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

Sergio Santamaría Carrasco and Roberto Cuervo Rosillo

ssantamaria94/ProfNER-SMM4H

ProfNER

SMM4H - SPANISH

Universidad Carlos III

Introduction

During situation of risk, 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.

Figure 1: Tweet example.

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 as well as cosine similarity computed over a subset of entities that serve as input for an encoder-decoder architecture with attention mechanism.

Dataset Description

10. 000 health-related tweets in spanish [1] annotated with proffessions and employment statuses by linguistic experts.

	Training	Validation	Training and Development
Words	32357	14836	39389
Characters	590	403	667

Table 1: Vocabulary statistics.

Next table describes the statistics of the dataset relevant for the proposed system.

	Training	Validation	Test
Tweets	6000	2000	2000

Table 2: Tweets compilation.

Method and System Description

Preprocessing

- Clear corpus
- Tokenization spaCy
- Annotations to BIOES schema

Tokenized sentece: Mary was born in Mississippi, and is currently living with daughter in Grand Island.

BIO representation: Mary/B-PATIENT was/O born/O in/O Mississipi/B-STATE ,/O and/O is/O currently/O living/O with/O daughter/O in/O Grand/B-CITY Island/I-CITY ./O

BIOES representation: Mary/S-PATIENT was/O born/O in/O Mississipi/S-STATE ,/O and/O is/O currently/O living/O with/O daughter/O in/O Grand/B-CITY Island/E-CITY ./O

Features

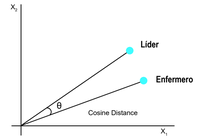
Words: Two different 300 dimensional representations based on pre-trained word embeddings [2] [3] and contextual embeddings generated with a fine-tuned BETO [4] .

Part-of-speech: An embedding representation of this feature is learned during training using the PoS-Tagging model provided by the Spacy.

Characters: Character-level embeddings of the words learned during training.

Syllables: Syllables-level embeddings of the words learned during training.

Cosine Similarity: Cosine similarity between BETO representation of the word to be analyzed and BETO embeddings of the entities found in the training and validation set



$$similarity(A,B) = \frac{A \cdot B}{\|A\| \times \|B\|} = \frac{\sum_{i=1}^{n} A_i \times B_i}{\sqrt{\sum_{i=1}^{n} A_i^2} \times \sqrt{\sum_{i=1}^{n} B_i^2}}$$

Architecture

Encoder-Decoder arquitecture with attention mechanism.

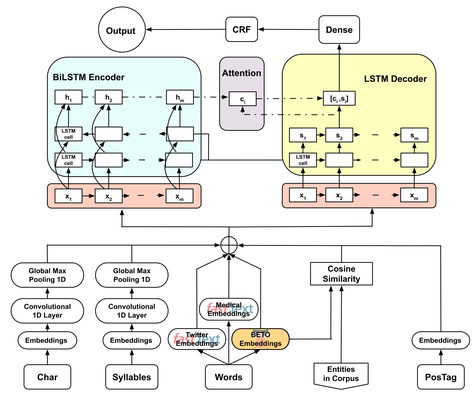


Figure 2: System Architecture.

Results

- Evaluation Metrics: Precision, Recall, F1-score.
- Training set was used for training the model while validation set was exploited to hyperparameter fine tuning.

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 3: Hyper-parameters details.

- Results with optimal parametric config.

	Precision	Recall	F1-Score
Validation	0.893	0.753	0.817
Test	0.883	0.77	0.823

Table 4: Final results obtained competition

Error Analisys

Three type of errors:

- Incorrect bondaries.
- Missing the entity mention (Missed).
- Incorrectly distinguishing the entity mention (Incorrectly distinguish).

Figure 3: Incorrect detections word cloud.

Figure 4: Missed word cloud.

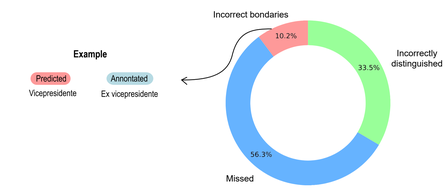


Figure 5: Error percentage.

Conclusion

Our encoder-decoder architecture with attention mechanism powered by word embeddings generates competitive results.

Future work:

- Data Augmentation techniques
- Other entities information as companies or organizations.

References

[1] Antonio Miranda-Escalada, Vicent Briva-Iglesias, Eu-làlia Farré, Salvador Lima López, Marvin Aguero,and Martin Krallinger. 2020.ProfNER corpus:gold standard annotations for profession detection inSpanish COVID-19 tweets. Funded by the Plan deImpulso de las Tecnologías del Lenguaje (Plan TL).

[2] Antonio Miranda-Escalada, Marvin Aguero, and Mar-tin Krallinger. 2021a. Spanish covid-19 twitter em-beddings in fasttext. Funded by the Plan de Impulsode las Tecnologías del Lenguaje (Plan TL).

[3] Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre,Jordi Armengol-Estapé, Siamak Barzegar, and Mar-tin Krallinger. 2020. Fasttext spanish medical em-beddings. Funded by the Plan de Impulso de lasTecnologías del Lenguaje (Plan TL).

[4] José Cañete, Gabriel Chaperon, Rodrigo Fuentes, Jou-Hui Ho, Hojin Kang, and Jorge Pérez. 2020. Span-ish pre-trained bert model and evaluation data. InPML4DC at ICLR 2020.