Lasige-BioTM at ProfNER: BiLSTM-CRF and contextual Spanish embeddings for Named Entity Recognition and Tweet Binary Classification

Pedro Ruas and **Vitor D. T. Andrade** and **Francisco M. Couto**
LASIGE, Faculdade de Ciências da Universidade de Lisboa
1749-016 Lisboa, Portugal
psruas@fc.ul.pt, fc49005@alunos.fc.ul.pt, fcouto@di.fc.ul.pt

Abstract

The paper describes the participation of the Lasige-BioTM team at sub-tracks A and B of ProfNER, which was based on: i) a BiLSTM-CRF model that leverages contextual and classical word embeddings to recognize and classify the mentions, and ii) on a rule-based module to classify tweets. In the Evaluation phase, our model achieved a F1-score of 0.917 (0,031 more than the median) in sub-track A and a F1-score of 0.727 (0,034 less than the median) in sub-track B.

1 Introduction

The track "ProfNER-ST: Identification of professions & occupations in Health-related Social Media" (Miranda-Escalada et al., 2021b) occurred in the context of the "Social Media Mining for Health Applications (#SMM4H) Shared Task 2021" (Magge et al., 2021), and included two different sub-tracks that focused on Spanish Twitter data:

- Track A – Tweet binary classification: to determine if a given tweet has a mention of occupation or not.
- Track B – Named Entity Recognition (NER) offset detection and classification: to recognise the span of mentions of occupations and classify them in the respective category.

This paper describes the participation of the Lasige-BioTM team in the aforementioned sub-tracks. We applied 8 different models NER models (4 supervised models based on BiLSTM-CRF architecture, 3 rule-based models) to predict entities for sub-track B and explored the impact of performing data augmentation in the training set. For sub-track A, we developed a rule-based model for tweet classification that was based on the NER output for sub-track B.

1.1 Related Work

According to Goyal et al. (2018), NER approaches can be divided in two categories: rule-based and machine learning-based, being the latter further subdivided into supervised, semi-supervised, unsupervised; other approaches combine aspects from the two categories and are thus designated by hybrid. The models with an architecture consisting of a bidirectional Long Short-Term Memory (BiLSTM) network and a Conditional Random Field (CRF) decoding layer are among the state-of-the-art approaches for the NER task. (Huang et al., 2015). For a comprehensive overview of the existing NER approaches please refer to Goyal et al. (2018) and, specifically for the biomedical domain, to Lamurias and Couto (2019).

2 Methodology

2.1 Corpus description

The ProfNER corpus (Miranda-Escalada et al., 2020) contains 10,000 health-related tweets in Spanish that were annotated by linguist experts with entities relative to professions, employment statuses, and other work-related activities and includes four categories: "PROFESION", "SITUACION_LABORAL", "ACTIVIDAD", and "FIGURATIVA". For sub-track A, a given tweet was assigned the label "1" if it included at least one entity belonging to any category, but for sub-track B only entities belonging to categories "PROFESION" and "SITUACION_LABORAL" were considered for evaluation.

2.2 Pre-processing

We performed data augmentation on the training set of the corpus using the Python library `nlpaug` (Ma, 2019). For example, considering the mentioned entity "médico" present in the training set, data augmentation consisted of substituting a random character by a keyboard character (i.e. replac-

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 108–111
June 10, 2021.

ing the character by a neighbour character in the keyboard in order to simulate a typing error character, since Twitter data is usually noisy: "médico" → "médLco"), by a random distance character ("médico" → "médicB"), and by a synonym (i.e. replacing the character by a synonym in the Spanish WordNet: "médico" → "dr."). The output of this step consisted of three additional training files besides the original training file, each one associated with the result of a type of augmentation.

2.3 MER

The first approach was based on `MER` (Couto and Lamurias, 2018), a minimal NER tagger that recognizes entities and the respective span in text according to a given lexicon. It is based on the text processing command-line tools `grep` and `awk`, and on an inverted recognition technique that uses the words in input text as patterns to match the lexicon words. Several lexicons were created and processed including: 1) mentions in "PROFESION" category in training set and its WordNet synonyms, 2) mentions in "PROFESION" category in training set and its WordNet synonyms, jointly with entities present in the Occupations gazetteer provided by the organisation (Asensio et al., 2021), 3) mentions in "SITUACION_LABORAL" category in training set and its WordNet synonyms, 4) entities in "ACTIVIDAD" category in train set and its WordNet synonyms, 5) entities in "FIGURATIVA" category in train set and its WordNet synonyms. The first model ("MER 1") included the lexicons 1, 3, 4, and 5, the second model ("MER 2") included the lexicons 2, 3, 4, and 5, the third model ("MER 3") was similar to the first one but the mention "sin" was filtered out. During Practice phase, we built the lexicons from the training set and used the validation set as the test set. For sub-task A, we developed a rule-based module to classify each tweet with the label "1" if at least one mention was recognized in the respective text, and with label "0" otherwise.

