Word Embeddings, Cosine Similarity and Deep Learning for Identification of Professions & Occupations in Health-related Social Media

Sergio Santamaría Carrasco and **Roberto Cuervo Rosillo**
Universidad Carlos III de Madrid,
Computer Science Department,
Av de la Universidad, 30,
28911, Leganés, Madrid, Spain
sesantam@pa.uc3m.es rcuervo@pa.uc3m.es

Abstract

ProfNER-ST focuses on the recognition of professions and occupations from Twitter using Spanish data. Our participation is based on a combination of word-level embeddings, including pre-trained Spanish BERT, as well as cosine similarity computed over a subset of entities that serve as input for an encoder-decoder architecture with attention mechanism. Finally, our best score achieved an F1-measure of 0.823 in the official test set.

1 Introduction

During situations of risk, such as the Covid-19 pandemic, detecting vulnerable occupations, be it due to their risk of direct exposure to the threat or due to mental health issues associated with work-related aspects, is critical to prepare preventive measures. These occupations can be detected through the analysis of tweets since Twitter has become a very useful tool to find reliable information. Due to the exponential growth of the use of this social network, natural language processing (NLP) techniques have become a crucial tool for unlocking this critical information.

This paper describes the participation of our team in ProfNER-ST (Miranda-Escalada et al., 2021b) challenge, 7b subtrack of the sixth Social Media Mining for Health Applications (SMM4HA) (Magge et al., 2021), which focuses on the recognition of professions and occupations from Twitter using Spanish data.

The core of the proposed system is based on an encoder-decoder architecture with attention mechanism successfully applied previously (Ali and Tan, 2019) for temporal expression recognition. This system combines several neural network architectures for the extraction of characteristics at a contextual level and a CRF for the decoding of labels. The proposed system reach a F1 score of 0.823.

2 Methods and system description

2.1 Pre-processing

We pre-process the text of the clinical cases taking into account different steps. First, the corpus are clean from urls. Secondly, the tweets are split into tokens using Spacy[1], an open-source library that provides support for texts in several languages, including Spanish. Finally, the text and its annotations are transformed into the CoNLL-2003 format using the BIOES schema (Ratinov and Roth, 2009).

2.2 Features

- **Words**: Two different 300 dimensional representations based on pre-trained word embeddings has been used with FastText (Bojanowski et al., 2016). Both have been selected for their contribution of domain-specific knowledge since the former have been generated from Spanish medical corpora (Soares et al., 2020) and the latter have been trained with Spanish Twitter data related to COVID-19 (Miranda-Escalada et al., 2021a). Contextual embeddings generated with a fine-tuned BETO (Cañete et al., 2020) model are also included, as these word representations are dynamically informed by the surrounding words improving performance.

- **Part-of-speech**: This feature has been considered due to the significant amount of information it offers about the word and its neighbors. It can also help in word sense disambiguation. The PoS-Tagging model used was the one provided by the Spacy. An embedding representation of this feature is learned during training, resulting in a 40-dimensional vector.

- **Characters**: We also add character-level embeddings of the words, learned during train-

[1] https://spacy.io/

ing and resulting in a 30-dimensional vector. These have proven to be useful for specific-domain tasks and morphologically-rich languages.

- **Syllables**: Syllable-level embeddings of the words, learned during training and resulting in a 75-dimensional vector is also added. Like character-level embeddings, they help to deal with words outside the vocabulary and contribute to capturing common prefixes and suffixes in the domain and correctly classifying words.

- **Cosine Similarity**: The BETO embeddings of the entities found in the training and validation set are used to calculate the cosine similarity between the BETO representation of the word to be analyzed, since previous work (Büyüktopaç and Acarman, 2019) has shown that could help to improve the results on data extracted from Twitter. This information is encoded as a 3717-dimensional vector.

2.3 Architecture

In the proposed system, shown in Figure 1, the character and syllable information is previously processed by a convolutional and global max pooling block, to be concatenated with the rest of the input features to serve as input to an encoder-decoder architecture with attention mechanism. The context vector as well as decoder outputs feeds a fully connected dense layer with $tanh$ activation function. The last layer (CRF optimization layer) consists of a conditional random fields layer selected due to the ability of the layer to take into account the dependencies between the different labels. The output of this layer provides the most probable sequence of labels.

The system has been developed in python 3 (Van Rossum and Drake, 2009) with Keras 2.2.4 (Chollet et al., 2015) and Tensorflow 1.14.0 (Abadi et al., 2016).

