IAI @ SocialDisNER : Catch me if you can! Capturing complex disease mentions in tweets

Aman Sinha*[1,2], Cristina García Holgado*[3], Marianne Clausel[1], Mathieu Constant[2]
[1]IECL, Université de Lorraine, Nancy, France ; [2]ATILF, Université de Lorraine, Nancy, France
[3]Lattice, École Normale Supérieure, Paris, France
{aman.sinha, marianne.clausel, mathieu.constant}@univ-lorraine.fr
cristina.gholgado@gmail.com

Abstract

Biomedical NER is an active research area today. Despite the availability of state-of-the-art models for standard NER tasks, their performance degrades on biomedical data due to OOV entities and the challenges encountered in specialized domains. We use Flair-NER framework to investigate the effectiveness of various contextual and static embeddings for NER on Spanish tweets, in particular, to capture complex disease mentions.

1 Motivation

In this paper, we present our system which recognizes disease mentions in Spanish tweets as part of the SocialDisNER challenge (Gasco et al., 2022). The motivation of our work is to:

(a) *Investigate the problem of identifying disease mentions in tweets* : Disease mentions occur with a variable length. While NER systems easily recognize simple-word entities, the task becomes exponentially difficult as the length increases (Dai, 2018; Shen et al., 2021), and also, it may complicate the gold annotation process. Moreover, entities can be composed of subgroups of entities (Dai et al., 2020). In this work, we deal with a particular linguistic context: tweet data. Here, we encounter an additional challenge as diseases mentions are specialized domain terms, but they occur in a social media platform characterized by informal language, comprising various noise-inducing elements such as hashtags, emojis, typos, and code-switching.

(b) *Investigate the performance of embedding resources* : Transfer learning (Pan and Yang, 2009) has been found to be very useful for downstream NLP tasks (Ruder et al., 2019). In this work, we explore the "capability" of different language models to use instilled knowledge from similar and different domains (Gururangan et al., 2020) and languages (Pfeiffer et al., 2020) pre-training datasets to learn NER on Spanish Twitter data, to see their effectiveness in addressing the above challenges.

GROUP	CONTEXTUAL	STATIC
domain	**xlrsc** (Lange et al., 2021) **rbbce** (Carrino et al., 2021) **sdf** (Chizhikova et al., 2022)	**es+clinical** **es+en+clinical**
multilingual	**xrl** (Conneau et al., 2019) **bbmcn** (Adelani, 2021) **wmn** (Tedeschi et al., 2021)	
es	**bscfn** (Cañete et al., 2020)	**es**
en	**bbucn** (Rawal, 2021)	

Table 1: Grouping of the different embeddings resources

2 Experimental Setup

2.1 Approach

To investigate the challenges of identifying complex disease mentions in tweets, we tested different embeddings using the Flair-NER (Akbik et al., 2019) framework using two types of models: Flair-S (Simple) and Flair-T (Transformers). Both of these models consist of a CRF-based SequenceTagger which takes input from a WordEmbedding/FlairEmbedding (Akbik et al., 2018) module for the static word embeddings (Flair-S), or from a TransformerWordEmbedding module for the contextual embeddings (Flair-T). We further defined four categories to group the different used embeddings as shown in Table 1.

2.2 Dataset

We used the GOLD SocialDisNER dataset (Gasco et al., 2022) provided by the organizers. It consists of a collection of health-related tweets from general users and a disease mention file annotated by medical experts. The disease mention file contains the annotated diseases mentions along with their begin and end offsets, and their corresponding

* Authors have equal contribution.

Proceedings of the 29th International Conference on Computational Linguistic, pages 85–89
October 12–17, 2022.

