Assessing multiple word embeddings for named entity recognition of professions and occupations in health-related social media

Vasile Păiș and **Maria Mitrofan**
Research Institute for Artificial Intelligence "Mihai Drăgănescu", Romanian Academy
Casa Academiei, Calea 13 Septembrie nr. 13, sector 5, Bucureşti, ROMÂNIA
vasile,maria@racai.ro

Abstract

This paper presents our contribution to the ProfNER shared task. Our work focused on evaluating different pre-trained word embedding representations suitable for the task. We further explored combinations of embeddings in order to improve the overall results.

1 Introduction

The ProfNER task (Miranda-Escalada et al., 2021b), part of the SMM4H workshop and shared task (Magge et al., 2021) organized at NAACL 2021, focused on identification of professions and occupations from health-relevant Twitter messages written in Spanish. It offered two sub-tasks: a) a binary classification task, deciding if a particular tweet contains a mention of an occupation, given the context, and b) extracting the actual named entities, by specifying the entity type, start and end offset as well as the actual text span.

Habibi et al. (2017) have shown that domain specific embeddings have an impact on the performance of a NER system. The ProfNER task is at a confluence between multiple domains. The classification sub-task suggests that tweets will actually contain not only health-related messages but probably also more general domain messages. However, the second task focuses on the analysis of health-related messages. Finally, social media can be regarded as a domain in itself. Therefore, our system was constructed on the assumption that word embeddings from multiple domains (general, health-related, social media) will have different impact on the performance of a NER system. We evaluated different pre-trained embeddings alone and in combination, as detailed in the next section.

Our interest for the task stemmed from our involvement with the CURLICAT[1] project for the CEF AT action, where NER in different domains (including health-related) is needed. Additionally, pre-trained word embeddings for Romanian language, such as Paiș and Tufiș (2018), are considered for suitability in different tasks within the European Language Equality (ELE)[2] project.

2 System description and results

We used a recurrent neural network model based on LSTM cells with token representation using pre-trained word embeddings and additional character embeddings, computed on the fly. The actual prediction is performed by a final CRF layer. For the implementation we used the NeuroNER[3] (Dernoncourt et al., 2017) package.

We considered the two sub-tasks to be intertwined. If a correct classification is given for the first sub-task, then this can be used in the second task to guide the NER process to execute only on the classified documents. However, also the reverse can be applied. A document containing correctly identified entities for the second sub-task should be classified as belonging to the domain of interest. We employed the second approach and first performed NER and then used this information for classification.

For the purposes of the NER sub-task we considered the following word embedding representations: Spanish Medical Embeddings[4] (Soares et al., 2019), Wikipedia Embeddings[5] (Mikolov et al., 2018), Twitter Embeddings[6] (Miranda-Escalada et al., 2021a). These were generated using the FastText toolkit (Bojanowski et al., 2017) and contain floating point vectors of dimension 300. The Spanish Medical Embeddings offers three variants

[1] https://curlicat-project.eu/

[2] http://www.european-language-equality.eu

[3] http://neuroner.com/

[4] https://zenodo.org/record/3744326#.YEbu950zZPZ

[5] https://fasttext.cc/docs/en/english-vectors.html

[6] https://zenodo.org/record/4449930#.YEbwUp0zZPY

Proceedings of the Sixth Social Media Mining for Health Workshop 2021, pages 128–130

Representation	P	R	F1
Medical	83.70	69.43	75.90
Twitter	82.92	71.58	76.83
Wiki	80.63	74.19	77.28
Twitter+Wiki	79.93	72.20	75.87
Twitter+Wiki (all)	81.90	72.96	77.17
Wiki+Twitter	80.86	75.27	77.96
Wiki+Twitter+Med	**83.84**	**75.73**	**79.58**

Table 1: Results of different word embeddings and combinations on the validation set for the NER subtask

Representation	P	R	F1
Medical	**92.38**	86.37	89.27
Twitter	92.05	87.42	89.68
Wiki	90.08	**89.52**	89.80
Twitter+Wiki	90.83	89.31	**90.06**
Twitter+Wiki (all)	91.67	87.63	89.60
Wiki+Twitter	89.68	89.31	89.50
Wiki+Twitter+Med	91.18	88.89	90.02

Table 2: Results of different word embeddings and combinations on the validation set for the Classification subtask

based on the SciELO[7] database of scientific articles, filtered Wikipedia (comprising the categories Pharmacology, Pharmacy, Medicine and Biology) and a reunion of the two datasets. For all three corpora, representations are available using CBOW and Skip-Gram algorithms, as described in Bojanowski et al. (2017). However we only used the Skip-Gram variants for our experiments, due to the availability of this type of pre-trained vectors for all the considered representations.