3 Results

During experimentation our team apply the standard measures, precision, recall, and micro-averaged F1-score, to evaluate the performance of our model.

While the training set (Miranda-Escalada et al., 2020) was used for training the model, the development set was exploited to hyperparameter fine tuning. In the prediction stage, we combined both sets

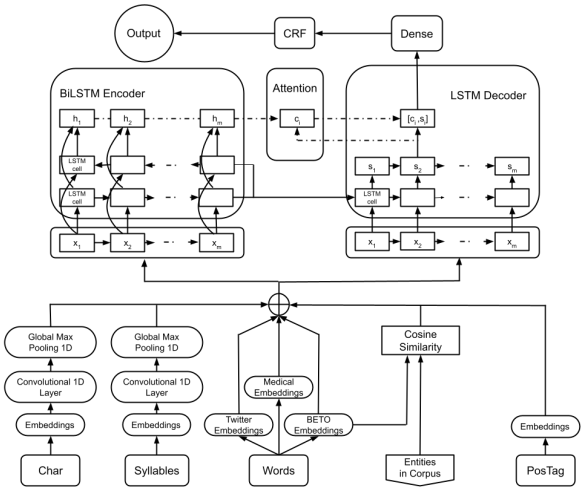

Figure 1: Architecture of the proposed model for profession and occupations recognition.

to training the model. The detailed hyper-parameter settings are illustrated in Table 1 'Opt.' denotes optimal.

Parameters	Tuned range	Opt
Train batch size	[8, 32, 64]	32
Epoch number	[2,3,4,5,6]	4
Dropout	[0.4, 0.5]	0.4
Max Seq Length	[50, 75, 100]	75
Learning rate	[0.01, 0.001, 0.0001]	0.001
Optimizer	-	Adam

Table 1: Hyper-parameters details.

With the optimal parametric configuration obtained during the experimentation, the model obtains the results shown in the Table 2.

	Precision	Recall	F-1 Score
Validation	0.893	0.753	0.817
Test	0.883	0.77	0.823

Table 2: Final results obtained in the competition.

4 Conclusion

In these working notes we describe our proposed system based on an encoder-decoder architecture with an attention mechanism powered by a combination of word embeddings that include pre-trained fine-tuned Spanish BERT embeddings.

Future work would explore different Data Augmentation techniques as well as other entities information, as companies or organizations, which could contain important information related to occupations.

References

Martín Abadi, Paul Barham, Jianmin Chen, Zhifeng Chen, Andy Davis, Jeffrey Dean, Matthieu Devin, Sanjay Ghemawat, Geoffrey Irving, Michael Isard, et al. 2016. Tensorflow: A system for large-scale machine learning. In *12th {USENIX} Symposium on Operating Systems Design and Implementation ({OSDI} 16)*, pages 265–283.

Mohammed NA Ali and Guanzheng Tan. 2019. Bidirectional encoder–decoder model for arabic named entity recognition. *Arabian Journal for Science and Engineering*, 44(11):9693–9701.

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2016. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*.

Onur Büyüktopaç and Tankut Acarman. 2019. Evaluation of cosine similarity feature for named entity recognition on tweets. In *International Conference on Man–Machine Interactions*, pages 125–135. Springer.

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*.

Francois Chollet et al. 2015. Keras.

Arjun Magge, Ari Klein, Ivan Flores, Ilseyar Alimova, Mohammed Ali Al-garadi, Antonio Miranda-Escalada, Zulfat Miftahutdinov, Eulàlia Farré-Maduell, Salvador Lima López, Juan M 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. In *Proceedings of the Sixth Social Media Mining for Health Applications Workshop & Shared Task*.

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, Vicent Briva-Iglesias, Marvin Agüero-Torales, Luis Gascó-Sánchez, and Martin Krallinger. 2021b. The profner shared task on automatic recognition of professions and 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*.

Lev Ratinov and Dan Roth. 2009. Design challenges and misconceptions in named entity recognition. In *Proceedings of the Thirteenth Conference on Computational Natural Language Learning (CoNLL-2009)*, pages 147–155.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Jordi Armengol-Estapé, Siamak Barzegar, and Martin Krallinger. 2020. Fasttext spanish medical embeddings. Funded by the Plan de Impulso de las Tecnologías del Lenguaje (Plan TL).

Guido Van Rossum and Fred L. Drake. 2009. *Python 3 Reference Manual*. CreateSpace, Scotts Valley, CA.

A Online Resources

The sources for the Troy&AbedInTheMorning participation are available via

- GitHub `https://github.com/ssantamaria94/ProfNER-SMM4H`,