2.4 BiLSTM-CRF

To implement the second approach, we resorted to the `FLAIR` framework (Akbik et al., 2019), and created an object of the class `SequenceTagger`, which instantiates a NER model with an architecture consisting of a BiLSTM network and a CRF decoding layer. LSTM are recurrent neural networks (RNNs), which include an input layer x representing features at time t, one or more hidden layers h, and an output layer y, which in the case of the NER task, represents a probability distribution over labels or tags at time t. A CRF network focus on the sentence level and also uses past and future tags/labels to predict the current one. The combination of a BiLSTM network with a CRF network has shown performance improvements over alternative architectures (Huang et al., 2015).

In the NER task, text needs to be tokenized and vectorized before being inputed to the neural network, which can be done leveraging pre-trained embeddings. FastText embeddings (Bojanowski et al., 2017) are an improvement over classic word embeddings, more concretely the skipgram model, by capturing sub-word information. `FLAIR` embeddings (Akbik et al., 2018) are contextual string embeddings that capture syntactic-semantic word features. We have explored the integration of different types of embeddings in the BiLSTM-CRF model through the `StackedEmbeddings` class:

- "Base" : `FLAIR` embeddings ("es-forward" and "es-backward") trained on Spanish Wikipedia (Akbik et al., 2018) + Spanish FastText embeddings
- "Twitter" : FastText Spanish COVID-19 Twitter Embeddings, provided by the organization (Miranda-Escalada et al., 2021a) (uncased version of the cbow model).
- "Medium" : `FLAIR` embeddings ("es-forward" and "es-backward") + Spanish FastText embeddings + FastText Spanish COVID-19 Twitter Embeddings

For the sub-track A, we applied a similar rule-based module as described in Section 2.3. If a model recognizes at least one entity in a given tweet in the context of sub-track B, the module assigns the label "1" to the respective tweet. If no entity is recognized in a given tweet, this receives the label "0". All the tweet IDs and respective label are then outputted in the predictions file for sub-track A.

2.4.1 Training

During Practice phase, we trained the models "Base" and "Twitter" on the original training file ("Base" and "Twitter"), and additionally, on the three files that resulted from the data augmentation step ("Base-aug" and "Twitter-aug"). During Evaluation phase, we merged the training and validation annotations, resulting in a file composed by 14,674 sentences for training and 1,630 sentences

for validation. The training parameters were set to: `hidden size` = 256, `Mini batch size` = 32, `Max epochs` = 55, `Patience` = 3.

3 Results and discussion

3.1 Practice phase

The performance of the referred models in the validation set for sub-tracks A and B are available in Table 1. The "Base" model trained on the original training file achieved the best performance in sub-tracks A and B: F1-scores (strict) of 0.908 and 0.716, respectively. Consequently, we selected this model for further training and application in the test set. The models trained on files resulting from data augmentation achieved lower performances compared with the respective versions trained exclusively on the original training file.

3.2 Evaluation phase

The results achieved by our model in the Evaluation phase and the median results for all competing teams are shown in Table 2. In sub-track A, our model achieved a F1-score of 0.917 (0.031 more than the median) and in sub-track our model achieved a F1-score of 0.727 (0.034 less than the median).

3.3 Error analysis

The model "Base", that uses contextual embeddings trained on a general corpora, obtained higher performance when comparing to the model "Twitter", although this latter model uses Twitter-specific embeddings, more concretely, FastText embeddings that were trained on Twitter data. For instance, consider the following tweet of the validation set: *"Ya que están sesionando la importante pero NO prioritaria #LeyDeAmnistia,será que también vean la cuestión de #Economia y #SaludParaTodos? Digo!Recuerden que su prioridad somos los millones que estamos indefensos ante el #COVID-19 y sin trabajo @MorenaSenadores #LeyDeAmnistiaNo https://t.co/DCiuqiBjEs"*. The model "Twitter" recognizes the mention "@MorenaSenadores" and assigns the "PROFESION" category to it, whereas the model "Base" does not recognize any mention, since is been able to assume in this context that the mention do not correspond to a profession, but instead to a Twitter handle. There is a mention with the string "senadores" classified as "PROFESION" in a tweet of the training set, which maybe leads the model "Twitter" to assume that the words "@MorenaSenadores" must also correspond to a mention, since the string is similar.

4 Conclusion

During the Practice Phase, we explored different approaches to participate in sub-tracks A e B of ProfNER: data augmentation on training set, and application of MER and a BiLSTM-CRF model for NER and further tweet classification. For the Evaluation phase we applied the BiLSTM-CRF model on the test set of ProfNER corpus and achieved F1-scores of 0.917 (0,031 more than the median) and in sub-track our model achieved a F1-score of 0.727 (0,034 less than the median). The code to run the experiments is available in our GitHub page[1]. For future work, we intend to perform hyper-parameter optimisation for the BiLSTM-CRF model, such as learning rate, hidden size, and specially the number of training epochs, since we had limited available time to perform the training of the model. We will also explore the use of different contextualised embeddings, since the models using this type of embeddings seem to achieve better performance compared to those using classical word embeddings. Besides, to improve tweet classification we will explore the application of Named Entity Linking tools (Lamurias et al., 2019) to link the recognized entities in sub-track B to structured vocabularies that contain hierarchical relationships between concepts, such as MeSH or DBpedia. This way, it will be possible to know the ancestors for a given entity, which will provide the context to effectively determine if the entity is associated with an occupation or not.

