A self-supervised domain-independent Named Entity Recognition using local similarity

Keerthi Vasan S A
Faculty of Engineering and Technology
Sri Ramachandra Institute of Higher Education and Research
Chennai

Uma Ranjan
Faculty of Engineering and Technology
Sri Ramachandra Institute of Higher Education and Research
Chennai

Abstract

Out-of-vocabulary words can be challenging for NER systems. We introduce a self-supervised system for Named Entity Recognition based on a few-shot annotated examples provided by experts. Subsequently, the rest of the words are tagged using the closest similarity match between the word embeddings of each category, generated in the same context as the original annotations. Additionally, we use a dual-threshold scheme to improve the robustness of the method. Our results show that this method outperforms current state-of-the-art methods in both accuracy and generalization.

1 Introduction

Named Entity Recognition (NER) is a sub-task of information extraction that involves identifying and categorizing entities in text into predefined categories, such as names of people, organizations, locations, medical terms, or quantities. NER is a critical tool in natural language processing (NLP), particularly in specialized domains such as medical literature and manufacturing.

In medical literature, NER can enable the use of Electronic Health Record (EHR) systems to retrieve previous case studies and support clinical decision-making. It also plays a vital role in telemedicine, where self-reported symptoms or preliminary investigation reports must be matched to specialists for remote diagnoses.

However, NER in specialized domains, particularly in medical documents, presents unique challenges. One primary issue is the lack of sufficient datasets. Many clinical documents, such as case summaries or clinical reports, cannot be made public due to patient privacy concerns, limiting the corpus available for training NER models. Additionally, medical texts often contain specialized vocabulary, abbreviations, and terminologies absent from general-purpose NLP datasets, leading to poor generalization by standard NER systems (Tsai et al., 2006).

Manual annotation of medical texts is labor-intensive and requires domain expertise, making it an expensive and slow process. Furthermore, the non-uniform nature of medical terminology across sub-disciplines adds complexity. For example, vocabulary in oncology differs significantly from that in cardiology. Consequently, even annotated datasets may not cover the full spectrum of terms across the medical field.

Traditionally, supervised learning methods have been used for NER recognition (Anandika and Mishra, 2019; Dash et al., 2024), which need a large corpus of annotated data. Even with these, the problem remains challenging due to the fact that there is no way to predict an unseen entity unless its context is included. Hence, this led to the use of context-based representations such as LSTM and its variants (Xu et al., 2018), including those with attention mechanisms (Shao et al., 2021). Medical corpus also consists of a wide variety of documents such as text books, case studies, self-reported symptoms by patients etc. where the contexts of the words varies widely. Further, the different sub-domains of medical literature such as cardiology, neurology etc., which use use very different words. The diversity of sub-domains, such as cardiology and neurology, further complicates the issue. Each modality must often be treated as a separate domain, necessitating different approaches.

2 Related Work

Named Entity Recognition (NER) in specialized domains has been extensively studied using various approaches. However, these methods face limitations, particularly in handling low-resource scenarios where labeled data is scarce.

Semi-supervised models are often employed to address the lack of high quality annotated data. These models generate pseudo-labels from limited

high quality labeled data, and subsequently used these pseudo-labels in the same manner as the high quality labels. This results in erroneous data and requires some kind of a post-filtering (Li and *et al*, 2020) or constructing a model based on the original high quality annotations (Liu et al., 2021). This has prompted the development of several few-shot learning techniques for NER in such domains (Fritzler et al., 2019).

Another alternative is transfer learning, where a model is pre-trained on a large, general-purpose dataset and then fine-tuned on a smaller, domain-specific dataset. Pre-trained models such as BERT (Bidirectional Encoder Representations from Transformers) and its variants (e.g., BioBERT, ClinicalBERT) belong to this class of methods.

A third approach to address this problem is the use of distant supervision models, where external knowledge bases (such as medical ontologies or databases) are used to automatically label the data. This also falls in the category of rule-based systems, and needs access to specialized ontologies which are expensive to create.

Character and word embeddings are another way to capture local context (Yang and Katiyar, 2020), and are a powerful way to capture usage of a word. However, embeddings trained on a source domain rarely do well on a target domain. Transfer learning methods, which have been very useful for classification methods, perform poorly on IR methods. Further, NER predictions based on word embeddings are typically noisy, and need further refinement. (Peng et al., 2021) uses Reinforcement Learning to refine and predict the instance.

