MEDINFO 2023 — The Future Is Accessible
J. Bichel-Findlay et al. (Eds.)

doi:10.3233/SHTI231053

Data Augmentation with Nearest Neighbor Classifier for Few-Shot Named Entity Recognition

Yao GE[a,1], Mohammed Ali AL-GARADI[b] and Abeed SARKER[a]
[a] *Department of Biomedical Informatics, School of Medicine, Emory University, Atlanta, Georgia*
[b] *Department of Biomedical Informatics, Vanderbilt University Medical Center, Nashville, U.S. state of Tennessee*
ORCiD ID: Yao Ge https://orcid.org/0000-0002-3323-7130, Mohammed Ali Al-Garadi https://orcid.org/0000-0002-6991-2687, Abeed Sarker https://orcid.org/0000-0001-7358-544X

Abstract. Few-shot learning (FSL) is a category of machine learning models that are designed with the intent of solving problems that have small amounts of labeled data available for training. FSL research progress in natural language processing (NLP), particularly within the medical domain, has been notably slow, primarily due to greater difficulties posed by domain-specific characteristics and data sparsity problems. We explored the use of novel methods for text representation and encoding combined with distance-based measures for improving FSL entity detection. In this paper, we propose a data augmentation method to incorporate semantic information from medical texts into the learning process and combine it with a nearest-neighbor classification strategy for predicting entities. Experiments performed on five biomedical text datasets demonstrate that our proposed approach often outperforms other approaches.

Keywords. Natural language processing, machine learning, few-shot learning, biomedical informatics

1. Introduction

Few-shot learning (FSL) is a class of machine learning methods that attempt to learn to execute tasks using small numbers of labeled training examples. Learning from small numbers of instances is challenging for machine learning models, although it is conceptually possible. For many NLP tasks, particularly within the medical domain, the availability of labeled data can be limited (e.g., for rare diseases) [1]. Even when large, labeled datasets that are created for targeted tasks can be difficult or impossible to share if they originate from medical sources due to restrictions associated with data privacy and patient security. The limitations of lexicon-based (e.g., lack of generalizability) and deep learning (e.g., need for large, labeled data) approaches provide motivation for the development of FSL methods that can effectively learn from small, labeled datasets [2].

In this paper, we propose a new method for FSL for named entity recognition (NER) that employs a semantic data augmentation module combined with a nearest neighbor classifier to solve data sparsity problems. We also explore the influences of

[1] Corresponding Author: Yao Ge, email: yao.ge@emory.edu.

different distance metrics. To evaluate our approach, we conducted experiments on five biomedical text datasets. Our results demonstrate that the proposed approach often outperforms other models.

2. Methods

The overarching aim of FSL-based NER systems is to learn from a small number of examples to label names of entities of interest in text documents. In the following subsections, we first introduce the encoding procedure for augmenting semantic information, then we present different distance metrics to explore the influences of methods for calculating similarities.

2.1. Data Augmentation with Nearest Neighbor Classifier

To generate contextual representations for all input tokens, we used a semantically augmented NER model [3] trained on the same source domain as an encoder. The architecture of this data augmentation method combined with the nearest neighbor classifier (DANN) is shown in Figure 1. It follows a popular neural architecture for supervised NER: a BERT-based NER model. For training these models on the source domain, we followed the settings described in Nie et al. [3]. After we obtained the pre-trained embeddings from the BERT-based NER model, for each token in the input sentence, we extracted the top m words that are most similar to the token based on cosine similarities obtained from GloVe.

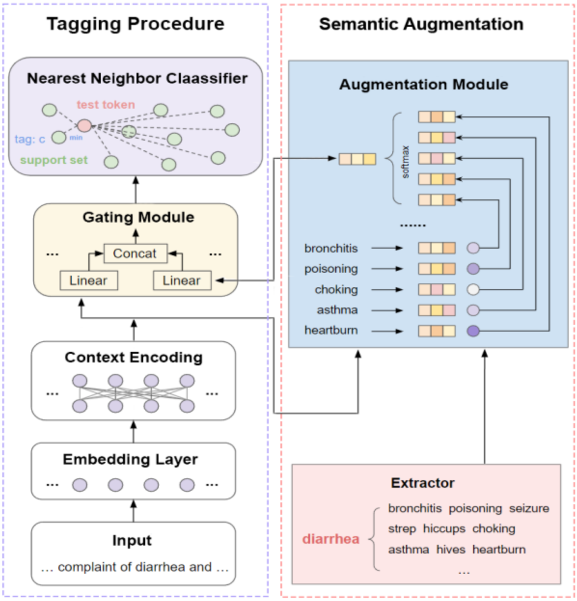

Figure 1. The overall architecture of our proposed model DANN: Data Augmentation combined with Nearest Neighbor classifier. An example sentence is shown, where the augmented semantic information for the word “diarrhea” are also passed for processing through the augmentation module and the gate module. After the gate module, a nearest neighbor classifier computes the similarity scores for tokens.

