NLP-CIC-WFU at SocialDisNER: Disease Mention Extraction in Spanish Tweets Using Transfer Learning and Search by Propagation

Antonio Tamayo
Instituto Politécnico Nacional
Centro de Investigación en Computación
Av. Juan de Dios Batiz,
s/n, 07320, Mexico City, Mexico
atamayoh2019@cic.ipn.mx

Diego Burgos
Wake Forest University
1834 Wake Forest Road,
Winston-Salem,
NC 27109,
Winston Salem, USA
burgosda@wfu.edu

Alexander Gelbukh
Instituto Politécnico Nacional
Centro de Investigación en Computación
Av. Juan de Dios Batiz,
s/n, 07320, Mexico City, Mexico
gelbukh@cic.ipn.mx

Abstract

Named entity recognition (e.g., disease mention extraction) is one of the most relevant tasks for data mining in the medical field. Although it is a well-known challenge, the bulk of the efforts to tackle this task have been made using clinical texts commonly written in English. In this work, we present our contribution to the SocialDisNER competition, which consists of a transfer learning approach to extracting disease mentions in a corpus from Twitter written in Spanish. We fine-tuned a model based on mBERT and applied post-processing using regular expressions to propagate the entities identified by the model and enhance disease mention extraction. Our system achieved a competitive strict F1 of 0.851 on the testing data set.

1 Motivation

Although there are several works for disease mention extraction, the bulk of them has been carried out for clinical texts written in English (Eftimov et al., 2017; Patra and Saha, 2013; Peng et al., 2019; Sachan et al., 2018; Lee et al., 2020; Akhtyamova, 2020; Alsentzer et al., 2019; Yasunaga et al., 2022; Gu et al., 2021). In this work, we present our contribution to the SocialDisNER competition (Gasco et al., 2022) at the SMM4H workshop, task 10 (Weissenbacher et al., 2022). Our system is focused on disease mention extraction from Twitter messages in Spanish. The nature of the texts written in this social network presents new challenges to the disease extraction task because misspellings are frequent (Magumba et al., 2018; Magge et al., 2021). Additionally, many disease mentions can be subsumed in hashtags, urls, or user names (Magumba et al., 2018), which, together with the above, makes it more difficult to identify entities than in common clinical texts as shown in Xiong et al. (2020); García-Pablos et al. (2020); Wang et al. (2019).

2 System description

In this work, we present a transfer learning approach using the model proposed by Tamayo et al. (2022), which is a version of multilingual BERT (Devlin et al., 2019) fine-tuned for disease mention extraction from clinical texts and we apply post-processing rules to extract diseases mentioned in a corpus of tweets in Spanish. Our system tackles the problem in three steps, namely, pre-processing, transfer learning, and post-processing. Below we describe each of them.

2.1 Pre-processing

To implement the fine-tuning process, the BIO scheme (Begin, Inside, Outside) (Ramshaw and Marcus, 1995) was used. Since the dataset provided by SocialDisNER is formatted in a different way, pre-processing was needed to take it to the BIO scheme. We used the disease mentions in the provided structured dataset as a reference to annotate disease mentions in each tweet with their corresponding labels in the BIO scheme. Tokenization was carried out using SpaCy (Honnibal and Montani, 2017) instead of a NER dedicated library such as SciSpacy (Neumann et al., 2019) because the former works for Spanish.

2.2 Transfer learning

We tackled disease mention extraction as a sequence labeling problem using the whole tweet as input, and the labels mentioned above as output. We randomly split partitions of the training dataset into training (75%) and validation (25%) sets. This partition was done iteratively five times with random seeds. Additionally, we carried out a hyperparameter tuning searching for the best model's configuration using a grid search for the epochs (3, 5, 7) and the learning rate (5e-03, 5e-05, 5e-07). 7 epochs and a learning rate of 5e-05 yielded the best results. With regard to the rest of hyperparameters, default values were kept. For this process,

Proceedings of the 29th International Conference on Computational Linguistic, pages 19–22
October 12–17, 2022.

we used a transformer library and the model available at Hugging Face[1]. Google Colab Pro with a GPU Tesla P100 with 27.3 gigabytes of available RAM was used to run all the experiments. The data we used for our training process together with the source code to replicate this work are available at a GitHub repository[2].