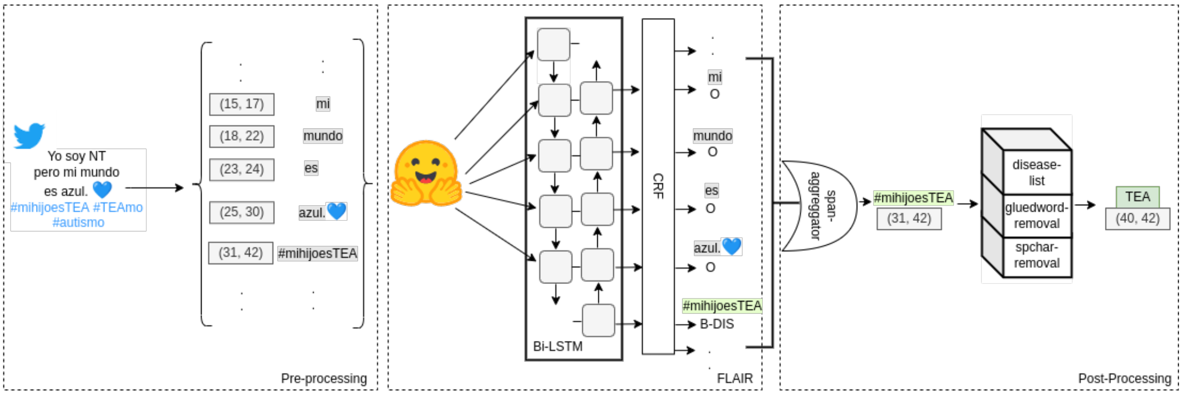

Figure 1: Experiment Pipeline

tweet file id. The dataset has 5000 training tweets (19425 mentions), 2500 validation tweets (4252 mentions), and 2500 test tweets files respectively.

2.3 Text-Processing

We perform a minimal pre-processing to keep all information in the text as we noticed a benefit from keeping the noise. We therefore propose the use of post-processing to obtain the strict spans from the identified disease mentions by removing different types of noise.

Pre-processing Tweets were tokenized on white space and converted into CoNLL format with no further pre-processing. Spans were generated for every token considering their character position in the text. BIO labels were assigned to every token matching the provided disease mentions in each tweet.

Post-processing Predictions were done over the tokenized tweets. The resulting predictions were converted into the mentions' file format where every detected disease mention span was aggregated. In a later post-processing step we fixed the aggregated spans. This fix of the original spans was performed on every detected disease mention when this contained any type of surrounding noise in the begin and/or end (See fig.1). This span-fix step determines the strict-f1 metric and includes three noise-treatment parts:

(a) Disease list: string-matching of disease mentions using an external custom list of disease words from a combination of the training disease mentions and online medical disease glossaries. This steps facilitates the removal of agglutinated words or attached noise (e.g. *#labioRojoContraLaMigraña*★ [0-26 → 18-25]).

(b) Glued words removal: removal of outer noise specific to Twitter hashtags when no disease mention is matched. The begin and the end of the mention is checked to remove this specific noise (e.g. *#DíaNacionalDelPárkinson* [0-24 → 16-24], *#TodasContraElCáncerDeMama* [0-26 → 15-26]).

(c) Special characters (Spchar) removal: removal of emojis, punctuation signs and other related characters when no disease mention is matched from the list (e.g. *#autismo*♡ [0-9 → 1-8]).

2.4 Implementation

For the Flair-T and Flair-S experiments, the models were trained for 50 epochs with a decaying learning rate (lr). Our setting were Flair-T (lr=5e-6, batch_size=4) and Flair-S (lr=1e-1, batch_size=2). The code is available at our Github repository.

	LMs	*Tag-F1*	*lenient-f1*	*strict-f1*
domain	**xrlsc**	**0.93**	**0.955**	**0.759**
	rbbce	0.93	0.951	0.757
	sdf	0.93	0.949	0.757
	es+clinical[†]	0.87	0.907	0.729
	es+en+clinical[†]	0.88	0.910	0.731
multilingual	**xrl**	0.93	0.950	0.756
	bbmcn	0.90	0.927	0.739
	wmn	0.89	0.925	0.735
es	**bscfn**	0.90	0.922	0.741
	es[†]	0.80	0.830	0.660
en	**bbucn**	0.87	0.914	0.725

Table 2: Final experiments results on Validation set; ([†]) correspond to Flair-S setting and rest of the entries correspond to Flair-T setting results.

