Named Entity Recognition in Scientific Texts Using Long Short-Term Memory Networks*

Ana Laura Lezama-Sánchez[1,*,†], Mireya Tovar Vidal[1,†]

[1]Benemerita Universidad Autonoma de Puebla, Av., 14 Sur Blvd., Puebla, ZIP Code: 72592, Mexico

Abstract

In this work, we present an approach for Named Entity Recognition in scientific texts using Long Short-Term Memory Neural Networks. We began with a text preprocessing step to clean and normalize the data, followed by the construction and training of a neural network model. The proposed model is evaluated on test data from GenoVarDis task of IberLEF 2024, and the results are assessed using a competency evaluation script.

Keywords

Deep learning, knowledge graphs, natural language processing

1. Introduction

The task of Named Entity Recognition (NER) in the medical field is crucial for advancing the understanding and management of clinical information. As the amount of available medical data grows exponentially, it becomes increasingly challenging for healthcare professionals to efficiently extract relevant information. This is where Natural Language Processing (NLP) techniques, specifically NER, come into play [1].

NER in the medical domain involves identifying and classifying specific entities such as disease names, medical procedures, medications, and patient names in unstructured medical texts. This process is fundamental for a variety of applications, including clinical information extraction for research, improving accuracy in diagnosis and treatment, and automating administrative tasks in the healthcare domain [2].

Traditional NER techniques in NLP relied on heuristic rules and predefined dictionaries, limiting their ability to capture the complexity and variability of medical language. However, with the advent of deep learning, especially architectures like Recurrent Neural Networks (RNNs), Long Short-Term Memory (LSTM), and Convolutional Neural Networks (CNNs), significant improvements in NER performance in medical texts have been achieved [3].

One of the main advantages of the deep learning approach is its ability to automatically learn representations of complex features from raw data. This means that deep learning-based NER models can capture patterns in medical language that may go unnoticed by traditional methods. Additionally, these models can easily adapt to different medical domains and specific clinical languages by fine-tuning model parameters [4].

To further improve NER performance in the medical domain, high-quality annotated datasets reflecting the variability and complexity of clinical language are crucial. Furthermore, the development of specific evaluation tools for medical NER, considering the unique characteristics of clinical texts, is essential for accurately measuring model performance and facilitating comparison between different approaches [5].

In this article, we propose applying an LSTM network for Named Entity Recognition, using datasets provided by the GenoVarDis task: NER in genomic variants and related diseases from iberLEF 2024 [6, 7].

IberLEF 2024, September 2024, Valladolid, Spain

*Corresponding author.

†These authors contributed equally.

yumita1102@gmail.com (A. L. Lezama-Sánchez); mireya.tovar@correo.buap.mx (M. T. Vidal)

https://ontologica.cs.buap.mx/ (M. T. Vidal)

0000-0002-2505-5150 (A. L. Lezama-Sánchez); 0000-0002-9086-7446 (M. T. Vidal)

Table 1
Datasets

Dataset	Documents
Train	417
Test	136
Dev	70

The reported results are a precision of 0.60%, an F1 score of 0.30%, and a recall of 0.20%, suggesting that the model has moderate performance in the task of recognizing named entities in this domain. Specifically, the low F1 score, and recall indicate that the model may be having difficulties capturing all relevant entities in the test data.

The paper structure is as follows: methodology is described in Section 2, results are presented in Section 3 and the conclusions and future work are outlined in Section 4.

2. Methodology

The implemented methodology begins with preprocessing the datasets. This includes tasks such as removing accents, standardizing text to lowercase, eliminating punctuation, special characters, numbers, and stopwords. Subsequently, we identified eight specific categories in the labeled text: Gene, NucleotideChange-BaseChange, DNAMutation, Disease, OtherMutation, Transcript, SNP and DNAAllele. Based on these categories, the model architecture was defined as a sequential neural network.

This network incorporates embedding layers to densely represent the text, followed by bidirectional LSTM layers to capture contextual information in both directions of the text. Additional dense layers with ReLU activation are included to learn nonlinear representations, along with a softmax activation output layer for multiclass classification of named entities.

The model's architecture includes two LSTM layers, each configured with distinct parameters. The first LSTM layer is bidirectional with 64 units and return_sequences=True, enabling the capture of contextual information in both directions of the temporal sequence. The second LSTM layer, also bidirectional but with 32 units, complements this structure for further refinement of learning. Both LSTM layers are encapsulated within the Bidirectional wrapper, facilitating effective bidirectional connections in sequence modeling. This configuration was selected to explore and compare the impact of different memory capacities and sequential learning on model performance.

After defining the architecture, the model is compiled using the sparse_categorical_crossentropy loss function, suitable for multiclass classification, and the Adam optimizer to optimize neural network weights. Following compilation, the model is trained on the training data. During training, the model adjusts its internal parameters to minimize the loss function and enhance accuracy in predicting input data.