Acknowledgements

This work was supported by FCT through funding of Deep Semantic Tagger (DeST) project (ref. PTDC/CCI-BIO/28685/2017) and LASIGE Research Unit (ref. UIDB/00408/2020 and ref. UIDP/00408/2020); and FCT through funding of PhD Scholarship, ref. 2020.05393.BD.

References

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*

[1] https://github.com/lasigeBioTM/LASIGE-participation-in-ProfNER

	Sub-track 7A			Sub-track 7B					
Model	**P**	**R**	**F1**	**P**	**R**	**F1**	**Rel-P**	**Rel-R**	**Rel-F1**
MER 1	0.621	0.767	0.687	0.399	0.535	0.457	0.565	0.668	0.612
MER 2	0.498	0.839	0.625	0.290	0.578	0.386	0.418	0.721	0.529
MER 3	0.621	0.767	0.687	0.472	0.535	0.501	0.668	0.667	0.667
Base	**0.941**	0.876	**0.908**	**0.795**	**0.651**	**0.716**	**0.901**	**0.738**	**0.811**
Base-aug	0.848	0.830	0.839	0.705	0.616	0.657	0.826	0.721	0.770
Twitter	0.895	0.874	0.884	0.721	0.616	0.664	0.856	0.730	0.788
Twitter-aug	0.786	0.904	0.841	0.597	0.611	0.604	0.737	0.755	0.746
Medium-aug	0.780	0.887	0.830	0.618	0.601	0.609	0.753	0.733	0.743

Table 1: Practice results for sub-track 7A (left) and sub-track 7B (right). P, R, and F1 refer to precision, recall, and F1-score (strict), respectively and Rel-P, Rel-R, and Rel-F1 refer to relaxed precision, relaxed recall, and relaxed F1-score, respectively

	Sub-track 7A			Sub-track 7B		
Model	**P**	**R**	**F1**	**P**	**R**	**F1**
Lasige-BioTM	**0.951**	**0.886**	**0.917**	0.814	0.657	0.727
Median	0.919	0.855	0.886	**0.842**	**0.727**	**0.761**

Table 2: Evaluation phase results for sub-tracks 7A and 7B. P, R, F1 refer to precision, recall, and F1-score (strict), respectively.

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.

Alejandro Asensio, Antonio Miranda-Escalada, Marvin Aguero, and Martin Krallinger. 2021. Occupations gazetteer - ProfNER & MEDDOPROF - occupations, professions and working status terms with their associated codes. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information.

Francisco M. Couto and Andre Lamurias. 2018. MER: a shell script and annotation server for minimal named entity recognition and linking. *Journal of Cheminformatics*, 10(1):58.

Archana Goyal, Vishal Gupta, and Manish Kumar. 2018. Recent Named Entity Recognition and Classification techniques: A systematic review. *Computer Science Review*, 29:21–43.

Zhiheng Huang, Wei Xu, and Kai Yu. 2015. Bidirectional lstm-crf models for sequence tagging.

Andre Lamurias and Francisco M Couto. 2019. Text Mining for Bioinformatics Using Biomedical Literature. In K. and Ranganathan, S., Gribskov, M., Nakai and C Schoonbach, editors, *Encyclopedia of Bioinformatics and Computational Biology, vol. 1*, January, pages pp. 602–61. Oxford: Elsevier.

Andre Lamurias, Pedro Ruas, and Francisco M. Couto. 2019. PPR-SSM: Personalized PageRank and semantic similarity measures for entity linking. *BMC Bioinformatics*, 20(1):1–12.

Edward Ma. 2019. Nlp augmentation. https://github.com/makcedward/nlpaug.

Arjun Magge, Ari Z. Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulalia Farre, Salvador Lima, Juan Banda, Karen O'Connor, Abeed Sarker, Elena Tutubalina, Martin Krallinger, Davy Weissenbacher, and Graciela Gonzalez-Hernandez. 2021. Overview of the sixth social media mining for health applications (#smm4h) shared tasks at naacl 2021.

Antonio Miranda-Escalada, Marvin Aguero, and Martin Krallinger. 2021a. Spanish covid-19 twitter embeddings in fasttext. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Antonio Miranda-Escalada, Vicent Briva-Iglesias, Eulàlia Farré, Salvador Lima López, Marvin Aguero, and Martin Krallinger. 2020. ProfNER corpus: gold standard annotations for profession detection in Spanish COVID-19 tweets. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Antonio Miranda-Escalada, Eulàlia Farré-Maduell, Salvador Lima López, Luis Gascó-Sánchez, Vicent Briva-Iglesias, Marvin Agüero-Torales, and Martin Krallinger. 2021b. The profner shared task on automatic recognition of occupation mentions in social media: systems, evaluation, guidelines, embeddings and corpora. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.