3 Results

Table 2 shows the results on validation set. The systems were evaluated on a span-level using two metrics: *lenient-f1* and *strict-f1*. We also evaluated the systems on a token-level for BIO-tag classification (*Tag-f1*). Our final test set submission **xrlsc**

	lenient-f1	*strict-f1*
xrlsc	**0.941**	0.647
Mean	0.844 ± 0.168	0.675 ± 0.246
Median	0.898	0.761

Table 3: Comparison of our best submitted system (xrlsc) on Test set with other participants submissions

obtained 0.941 (*lenient-f1*) and 0.647 (*strict-f1*). Although *tag-f1* and *lenient-f1* reported similar scores, we observed a decrease on *strict-f1*. Compared to other submissions, our system performed ∼10% better over mean w.r.t. lenient-f1 whereas, it falls ∼2.5% short w.r.t. strict-f1 (see Table 3).

4 Discussion

`Model`	*without PP*	*with PP*
xrlsc	0.318	0.759
rbbce	0.319	0.757
sdf	0.318	0.757
es+clinical	0.308	0.729
es+en+clinical	0.313	0.731
xrl	0.317	0.756
bbmcn	0.311	0.739
wmn	0.306	0.735
bscfn	0.316	0.741
es	0.294	0.660
bbucn	0.298	0.725

Table 4: Effect of post-processing (PP) with metric *strict-f1* on Validation set

Impact of post-processing on span aggregation While the models were efficient at identifying the disease mentions on a token-level (as indicated by lenient-f1 scores in Table 2), the presence of noise in the extracted disease mentions leads to low strict-f1 scores without post-processing as shown in Table 4. The post-processing strategy of noise removal[1] increases significantly the strict-f1 by more than 40% (except for es-model). However, there is still a noticeable difference with the lenient-f1 which shows that this step encountered limitations to capture properly the strict spans. String-matching to remove surrounding noise on target entities is an effective strategy but it is however limited to known diseases from the provided list. When encountering unknown detected mentions, it becomes an arduous task due to the shifting and variation of different agglutinated noise-words (e.g. #ElVPHEsCosaDeTodos / #vacunavph / #DiaInternacionalContraelVPH). We also associate the reason of a lower strict-f1 to a displacement of the tokens' spans on the dataset due to the interference of special characters generated by multi-character emojis and a few encoding conflicts from the tweets extraction process. On further analysis, we found 170 tweet files potentially affected by this issue out of 2500 files in the validation set which comprises 7.92% of the total disease mentions.

Complex and discontinuous named entities Such entities are particularly problematic as they can lead to multiple disease mention identification. Variable context boundaries (e.g. *#dolorneuropatico en #COVID19?* → single or multiple disease mentions?) may lead to an alteration of the computation of the spans. We found that the error for capturing long and discontinuous entities was 26.11% lower for Flair-T models than Flair-S models. For noisy and agglutinated words such as *#HablemosDeVIH* or *#diabetestipo2*, Flair-T are more effective than Flair-S models. Besides the small number of errors on Flair-T in this sense, we noted a negative effect of agglutinated words with emojis (e.g. #*autismo*♡).

Transformation of entities With the character limit of tweets and the length of certain entities (e.g. *Enfermedad pulmonar intersticial difusa*), we encounter an increased use of acronyms (e.g. *#EPID*), which can be challenging for NER systems. We observed Flair-T models were 49.2% less prone to fail in detecting diseases' acronyms. Other transformations include flexions or verbal derivations (e.g. *resfriado*→*resfriarse*), where Flair-T was found to be consistently more effective.

Multilinguality Domain- and multilingual-specific embeddings have a comparable performance (refer Table 2). Both performed better than es-models as they benefited from the knowledge from the clinical datasets (Lange et al., 2021) that were used to pre-train them. Irrespective of the adaptive fine-tuning, en-specific models performed lower than es-specific models. Their marginal difference can be attributed to common standard disease names used by users on Twitter.