In conclusion, this approach enables analysis and classification of specific entities in the labeled corpus, contributing to the discovery of pertinent information in text pertaining to genetics and molecular biology.

3. Results

The dataset comprises the translation and manual curation of the documents from the tmVar32 annotations (PubMed3 abstracts) with their associated diseases and symptoms, which are presented in Table 1.

When training our LSTM model on the entire training dataset and subsequently testing it on the test dataset provided by Agüero-Torales et al., 2024 [6], we achieved a precision score of 0.60, placing us sixth among all participants. However, our F_1 score was 0.30 and recall was 0.20, placing us last

Table 2
Results obtained by each participant

User	F_1	Precision	Recall
Ander.martinez	0.82 (1)	0.82 (1)	0.81 (1)
VictorMov	0.79 (2)	0.79 (2)	0.79 (2)
ELiRF-VRAIN	0.73 (3)	0.77 (3)	0.69 (3)
Milímetro98	0.54 (4)	0.61 (5)	0.49 (4)
Orlandxrf	0.53 (5)	0.73 (4)	0.41 (6)
GuillemGSubies	0.42 (6)	0.43 (8)	0.42 (5)
mmaguero	0.31 (7)	0.59 (7)	0.21 (7)
Antares-Amazel	**0.30 (8)**	**0.60 (6)**	**0.20 (8)**

compared to all participants. During the tests conducted, initially with 10 epochs, we concluded that adding 40 epochs and some additional layers to the model resulted in lower precision compared to the results obtained in the development phase, where we achieved a precision of 0.64, an F_1 score of 0.33, and a recall of 0.22.

In Table 2, the results obtained by each participant in the task of iberLEF 2024 are shown. Our results are marked in bold with the nickname Antares-Amazel.

4. Conclusions and Future Work

In this article, the proposal was to use an LSTM network for Named Entity Recognition (NER) in texts related to genomic variants and diseases, utilizing data from the GenoVarDis challenge of iberLEF 2024 [6]. The reported results show a precision of 0.60%, an F_1 score of 0.30%, and a recall of 0.20%, indicating a moderate performance of the model in this specific task.

However, the low F_1 score, and recall suggest that the model faces difficulties in capturing all relevant entities in the test data, reflecting the inherent complexity and linguistic diversity in this specialized field.

As future work, it is suggested to improve the model architecture using advanced approaches that consider contextual and semantic relationships in biological texts, enrich the dataset with more labeled and diverse examples, apply data augmentation techniques to enhance model training, and explore hybrid or ensemble models to improve the precision and generalization capability of the NER system in the domain of genomic variants and related diseases.

5. Acknowledgments

This work was carried out at the Ontological Engineering Laboratory at the Benemerita Universidad Autonoma de Puebla, under the guidance of Ph.D. Mireya Tovar Vidal, who provided all necessary resources.

References

[1] X. Zeng, Y. Luo, Machine learning and natural language processing for genomic variants information extraction, https://www.mdpi.com/2073-4425/11/3/323 11 (2020).

[2] S. Locke, A. Bashall, S. Al-Adely, J. Moore, A. Wilson, G. B. Kitchen, Natural language processing in medicine: a review, Trends in Anaesthesia and Critical Care 38 (2021) 4–9.

[3] S. Naseer, M. M. Ghafoor, S. bin Khalid Alvi, A. Kiran, S. U. Rahmand, G. Murtazae, G. Murtaza, Named entity recognition (ner) in nlp techniques, tools accuracy and performance., Pakistan Journal of Multidisciplinary Research 2 (2021) 293–308.

[4] V. Moonsamy, X. Zhang, J. Baguley, Convolutional neural networks for named entity recognition in clinical text., Proceedings of the 2018 IEEE International Conference on Bioinformatics and Biomedicine (BIBM), 548-551. https://ieeexplore.ieee.org/document/8621392 (2018).
[5] G. Lample, M. Ballesteros, S. Subramanian, K. Kawakami, C. Dyer, Neural architectures for named entity recognition, arXiv preprint arXiv:1603.01360 (2016).
[6] M. M. Agüero-Torales, C. Rodríguez Abellán, M. Carcajona Mata, J. I. Díaz Hernández, M. Solís López, A. Miranda-Escalada, S. López-Alvárez, J. Mira Prats, C. Castaño Moraga, D. Vilares, L. Chiruzzo, Overview of genovardis at iberlef 2024: Ner of genomic variants and related diseases in spanish, Procesamiento del Lenguaje Natural 73 (2024).
[7] L. Chiruzzo, S. M. Jiménez-Zafra, F. Rangel, Overview of IberLEF 2024: Natural Language Processing Challenges for Spanish and other Iberian Languages, in: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2024), co-located with the 40th Conference of the Spanish Society for Natural Language Processing (SEPLN 2024), CEUR-WS.org, 2024.