CLaCLab at SocialDisNER: Using Medical Gazetteers for Named-Entity Recognition of Disease Mentions in Spanish Tweets

Harsh Verma, Parsa Bagherzadeh, Sabine Bergler
CLaC Labs, Concordia University
{h_ver, p_bagher, bergler} @cse.concordia.ca

Abstract

This paper summarizes the CLaC submission for SMM4H 2022 Task 10 which concerns the recognition of diseases mentioned in Spanish tweets. Before classifying each token, we encode each token with a transformer encoder using features from Multilingual RoBERTa Large, UMLS gazetteer, and DISTEMIST gazetteer, among others. We obtain a strict F1 score of 0.869, with competition mean of 0.675, standard deviation of 0.245, and median of 0.761. Our code is available here.

1 Motivation

Finding mentions of diseases in tweets in all languages has become an important tool for epidemiologists, especially in times of a pandemic. SocialDisNER (Gasco et al., 2022) at SMM4H 2022 concerns the recognition of disease mentions in Spanish tweets. A disease mention can include both lay and professional language and may be located in hashtags or usernames as well as the tweet text. Disease entities are underlined in Example 1:

(1) *Pasos sábado 23 de mayo! ... Sumemos 1 millón de pasos por la epilepsia, #1MillonDePasos, #Epilepsia, #ADosMetrosDeDistancia #InvestigaEpilepsia @RetoDravet #juntosesmejor @FundacionDravet*

2 System

Our contribution consists of a simple pipeline incorporating four main components: a standard tokenizer for splitting hashtags and username tokens with an ad hoc extension for disease recognition, word embeddings for Spanish and English from RoBERTa Large, four gazetteer lists extracted from relevant domain resources and represented as one hot vectors, and a linear classifier that produces BIO (*beginning, inside, outside*) tags for each token for recognition of multi-token disease mentions. The simplicity of this pipeline and its knowledge injection from readily available domain resources rather than training purely from training data make our system's strength.

Tokenization Sequence analysis classifying each token into BIO tags requires good tokenization. We pre-process the data using ANNIE (Cunningham et al., 2002) and the Twitter English Tokenizer (Bontcheva et al., 2013).

Hashtags and usernames usually include several words strung together without separating characters, e.g. #InvestigaEpilepsia. Since disease mentions can occur in these word composites, their individual word components have to be separated. Camel casing makes splitting this example into *Investiga* and *Epilepsia* easy, but for hashtags like #nosolohaycovid only knowledge of words in Spanish leads to the desired split into *no*, *solo*, *hay* and *covid*. Because disease names are not always part of the general vocabulary of a language, disease gazetteers are a lightweight means to inject such domain knowledge into a general machine learning system.

A gazetteer lists entities of a particular type. For example, a disease gazetteer would contain a list of disease names like COVID, Hepatitis, Epilepsia etc. Ad hoc gazetteer lists that are written for a specific task often suffer from limited coverage. We circumvent our own limitations by using four gazetteer lists compiled from extant resources:

GoldGaz: disease names compiled from the gold annotations of SocialDisNER training and validation data

DistemistGaz: (Miranda-Escalada et al., 2022) Spanish disease gazetteer compiled from Snomed-CT (Donnelly et al., 2006)

SilverGaz: disease gazetteer compiled from silver standard data made available for SocialDisNER (Gasco et al., 2022)

UmlsGaz: disease gazetteer of Spanish and English disease terms we compiled from UMLS (Bodenreider, 2004). Semantic type T047 represents *Disease or Syndrome*, identifying all UMLS concepts representing diseases.

Twitter tokenizer extension The GATE Twitter English Tokenizer (Bontcheva et al., 2013) identifies hashtags and usernames, but doesn't split them. To split usernames and hashtags, we identify the longest substring in a hashtag/username that matches an entity in our gazetteer, then treat that substring as a separate token. The extension matches first against GoldGaz, then DistemistGaz, followed by UmlsGaz, and finally SilverGaz.

Classification We frame the named entity recognition (NER) task as a sequence labeling task with the BIO (*beginning, inside, outside*) tagging scheme. In other words, we classify each token as either B (token is the beginning of a disease name), I (token lies inside a disease name) or O (token lies outside a disease name).

Model Preprocessing produces S, a list of t tokens. S is fed into a pretrained XLM RoBERTa Large (Conneau et al., 2020) model, which returns the matrix $\mathbf{H} \in \mathbb{R}^{t \times d}$, with $d = 1024$. The ith row of $\mathbf{H}$ is a vector of size d and corresponds to the contextualized embedding of the ith token.