Gold Annotations Domain embeddings helped to identify unknown diseases and possible incom-

[1] We use blue font color to distinguish between target entities and noise

plete spans on the gold annotations. This raises the question of what sequence we can consider a disease and to what extent we determine its span. While our error analysis showed identification of words not strictly corresponding to diseases (e.g. *esquizofrenia cultural*), new ones were identified (e.g. *Alteraciones cutáneas*), in both Spanish and English (e.g. *#epilepsywarrior*), and the complete span was captured (e.g. *esteatosis hepática grado II-III*) when a gold annotation was incomplete. We found this to be an interesting strategy to explore to improve the quality of gold annotations.

5 Conclusion

In this paper, we studied the existing problems for Bio-Medical NER on Twitter data and further shown the effectiveness of contextualized embeddings. In addition, we observed a potential in these systems to improve the quality of gold annotations. Regardless of the high performance of the presented systems, post-processing remains a key step to achieve high quality extractions of target entities. In future work, we would like to investigate more effective post-processing strategies and a finer tagging schema such as nested annotations for a more accurate detection of complex disease mentions.

References

David Adelani. 2021. bert-base-multilingual-cased-ner-hrl.

Alan Akbik, Tanja Bergmann, Duncan Blythe, Kashif Rasul, Stefan Schweter, and Roland Vollgraf. 2019. FLAIR: An easy-to-use framework for state-of-the-art NLP. In *NAACL 2019, 2019 Annual Conference of the North American Chapter of the Association for Computational Linguistics (Demonstrations)*, pages 54–59.

Alan Akbik, Duncan Blythe, and Roland Vollgraf. 2018. Contextual string embeddings for sequence labeling. In *COLING 2018, 27th International Conference on Computational Linguistics*, pages 1638–1649.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario.

José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Spanish pre-trained bert model and evaluation data. In *PML4DC at ICLR 2020*.

Mariia Chizhikova, Jaime Collado-Montañéz, Pilar López-Úbeda, Manuel C. Díaz-Galiano, L. Alfonso Ureña-López, and M. Teresa Martín-Valdivia. 2022. Sinai at clef 2022: Leveraging biomedical transformers to detect and normalize disease mentions. pages 265–273.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2019. Unsupervised cross-lingual representation learning at scale. *CoRR*, abs/1911.02116.

Xiang Dai. 2018. Recognizing complex entity mentions: A review and future directions. In *Proceedings of ACL 2018, Student Research Workshop*, pages 37–44, Melbourne, Australia. Association for Computational Linguistics.

Xiang Dai, Sarvnaz Karimi, Ben Hachey, and Cecile Paris. 2020. An effective transition-based model for discontinuous NER. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 5860–5870, Online. Association for Computational Linguistics.

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*.

Suchin Gururangan, Ana Marasović, Swabha Swayamdipta, Kyle Lo, Iz Beltagy, Doug Downey, and Noah A Smith. 2020. Don't stop pretraining: adapt language models to domains and tasks. *arXiv preprint arXiv:2004.10964*.

Lukas Lange, Heike Adel, Jannik Strötgen, and Dietrich Klakow. 2021. Clin-x: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

Jonas Pfeiffer, Ivan Vulić, Iryna Gurevych, and Sebastian Ruder. 2020. Mad-x: An adapter-based framework for multi-task cross-lingual transfer. *arXiv preprint arXiv:2005.00052*.

Sam Rawal. 2021. bert-base-uncased-clinical-ner.

Sebastian Ruder, Matthew E Peters, Swabha Swayamdipta, and Thomas Wolf. 2019. Transfer learning in natural language processing. In *Proceedings of the 2019 conference of the North American chapter of the association for computational linguistics: Tutorials*, pages 15–18.

Yongliang Shen, Xinyin Ma, Zeqi Tan, Shuai Zhang, Wen Wang, and Weiming Lu. 2021. Locate and label: A two-stage identifier for nested named entity recognition. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 2782–2794, Online. Association for Computational Linguistics.

Simone Tedeschi, Valentino Maiorca, Niccolò Campolungo, Francesco Cecconi, and Roberto Navigli. 2021. WikiNEuRal: Combined neural and knowledge-based silver data creation for multilingual NER. In *Findings of the Association for Computational Linguistics: EMNLP 2021*, pages 2521–2533, Punta Cana, Dominican Republic. Association for Computational Linguistics.