2.3 Post-processing plus search by propagation

Post-processing was carried out through a custom Python script to clean up and format the output as follows: 1) Because mBERT works with a subword tokenization system, we decoded the output that contained subwords. 2) We concatenated all the named entities detected by the model one after the other. This means that if the model detected a named entity whose final character position (or final character position plus one) concurred with the first position of the next named entity detected, our system considered that these two entities were part of one single entity. This was necessary because the model extracts parts of some entities separately. 3) We also applied simple but effective post-processing based on some orthographic and grammatical rules which are detailed in Table 1. 4) Under the assumption that SocialDisNER participants were required to extract all the mentions of a disease mention occurring in a tweet, we used the entities extracted by the model to identify and extract any repetitions of said entities in the same document. In order to retrieve misspelled mentions or mentions subsumed by hashtags, urls, or user names, we carried out a search by propagation applying the following steps: a) lowercase both the entity identified by the model and the tweet, b) concatenate multi-word entities, c) delete accents, and d) search entity occurrences throughout the tweet. Lastly, since we work with the BIO scheme, the last post-processing step consisted of decoding the predictions to put them in the data format required by SocialDisNER.

3 Results and error analysis

The results achieved by our model on the validation dataset can be seen in Table 2.

Likewise, for future reference and improvement, we present the following brief error analysis. First, our system extracts false positives. They are meaningful entities, but they are not in the gold standard (e.g., EFyC, *formación diabetológica*). Second, the model truncates some entities (e.g., *problemas de apre* insted of *problemas de aprendizaje*, *respiratorias crónicas* instead of *Enfermedades respiratorias crónicas*). The first example of the latter type of error is caused by the nature of the mBERT model which works with subwords tokenization. Finally, we consider that there are some errors resulting from an incorrect tagging of the dataset (e.g., our model extracts the entity *cáncer* but in the gold standard appears *niñas y niños que hay hoy con cáncer*).

If the disease mention detected . . .	**. . . then apply this rule**
1. Starts with punctuation mark	1. Delete the match and adjust the entity's beginning index
2. Contains a mark of new line	2. Replace the match with a space
3. Contains a space before and/or after a hyphen or a parenthesis	3. Delete the space(s) and adjust the entity's ending index
4. Ends with non-content words or punctuation marks	4. Delete the match and adjust the entity's ending index
5. Concurs with non-content words or punctuation/hashtag marks	5. Leave out of the entities detected

Table 1: Post-processing rules

Model	**P**	**R**	**F1**
mBERT + post-processing	0.861	0.876	0.868

Table 2: Results (5-iteration mean) on the development dataset

4 Conclusions

In this work, we presented a system based on mBERT following a fine-tuning approach plus simple post-processing and search by propagation to extract disease mentions from Tweets in Spanish. We achieved competitive results with a strict F1 of 0.851, Precision of 0.842, and Recall of 0.860 on the test dataset of the SocialDisNER competition.

[1] The model is available at: https://bit.ly/3zGlxWy

[2] https://github.com/ajtamayoh/NLP-CIC-WFU-Contribution-to-SocialDisNER-shared-task-2022.git

Acknowledgements

This work was done with partial support from the Mexican Government through the grant A1-S-47854 of CONACYT, Mexico, grants 20220852 and 20220859 of the Secretaría de Investigación y Posgrado of the Instituto Politécnico Nacional, Mexico. The authors thank the CONACYT for the computing resources brought to them through the Plataforma de Aprendizaje Profundo para Tecnologías del Lenguaje of the Laboratorio de Supercómputo of the INAOE, Mexico and acknowledge the support of Microsoft through the Microsoft Latin America PhD Award.

References

Liliya Akhtyamova. 2020. Named entity recognition in spanish biomedical literature: Short review and bert model. In *2020 26th Conference of Open Innovations Association (FRUCT)*, pages 1–7. IEEE.

Emily Alsentzer, John R Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew McDermott. 2019. Publicly available clinical bert embeddings. *arXiv preprint arXiv:1904.03323*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. 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, Volume 1 (Long and Short Papers)*, pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.