In this work, we bypass the problem of source-target mismatch by estimating word embeddings directly from the target dataset. This ensures that the context of the words corresponds to the document currently analyzed. Since we create embeddings from a single document, the process is extremely fast, compared to creating a word embedding from a large corpus.

Creating embeddings from a small dataset can result in noisy entities. We overcome this by introducing a mechanism by which comparisons are done with all the entities of a label. We also introduce a dual threshold mechanism for computing similarity. This increases the robustness of our method, and achieves much higher accuracy than state of the art methods. This also enables our approach to be used for any domain without the need for transfer learning.

Since we generate embeddings as well as labels solely from a target document, our method can also address the heterogeneity of documents (text books, journal articles, case reports, short case summaries) that are written in very different styles, which would otherwise need specific models for each style.

3 Proposed Method

In this work, we propose a few-shot self-supervised approach where the medical expert only needs to annotate a small sample of words within a document. A Word2Vec model is trained on the document, and vector representations generated for all the nouns in the document. The annotated labels serve as a reference, and their embeddings are compared with that those of the unannotated words in the same document. The rationale behind this approach is that Out-of-vocabulary words derive the most relevant context from the document itself. By leveraging the style and context patterns within the document, the model can infer the meanings of new or rare terms based on their similarity to annotated words.

Once the annotated words in each category are embedded, we compute the cosine similarity between the vectors of the unknown words and those that have already been tagged using manual annotation. Cosine similarity measures the angle between two vectors in a high-dimensional space, and is widely used in NLP tasks to measure the semantic similarity between words. Words that are more similar (i.e., those with higher cosine similarity scores) are likely to belong to the same category.

We calculate the cosine similarity between each untagged word and the words that have already been assigned a category. We then assign a category to each untagged word based on a combination of thresholds as follows :

Let $\mathcal{L}_1, \mathcal{L}_2, \ldots, \mathcal{L}_n$ be the set of entity labels, each of which is a collection of d-dimensional word embeddings $\{w_j^i\}_{j=1}^{N_i}$.

Let $\mathbf{u} \in \mathbf{R}^d$ be the new word to be assigned a label.

Let $\text{sim}(\mathbf{u}, \mathbf{v})$ be the cosine similarity between two embedding vectors $\mathbf{u}$ and $\mathbf{v}$

Two thresholds are defined, a minimum similarity threshold τ_{min} and a maximum similarity threshold τ_{max}.

For a target word $\mathbf{u}$, a label $\mathcal{L}_i$ belongs to the

set $\mathcal{C}(\mathbf{u})$ of candidate labels for $\mathbf{u}$ if and only if it satisfies both the following conditions:

- *Minimum similarity condition:*

$$\min_{w_j^i \in L_i} \text{sim}(u, w_j^i) > \tau_{min}$$

- *Maximum similarity condition:*

$$\max_{w_j^i \in L_i} \text{sim}(u, w_j^i) > \tau_{max}$$

The label corresponding to the target word $\mathbf{u}$ is defined as $\arg\max_i\{max_j \text{sim}(\mathbf{u}, w_j^i\}$, where $i \in \mathcal{C}(\mathbf{u})$. If $\mathcal{C}(\mathbf{u}) = \Phi$, the word is not assigned a label. The methodology is illustrated in Figure 1

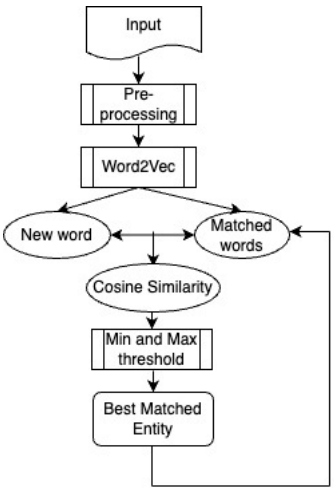

Figure 1: Proposed Method

The thresholds are hyperparameters and ensure the robustness of similarity between annotated labels and target word. A min threshold requires that the embedding of the new word exceed the minimum similarity with each of the annotated words in the category. A max threshold stipulates that the new word exceed the max threshold for at least one word in the category. The use of the min threshold removes the entities which were wrongly assigned to an category, and improves the false positive rate. By adjusting the thresholds, the system can be made more or less conservative in its predictions, allowing for a balance between precision and recall. The thresholds are currently chosen by trial and error.