We use the GATE ANNIE Gazetteer plugin to find tokens that match terms in UmlsGaz exactly (case-insensitively), creating $\mathbf{G}_{umls} \in \mathbb{R}^{t \times 2}$, a matrix where each row is a one hot vector indicating whether the ith token matches with *UmlsGaz* or not. Similarly, we create $\mathbf{G}_{silver}$ and $\mathbf{G}_{distemist}$ corresponding to *SilverGaz* and *DistemistGaz* respectively. Then, $\mathbf{H}$ is row-wise concatenated with $\mathbf{G}_{umls}$, $\mathbf{G}_{distemist}$, and $\mathbf{G}_{silver}$ to produce $\mathbf{Z} \in \mathbb{R}^{t \times 1030}$. $\mathbf{Z}$ is then fed into a 6-layer Transformer Encoder (Vaswani et al., 2017) with positional encoding and 10 attention heads per layer. The Transformer Encoder outputs $\mathbf{Y} \in \mathbb{R}^{t \times 1030}$ where the ith row represents the encoded representation of the ith token. $\mathbf{Y}$ is fed into a linear classifier which produces $\mathbf{I} \in \mathbb{R}^{t \times 3}$, classifying each token into one of the 3 classes corresponding to the BIO tagging scheme. This represents our submission system M_{sub}.

Training XLM RoBERTa large is fine-tuned on the training data using the Adam optimizer (Kingma and Ba, 2015) with a learning rate of 1e-5 and early stopping. Training for 10 epochs takes 2.5 hours on one Nvidia RTX 3090 gpu.

3 Ablation and test results

The evaluation is by strict F1 score, requiring exactly matching gold spans without partial credit.

Our submission model M_{sub} was described above in Section 2.2. Let $M_{no\text{-}gaz}$ be the same model as M_{sub} but without the one-hot features generated from gazetteer matching. Let $M_{no\text{-}tok}$ be the same model as M_{sub} but without the tokenizer extension using disease gazetteers (hashtags and usernames are tokenized with the GATE Hashtag Tokenizer only). Finally, let $M_{RoBERTa}$ be the baseline model without one-hot features, custom tokenization, and transformer encoder – the output of the language model $\mathbf{H}$ is fed directly into the classifier.

Model	F1 score	Precision	Recall
$M_{RoBERTa}$	0.880	0.856	0.905
$M_{no\text{-}gaz}$	0.888	0.875	0.902
$M_{no\text{-}tok}$	0.888	0.884	0.892
M_{sub}	0.892	0.882	0.900

Table 1: Ablation on development set

We see that M_{sub} improves on $M_{RoBERTa}$ by only 1.2%. Omitting gazetteers or the tokenizer extension each looses only 0.4% over M_{sub}. $M_{no\text{-}gaz}$ looses precision while $M_{no\text{-}tok}$ looses recall against the submission system.

Table 2 shows competition performance of M_{sub} on the test set. We see that performance on the test set decreased by only 2.3%, highlighting the robustness of our technique.

Model	F1 score	Precision	Recall
M_{sub}	0.869	0.851	0.888
Mean	0.675	0.680	0.677
Std. dev.	0.246	0.245	0.254
Median	0.761	0.758	0.780

Table 2: Competition results on test set

4 Conclusion

For the task of detecting disease mentions, a general large, pre-trained language model (XLM RoBERTa large) was enhanced with task-oriented preprocessing (splitting hashtags into component words) and lookup in available quality word lists.

This combination was effective and in competition placed our system nearly 20% above the mean.

References

Olivier Bodenreider. 2004. The unified medical language system (UMLS): integrating biomedical terminology. *Nucleic acids research*, 32(suppl_1):D267–D270.

Kalina Bontcheva, Leon Derczynski, Adam Funk, Mark A Greenwood, Diana Maynard, and Niraj Aswani. 2013. Twitie: An open-source information extraction pipeline for microblog text. In *Proceedings of the international conference recent advances in natural language processing RANLP 2013*, pages 83–90.

Alexis Conneau, Kartikay Khandelwal, Naman Goyal, Vishrav Chaudhary, Guillaume Wenzek, Francisco Guzmán, Edouard Grave, Myle Ott, Luke Zettlemoyer, and Veselin Stoyanov. 2020. Unsupervised cross-lingual representation learning at scale. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 8440–8451, Online. Association for Computational Linguistics.

Hamish Cunningham, Diana Maynard, Kalina Bontcheva, and Valentin Tablan. 2002. GATE: A Framework and Graphical Development Environment for Robust NLP tools and applications. In *Proceedings of the 40th Anniversary Meeting of the Association for Computational Linguistics (ACL'02)*.

Kevin Donnelly et al. 2006. Snomed-ct: The advanced terminology and coding system for ehealth. *Studies in health technology and informatics*, 121:279.

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*.

Diederik P Kingma and Jimmy Ba. 2015. Adam: A method for stochastic optimization. In *Proceedings of ICLR 2015*.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Working Notes of Conference and Labs of the Evaluation (CLEF) Forum. CEUR Workshop Proceedings*.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Proceedings of Advances in neural information processing systems NIPS 2017*.