Since not all extracted words are helpful, the augmentation module is used with an attention mechanism to weigh the semantic information carried by the extracted words depending on their contributions in various contexts. After the semantic augmentation module, a gate module is applied since the contribution of the obtained augmented semantic information to the NER task varies in different contexts. Particularly, we used a gate to control the information flow by:

$$g = \sigma\left(W_1 \cdot h_i + W_2 \cdot v_i + b_g\right) \tag{1}$$

where $\boldsymbol{W_1}$ and $\boldsymbol{W_2}$ are trainable matrices and $\mathbf{b_g}$ is the corresponding bias term.

At inference, given a test example $\mathbf{x} = \{x_t\}_1^T$ and a K-shot entity support set $S = \left\{\left(\mathrm{x}_{\mathrm{n}}^{(\mathrm{sup})}, \mathrm{y}_{\mathrm{n}}^{(\mathrm{sup})}\right)\right\}_{\mathrm{n}=1}^{\mathrm{N}}$ comprising $\boldsymbol{N}$ sentences, we employed a token embedder $\mathrm{f}_\theta(\mathrm{x}) = \hat{\mathrm{x}}$ to obtain contextual representations for all tokens in their respective sentences. Next, different distance metrics are used for computing similarities between tokens in the nearest neighbor classification.

2.2. *Distance Metrics*

We experimented with four methods: squared Euclidean distance, Manhattan distance, infinity norm distance, and 3-norm distance—all commonly used measures of distance.

The Euclidean distance between two points in Euclidean space is the length of a line segment between the two points. In Euclidean space, Euclidean distance (2-norm distance) is usually used to compute the distance between two points. Other distances, based on other norms, are sometimes used instead. The 2-norm distance is the Euclidean distance, the 1-norm distance is called Manhattan distance, because it is the distance a car would drive in a city laid out in square blocks (if there are no one-way streets). The infinity norm distance is also called Chebyshev distance. The p-norm is rarely used for values of p other than 1, 2, and infinity, so in our experiment, we only tried 3-norm.

2.3. *Datasets and Comparison Models*

We conducted our experiments on five medical text datasets, which included MIMIC III (Medical Information Mart for Intensive Care) dataset [4] and four additional datasets from different shared tasks: (i) the N2C2 2018 shared task track 2 [5]; (ii) the I2B2 2014 shared task [6]; (iii) the BioNLP 2016 shared task [7]; and (iv) the SMM4H (Social Media Mining for Health Applications) 2021 shared task 1b [8].

We used the NNShot model [9] as our comparison system since it was the top FSL-based NER model from in our recent benchmarking work [10]. The backbone of NNShot is the Nearest Neighbor Classifier, and it uses Euclidean distance to compute a similarity score between a text segment in the test set and all tokens in the support set.

3. Results

Table 1 shows the F_1-scores of our proposed model DANN with different distance metrics on five medical datasets. From the table, we see that our experimental results outperform the comparison model on most tasks, and our model's performance is the best one on the I2B2 2014 dataset. On other datasets with different characteristics, the performance of our model is also close to that of the benchmark models, except for the N2C2 2018 dataset. This is probably because the size of the N2C2 2018 dataset is the largest amongst those included in this study, and hence, the number of occurrences of the "O" entity type is much higher than in other datasets. This perhaps leads to the introduction of more noise. The table also shows that for the social media dataset (SMM4H 2021), none of the models can make accurate predictions with few samples. Previous research has shown that social media based medical NLP datasets are more difficult to obtain high performances as social media data has specific characteristics that make NLP challenging, such as the presence of misspellings and colloquial expressions.

Table 1. F$_1$-scores of our proposed DANN models with four different distance metrics on five medical datasets compared with NNShot. The best performance of our models in 5-shot settings and 1-shot settings has been highlighted in bold and underlined.

Models	Training Size	N2C2 2018	I2B2 2014	MIMIC III	BioNLP 2016	SMM4H 2021
NNShot (few-shot model)	5-shot	**25.29**	19.73	19.51	**28.88**	0.00
	1-shot	16.70	16.35	15.37	6.42	0.00
DANN + Squared Euclidean distance	5-shot	0.21	25.18	19.34	24.02	0.00
	1-shot	2.25	11.95	9.55	22.68	0.00
DANN + Manhattan distance	5-shot	0.16	**27.29**	**19.68**	24.21	0.00
	1-shot	1.95	10.81	5.38	22.97	0.00
DANN + Infinity norm distance	5-shot	0.13	16.99	16.99	23.97	0.00
	1-shot	1.68	16.99	9.70	22.84	0.00
DANN + 3-norm distance	5-shot	0.14	22.87	18.98	23.93	0.00
	1-shot	2.25	13.92	10.54	23.16	0.00

For the same settings of the DANN model, we can horizontally compare five methods for calculating similarity. From Table 1, we see that Manhattan distance performs relatively well in the 5-shot setting, slightly outperforming other distance metrics on three datasets. Meanwhile, the best-performing distance method in the 1-shot setting is the least frequently used 3-norm distance metric, which performs the best on three datasets.