The advantages of this approach are the following

- Reduced need for extensive annotation, and preservation of the high quality annotations. Since only a small number of words in each document need to be annotated, this method significantly reduces the burden on medical experts. This also resonates with the workflow of physicians, where they are used to annotating only a few examples. Further, we do not combine generated labels with actual annotated labels, thereby preserving the quality of the annotations.

- local contextual similarity by using word embeddings generated from the same document. This enables the model to handle a wide variety of document lengths.

- Robustness : A combination of two thresholds are used to ensure the balance between precision and recall.

4 Results

We evaluated our method on different levels of manual annotation (80%, 50%, and 25%) and on two different lengths of texts - long texts consisting of 70-80 sentences and short texts consisting of 30-40 sentences. Since our method generates embeddings from the same document, we investigate the effect of the length of the document. The two lengths we have chosen correspond to the typical lengths present in descriptions obtained in text books or protocol documents (long texts), while the shorter texts correspond to case reports or case summaries.

We also studied the performance on two different sub-domains of medical field - cardiology and ophthalmology. The different domains are considered here only for human evaluation. The actual domains are not relevant for the performance of the algorithm.

4.1 Cardiology

We took two sample texts in cardiology - one short and one long. We then used different percentages of annotation. For example, 80% annotation means 80% of the entities in the text are annotated, and the task during testing is to check how many of the entities are labeled correctly. The F1 score for these cases is presented in Table 1 (short text) and Table 2 (long text). The results are also compared with the Baseline Bloom-CRF model implemented in Spacy 2.0.

4.2 Ophthalmology

We evaluate the results of the algorithm on two types of texts in typical ophthalmology settings -

Entity	80%	50%	25%	25% (Base)
Risk Factor	0.75	0.8	0.62	0.33
Anatomy	0.75	0.4	0.56	0.33
Procedure	0.75	0.6	0.69	0.33
Condition	0.75	0.6	0.62	0.46

Table 1: F1 scores on Cardiology short text

Entity	80%	50%	25%	25% (Base)
Risk Factor	0.75	0.9	0.81	0.46
Anatomy	0.75	0.5	0.5	0.33
Procedure	0.75	0.6	0.5	0.33
Condition	0.5	0.7	0.5	0.66

Table 2: F1 scores on Cardiology Long text

short and long, each with 80%, 50% and 25% annotations. Tables 3 and 4 indicate the results on short text and long text respectively. We also compare these with the Baseline Bloom-CRF model implemented in Spacy 2.0.

4.3 Ablation study

We perform an ablation study to determine which component causes the improved performance. Towards this, we use the manual annotations to train a state-of-the-art system using Space's 2.0. The algorithm fundamentally creates a Bloom embedding of the annotated words, and uses Conditional Random Fields to predict the entity tags. We used this component to study the effect of the manual annotations we supplied, to see if these annotations would work just as well in any other system. We found that the Bloom-CRF model of Spacy 2.0 was not able to generalize from these manual annotations. In fact, it even missed some of the tags that were manually annotated. On the other hand, our system captured all of the seen entities, and many of the unseen entities as well. Sample results are presented in Figures 2 and 3.

Entity	80%	50%	25%	25% (Base)
Anat. Str.	1.0	0.83	0.75	0.33
Procedure	0.5	0.75	0.67	0.4
Symptom	0.5	0.75	0.67	0.4
Condition	1.0	0.67	0.5	0.33

Table 3: F1 scores on Ophthalmology Short text

Entity	80%	50%	25%	25% (Base)
Anat. Str.	0.75	0.7	0.88	0.3
Procedure	0.75	0.8	0.75	0.66
Symptom	0.75	0.7	0.5	0.46
Condition	1.0	0.6	0.44	0.33

Table 4: F1 scores on Ophthalmology Long text

This shows that our method of computing similarity is superior to the traditional supervised training method on few shot examples.

