Occupation Recognition and Normalization in Clinical Notes

Kaushik Acharya[1][0000−0003−2759−3646]

Philips India Limited, Bangalore, Karnataka 560045, India
acharya.kaushik@gmail.com

Abstract. This paper describes the system submitted in MEDical DOcuments PROFessions recognition (MEDDOPROF) shared task, which is part of Iberian Languages Evaluation Forum (IberLeF) 2021 tasks. The Named Entity Recognition (NER) model built using Conditional Random Field, detects occupation and employment status entities in the Spanish medical documents. The entities are mapped to their code using vector embedding similarity of mention text with the code label text. The model obtains F-score of 0.635 for NER and 0.566 for the normalization task.

Keywords: Named Entity Recognition · Entity Linking · Vector Embedding Similarity.

1 Introduction

There have been several studies [1,10] showing correlation of socio-demographic factors on physical and mental health, habits and lifestyle choices. Occupation and employment status are among the prime factors. Therefore, it becomes important to automatically identify its mention in clinical free text.

The MEDDOPROF shared task [7] focuses on the automatic detection of these factors as well as their mapping to standard code in medical documents.

MEDDOPROF shared task comprised of three tracks:

- Track 1: **MEDDOPROF-NER**
- Track 2: **MEDDOPROF-CLASS**
- Track 3: **MEDDOPROF-NORM**

MEDDOPROF-NER Track 1 is a named entity recognition problem which requires finding mentions of occupations and classifying them as:

- *Profession*: label PROFESION.
- *Employment Status*: label SITUACION_LABORAL.
- *Activity*: label ACTIVIDAD.

IberLEF 2021, September 2021, Málaga, Spain.

MEDDOPROF-CLASS Track 2 requires determining to whom the occupation mention belongs to:

- *Patient*: label PACIENTE.
- *Family Member*: label FAMILIAR.
- *Health Professional*: label SANITARIO.
- *Someone else*: label OTROS.

MEDDOPROF-NORM Track 3 is an entity linking problem which requires mapping predicted entities to the codes. The codes are unique concept identifiers from

- European Skills, Competences, Qualifications and Occupations (ESCO)
- SNOMED-CT

The shared task consists of clinical text written in Spanish. As per Instituto Cervantes's 2019 yearbook, Spanish is the world's second-most spoken native language with 480+ million native speakers (https://www.languagemagazine.com/2019/11/18/spanish-in-the-world/). According to their 2014 report (https://en.wikipedia.org/wiki/List_of_countries_where_Spanish_is_an_official_language), it is the official language for 20 sovereign states and one dependent territory, totaling population around 442 million. Hence, building NLP systems for Spanish can have a significant societal impact.

System described in this paper participated in MEDDOPROF-NER and MEDDOPROF-NORM. Source code has been shared on github (https://github.com/kaushikacharya/clinical_occupation_recognition).

2 Model Description

2.1 Data

The corpus was annotated with profession and employment status in BRAT Standoff format (https://brat.nlplab.org/standoff.html) by a team composed of linguists and clinical experts. These clinical cases were sourced from different specialties.

MEDDOPROF-NER For each clinical case, clinical note is stored in text file (.txt) and annotation(s) in annotation file (.ann). An example BRAT annotation is shown in Fig. 1.

MEDDOPROF-NORM Code corresponding to each of the annotated entity in train set was provided in a tab-separated file (.tsv). Additionally, a code reference list was provided with code and its corresponding label and alternative label(if available). Primarily, professions are mapped to ESCO while working statuses and activities are mapped to SNOMED-CT.

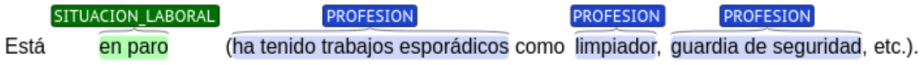

Fig. 1. Example BRAT annotation with profession and employment status labels. *English Translation*: He is unemployed (he has had sporadic jobs as a cleaner, security guard, etc.).