4. Discussion

The essence of our method is to change the input from the simple embedding generated by the BERT model to a more complex generation method. Specifically, we use a data augmentation module, based on the nearest neighbor classifier. In this experiment, we used a BERT-based NER model to generate encodings, and used GloVe to select words that are similar to the input tokens. These are both good mechanisms for obtaining word vectors, but they have no unique advantages for medical data. Therefore, we also experimented with more domain-specific models such as BioBERT and ClinicalBERT to try to obtain the representations of tokens which are learned from medical or scientific data. These experiments, however, did not produce results better than other approaches.

5. Conclusions

FSL approaches have substantial promise for NLP in the medical domain as many medical datasets naturally have low numbers of annotated instances. We focused on applying a semantic augmentation module with an attention mechanism for leveraging the semantic information from the extracted similar words, then employed nearest neighbor learning at the inference stage. The results show that our model mostly outperforms other methods on one dataset. In the future, we will explore the possibility of incorporating domain knowledge from the Unified Medical Language System (UMLS), and potential opportunities for multi-modal data augmentation. We will also experiment with social media specific pretrained models, such as BERTweet for improving performance on social media datasets, such as SMM4H.

Acknowledgements

This research was supported by the National Institute on Drug Abuse (R01DA057599). The content is solely the responsibility of the authors.

References

[1] Hofer M, Kormilitzin A, Goldberg P, Nevado-Holgado A. Few-shot learning for named entity recognition in medical text. arXiv preprint arXiv:1811.05468. 2018 Nov 13, doi:10.48550/arXiv.1811.05468.

[2] Li W, Wang L, Xu J, Huo J, Gao Y, Luo J. Revisiting local descriptor-based image-to-class measure for few-shot learning. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition, 2019. p. 7260-7268, doi: 10.48550/arXiv.1811.05468.

[3] Nie Y, Tian Y, Wan X, Song Y, Dai B. Named entity recognition for social media texts with semantic augmentation. 2020. arXiv preprint arXiv:2010.15458. doi:10.48550/arXiv.2010.15458.

[4] Johnson AE, Pollard TJ, Shen L, Lehman LW, Feng M, Ghassemi M, Moody B, Szolovits P, Anthony Celi L, Mark RG. MIMIC-III, a freely accessible critical care database. Sci data. 2016 May;3(1):1-9, doi: 10.1038/sdata.2016.35.

[5] Henry S, Buchan K, Filannino M, Stubbs A, Uzuner O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. J Am Med Inform Assoc. 2020 Jan;27(1):3-12, doi: 10.1093/jamia/ocz166.

[6] Stubbs A, Uzuner Ö. Annotating longitudinal clinical narratives for de-identification: The 2014 i2b2/UTHealth corpus. J Biomed Inform. 2015 Dec. 58: S20-9, doi: 10.1016/j.jbi.2015.07.020.

[7] Chaix E, Dubreucq B, Fatihi A, Valsamou D, Bossy R, Ba M, Deléger L, Zweigenbaum P, Bessieres P, Lepiniec L, Nédellec C. Overview of the regulatory network of plant seed development (seedev) task at the bionlp shared task 2016. In BioNLP Shared Task; 2016 Aug. p. 113. The Association for Computational Linguistics.

[8] Weissenbacher D, Sarker A, Magge A, Daughton A, O'Connor K, Paul M, Gonzalez G. Overview of the fourth social media mining for health (SMM4H) shared tasks at ACL 2019. Proceedings of the fourth social media mining for health applications (SMM4H) workshop and shared task; 2019. p. 21-30, doi: 10.18653/v1/W19-3203.

[9] Yang Y, Katiyar A. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. 2020 Oct. arXiv preprint arXiv:2010.02405, doi: 10.48550/arXiv.2010.02405.

[10] Ge Y, Guo Y, Yang YC, Al–Garadi MA, Sarker A. A comparison of few-shot and traditional named entity recognition models for medical text, Proceedings of the 10th IEEE International Conference on Healthcare Informatics; 2022 Jun. p. 84-89, doi: 10.1109/ICHI2022.2022.00024.