Figure 2: Results on cardiology text

Figure 3: Results on Ophthalmology text

We next implemented a single threshold based system which checked for the closest annotated entity, subject to a lower threshold. This method yielded a high precision for non-entities, and a high recall for entities but yielded a precision of 0.33 for non-entities and a precision of 0.6 for entities, suggesting that the method was missing tags on unseen entities. Even with a single threshold, the performance was better than the previous results reported on closest embedding approach (Yang and Katiyar, 2020).

5 Conclusion

We present a self-supervised few-shot domain-independent NER model which uses the local context present within the document along with a limited number of high quality annotations. To our knowledge, this is the first work that does not require model retraining on an external corpus, or specific domain-based rules. This is also, to our knowledge, the first work which focuses on sub-domains of a medical field for NER.

We achieve very good accuracy on both short and long documents, and demonstrate its use in two different sub-domains without re-training.

Further, we present results on documents written in very different styles, which could correspond to the different applications and workflows in the medical field. While the procedural documents may be used to enhance the use of EHR and other digital health initiatives, short texts written in the style of self-reported symptoms can be used to assign the right specialist in telemedicine applications.

Our method further overcomes the need for external domain-specific re-training or corpuses which need to match the style of the current text. It can also be used in situations where recent up-to-date corpuses are hard to create.

Limitations

One of the main limitations of the paper is the variable performance on texts. Since the method uses the local context present within the text, its performance depends on the distribution of words of each entity present in the text. This also explains why the performance does not linearly scale with the amount of manual annotation for all entities. It has been observed that a text must have at least 3-4 entities of a category to perform well.

References

Amrita Anandika and Smita Prava Mishra. 2019. A study on machine learning approaches for named entity recognition. In *2019 International Conference on Applied Machine Learning (ICAML)*, pages 153–159.

Adyasha Dash, Subhashree Darshana, Devendra Kumar Yadav, and Vinti Gupta. 2024. A clinical named entity recognition model using pretrained word embedding and deep neural networks. *Decision Analytics Journal*, 10:100426.

Alexander Fritzler, Varvara Logacheva, and Maksim Kretov. 2019. Few-shot classification in named entity recognition task. In *Proceedings of the 34th ACM/SIGAPP Symposium on Applied Computing*, SAC '19, page 993–1000, New York, NY, USA. Association for Computing Machinery.

Zz Li and Dw Feng *et al*. 2020. Learning to select pseudo labels: a semi supervised method for named entity recognition. *Front. Inform. Tech. Electron Eng*, 21:903–916.

Yuanyuan Liu, Xiang Li, Jiaxin Shi, Lei Zhang, and Juanzi Li. 2021. Named entity recognition using a semi-supervised model based on bert and bootstrapping. In *Knowledge Graph and Semantic Computing: Knowledge Graph and Cognitive Intelligence*, pages 54–63, Singapore. Springer Singapore.

Shi Peng, Yong Zhang, Zhengyun Wang, Dingkang Gao, Feng Xiong, and Haoyang Zuo. 2021. Named entity recognition using negative sampling and reinforcement learning. In *2021 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 714–719.

Yinan Shao, Jerry Chun-Wei Lin, Gautam Srivastava, Alireza Jolfaei, Dongdong Guo, and Yi Hu. 2021. Self-attention-based conditional random fields latent variables model for sequence labeling. *Pattern Recognition Letters*, 145:157–164.

Richard Tzong-Han Tsai, Cheng-Lung Sung, Hongjie Dai, Hsieh-Chuan Hung, Ting-Yi Sung, and Wen-Lian Hsu. 2006. Nerbio: Using selected word conjunctions, term normalization, and global patterns to improve biomedical named entity recognition. *BMC bioinformatics*, 7 Suppl 5:S11.

Kai Xu, Zhanfan Zhou, Tianyong Hao, and Wenyin Liu. 2018. A bidirectional lstm and conditional random fields approach to medical named entity recognition. In *Proceedings of the International Conference on Advanced Intelligent Systems and Informatics 2017*, pages 355–365, Cham. Springer International Publishing.

Yi Yang and Arzoo Katiyar. 2020. Simple and effective few-shot named entity recognition with structured nearest neighbor learning. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 6365–6375, Online. Association for Computational Linguistics.