicability of the proposed

methods in real-world applications and to study how they can be further improved and refined in a clinical setting through the feedback from healthcare workers and customers.

5. Conclusions

Both word and medical embedding-based CNNs and transformers performed competitively in medical CWI. Experiments suggested that one-hot encoded character embeddings are insufficient for deep learning models to achieve their best potential. These conclusions guide next steps towards medical CWI with limited training data.

Acknowledgements

This research was funded by and has been delivered in partnership with Our Health in Our Hands (OHIOH), a strategic initiative of the ANU, which aims to transform health care by developing new personalized health technologies and solutions in collaboration with patients, clinicians and health-care providers. We gratefully acknowledge the funding from the ANU School of Computing for the first author's PhD studies.

References

[1] Aroyehun ST, Angel J, Alvarez DA, Gelbukh A. Complex word identification: Convolutional neural network vs. feature engineering. In: Proceedings of the thirteenth workshop on innovative use of NLP for building educational applications, 2018, pp. 322-327.
[2] Bingel J, Paetzold G, Søgaard A. Lexi: A tool for adaptive, personalized text simplification. In: Proceedings of the 27th International Conference on Computational Linguistics, 2018, pp. 245-258.
[3] Gooding S, Kochmar E. Complex word identification as a sequence labelling task. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics, 2019, pp. 1148-1153.
[4] Johnson M, Anderson P, Dras M, Steedman M. Predicting accuracy on large datasets from smaller pilot data. In: Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers), 2018, pp. 450-455.
[5] Paszke A, Gross S, Massa F et al. PyTorch: An imperative style, high-performance deep learning library.
[6] Pennington J, Socher R, Manning CD. Glove: Global vectors for word representation. In: Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP), 2014, pp. 1532-1543.
[7] Rubin DL, Moreira DA, Kanjamala P, Musen MA. BioPortal: A web portal to biomedical ontologies. In: AAAI Spring Symposium: Symbiotic Relationships between Semantic Web and Knowledge Engineering, 2008, pp. 74-77.
[8] Shardlow M. A survey of automated text simplification. International Journal of Advanced Computer Science and Applications 4 (2014), 58-70.
[9] Suominen H, Kelly L, Goeuriot L. Scholarly influence of the conference and labs of the evaluation forum eHealth initiative: Review and bibliometric study of the 2012 to 2017 outcomes. JMIR research protocols 7 (2018), e10961.
[10] Vaswani A, Shazeer N, Parmar N et al. Attention is all you need. Advances in Neural Information Processing Systems 30 (2017), 5998-6008.
[11] Wang Y, Liu S, Afzal N et al. A comparison of word embeddings for the biomedical natural language processing. Journal of biomedical informatics 87 (2018), 12-20.
[12] Zhang Y, Chen Q, Yang Z, Lin H, Lu Z, BioWordVec, improving biomedical word embeddings with subword information and MeSH. Scientific data 6 (2019), 1-9.