We first experimented with individual representations and then began experimenting with sets of two embeddings concatenated. For the words present in the first considered embedding we added the corresponding vector from the second embedding or a zero vector. This provided an input vector of size 600 (resulting from concatenating two vectors of size 300 each), which required the adaptation of the network size accordingly. Additionally we considered a full combination of Twitter and Wikipedia embeddings, placing zero-valued vectors if words were also missing from the first embedding. A final experiment was conducted on a concatenation of 3 embeddings (total vector size 900). Results on the validation set are presented in Table 1 and Table 2, while results on the test set are in Table 3.

[7] https://scielo.org/

Representation	NER F1	Classification F1
Medical	73.60	86.43
Twitter	74.60	88.04
Wiki	75.40	88.72
Twitter+Wiki	76.20	**88.98**
Wiki+Twitter	75.70	88.38
Twitter+Wiki (all)	75.30	88.24
Wiki+Twitter+Med	**78.50**	88.81

Table 3: Results of different word embeddings and combinations on the test set for both subtasks

Given the word embeddings size (300, 600 and 900, depending on the experiment), the neural network was changed to have a token LSTM hidden layer of the same size. Other hyper-parameters, common to all experiments, are: character embedding of size 25, learning rate of 0.005, dropout rate 0.5 and early stopping if no improvement was achieved for 10 epochs.

Experiments show that given the recurrent neural architecture used, the best single embeddings results, considering overall F1 score, for both subtasks are provided by the Wikipedia embeddings (a general domain representation). However, the Medical Embeddings seem to achieve higher precision. Considering the NER task, the combination of Wikipedia and Twitter achieves the highest F1 from the two embeddings experiments, while the three embeddings combination provides the final best score.

For the first subtask we used the predictions given by a NER model and considered a tweet with at least one recognized entity to belong to the domain required by the subtask. In order to improve recall we further extracted a list of professions from the training set of the NER subtask. This list was filtered and we removed strings that tend to appear many times in tweets labelled "0" in the training set belonging to the classification task. The filtered list was applied in addition to the NER information and texts that had no extracted entities were labelled "1" if they contained any string from the list. This allowed us to further increase the classifier's performance.

3 Conclusions

We investigated the suitability of different representations for analysing text from the health domain in social media, particularly Twitter messages.

Contrary to our initial assumption, a general domain representation (Wikipedia based) provided the best NER results, considering single representations. However, a combination of word embeddings achieved the highest F1 score. For both validation and test datasets, the best models considering F1 are a combination of Twitter and Wikipedia for the NER task and a combination of all three models for the classification task. We consider this to be explainable by the characteristic of social media messages where people do not necessarily restrict their language to in-domain vocabulary (in this case health related) but rather mix in-domain messages with out-of-domain messages or even combine in the same message sentences from multiple domains.

Acknowledgements

This research was supported by the EC grant INEA/CEF/ICT/A2018/28592472 for the Action No: 2019-EU-IA-0034 entitled "Curated Multilingual Language Resources for CEF.AT" (CURLICAT) and by the European Language Equality (ELE) project.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. arXiv:1607.04606.

Franck Dernoncourt, Ji Young Lee, and Peter Szolovits. 2017. NeuroNER: an easy-to-use program for named-entity recognition based on neural networks. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, pages 97–102, Copenhagen, Denmark. Association for Computational Linguistics.

Maryam Habibi, Leon Weber, Mariana Neves, David Luis Wiegandt, and Ulf Leser. 2017. Deep learning with word embeddings improves biomedical named entity recognition. *Bioinformatics*, 33(14):i37–i48.

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*.

Tomas Mikolov, Edouard Grave, Piotr Bojanowski, Christian Puhrsch, and Armand Joulin. 2018. Advances in pre-training distributed word representations. In *Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018)*.

Antonio Miranda-Escalada, Marvin Aguero, and Martin Krallinger. 2021a. Spanish covid-19 twitter embeddings in fasttext. *Zenodo*.

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*.

Vasile Pais and Dan Tufiș. 2018. Computing distributed representations of words using the corola corpus. *Proceedings of the Romanian Academy, Series A*, 19(2):403–410.

Felipe Soares, Marta Villegas, Aitor Gonzalez-Agirre, Martin Krallinger, and Jordi Armengol-Estapé. 2019. Medical word embeddings for Spanish: Development and evaluation. In *Proceedings of the 2nd Clinical Natural Language Processing Workshop*, pages 124–133, Minneapolis, Minnesota, USA. Association for Computational Linguistics.