Tome Eftimov, Barbara Koroušić Seljak, and Peter Korošec. 2017. A rule-based named-entity recognition method for knowledge extraction of evidence-based dietary recommendations. *PloS one*, 12(6):e0179488.

Aitor García-Pablos, Naiara Perez, and Montse Cuadros. 2020. Vicomtech at cantemist 2020. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2020), CEUR Workshop Proceedings*, volume 17, page 25.

Luis Gasco, Darryl Estrada-Zavala, Eulàlia Farré-Maduell, Salvador Lima-López, Antonio Miranda-Escalada, and Martin Krallinger. 2022. Overview of the SocialDisNER shared task on detection of diseases mentions from healthcare related and patient generated social media content: methods, evaluation and corpora. In *Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task*.

Yu Gu, Robert Tinn, Hao Cheng, Michael Lucas, Naoto Usuyama, Xiaodong Liu, Tristan Naumann, Jianfeng Gao, and Hoifung Poon. 2021. Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 3(1):1–23.

Matthew Honnibal and Ines Montani. 2017. spacy 2: Natural language understanding with bloom embeddings, convolutional neural networks and incremental parsing. *To appear*, 7(1):411–420.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 36(4):1234–1240.

Arjun Magge, Davy Weissenbacher, Karen O'Connor, Matthew Scotch, and Graciela Gonzalez-Hernandez. 2021. Seed: Symptom extraction from english social media posts using deep learning and transfer learning. *medRxiv*.

Mark Abraham Magumba, Peter Nabende, and Ernest Mwebaze. 2018. Ontology boosted deep learning for disease name extraction from twitter messages. *Journal of Big Data*, 5(1):1–19.

Mark Neumann, Daniel King, Iz Beltagy, and Waleed Ammar. 2019. ScispaCy: Fast and robust models for biomedical natural language processing. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 319–327, Florence, Italy. Association for Computational Linguistics.

Rakesh Patra and Sujan Kumar Saha. 2013. A kernel-based approach for biomedical named entity recognition. *The Scientific World Journal*, 2013.

Yifan Peng, Shankai Yan, and Zhiyong Lu. 2019. Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In *Proceedings of the 18th BioNLP Workshop and Shared Task*, pages 58–65, Florence, Italy. Association for Computational Linguistics.

Lance Ramshaw and Mitch Marcus. 1995. Text chunking using transformation-based learning. In *Third Workshop on Very Large Corpora*.

Devendra Singh Sachan, Pengtao Xie, Mrinmaya Sachan, and Eric P Xing. 2018. Effective use of bidirectional language modeling for transfer learning in biomedical named entity recognition. In *Machine learning for healthcare conference*, pages 383–402. PMLR.

Antonio Tamayo, Diego A. Burgos, and Alexander Gelbukh. 2022. mbert and simple post-processing: A baseline for disease mention detection in spanish. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

Zhiguo Wang, Patrick Ng, Xiaofei Ma, Ramesh Nallapati, and Bing Xiang. 2019. Multi-passage bert: A globally normalized bert model for open-domain question answering. *arXiv preprint arXiv:1908.08167*.

Davy Weissenbacher, Ari Z. Klein, Luis Gascó, Darryl Estrada-Zavala, Martin Krallinger, Yuting Guo, Yao Ge, Abeed Sarker, Ana Lucia Schmidt, Raul Rodriguez-Esteban, Mathias Leddin, Arjun Magge, Juan M. Banda, Vera Davydova, Elena Tutubalina, and Graciela Gonzalez-Hernandez. 2022. Overview of the seventh social media mining for health applications (#smm4h) shared tasks at coling 2022. In *In Proceedings of the Seventh Social Media Mining for Health (#SMM4H) Workshop and Shared Task.*

Ying Xiong, Yuanhang Huang, Qingcai Chen, Xiaolong Wang, Yuan Nic, and Buzhou Tang. 2020. A joint model for medical named entity recognition and normalization. *Proceedings http://ceur-ws. org ISSN*, 1613(0073):17.

Michihiro Yasunaga, Jure Leskovec, and Percy Liang. 2022. LinkBERT: Pretraining language models with document links. In *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 8003–8016, Dublin, Ireland. Association for Computational Linguistics.