2.2 Named Entity Recognition

A Linear-chain Conditional Random Fields (CRF) [5] classifier was trained to recognize the named entities. Parameter estimation is done using Limited-memory BFGS (L-BFGS) [8] optimization algorithm. L-BFGS belongs to the family of quasi-Newton methods that approximates the Broyden–Fletcher–Goldfarb–Shanno algorithm (BFGS) using a limited amount of computer memory. The classifier is trained with both L1 and L2 regularization (coefficient values for both as 0.1). The CRF model is implemented using sklearn-crfsuite, which is a wrapper over CRFsuite [9].

Features Features are extracted using spaCy [3], an open source software library for natural language processing, written in Python and Cython. *es_core_news_sm* (https://spacy.io/models/es) is the model used that was trained on news and media genre text. These features can be categorized as follows:

- **Lexical features**
  - Unigrams (Current and immediate neighbor words)
  - String case
- **Parts of speech (POS) features**
  - Current word's POS
  - Prev and Next word's POS
  - Governor word's POS
- **Dependency parse features**
  - Governor words
  - Dependency type of current word
  - Dependency type of Governor word

Example In the sentence shown in Fig. 2, **seguridad** produces the following features:

- **current word POS**: NOUN
- **dependency tag**: nmod
- **parent dependency tag**: appos

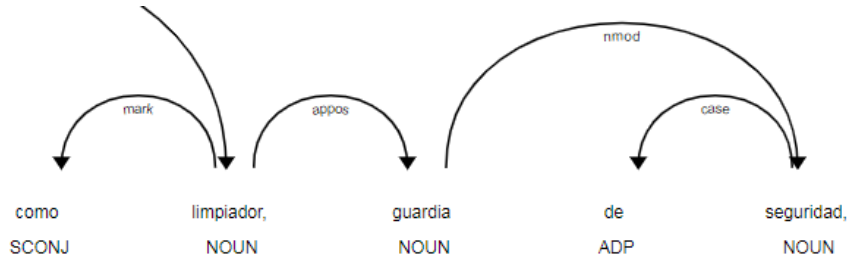

Fig. 2. POS & Dependency parse for the sentence in Fig. 1

2.3 Entity Linking

Mapping predicted entities to its relevant code has been solved using the vector embedding similarity approach. Vector embedding for the entities and text corresponding to the codes have been generated using fastText's trained model [2] for Spanish. For each predicted entities, code is assigned which has the highest cosine similarity.

3 Experimental Setup

3.1 Data

Train test split of corpus for the shared task is displayed in Table 1. Count statistics of the entities is displayed in Table 2.

Table 1. Dataset count statistics.

Dataset	Clinical Cases	Sentences
Training	1500	49932
Test	344	9671

Table 2. Count statistics of the entities.

Occupation Entity	Count(Train set)	Count(Test set)
PROFESION	2513	693
SITUACION_LABORAL	1010	356
ACTIVIDAD	109	28

3.2 External libraries

The system utilized the following libraries:

- fastText (version: 0.9.2) (https://github.com/facebookresearch/fastText)
- sklearn-crfsuite (version: 0.3.6) (https://sklearn-crfsuite.readthedocs.io/)
- spaCy (version: 3.0.5) (https://github.com/explosion/spaCy)

Their usage have been explained in Section 2.2, 2.2 and 2.3.

3.3 Evaluation Metrics

For both MEDDOPROF-NER and MEDDOPROF-NORM tasks; *precision*, *recall* and *F-score* are computed for each of the clinical case. These metrics are then summarized over the corpus using micro-average.

4 Results

4.1 Quantitative Findings

Table 3 shows the micro-average metrics for MEDDOPROF-NER on both train and test set using the evaluation script provided by the task organizers. Results for MEDDOPROF-NORM is shown in Table 4.

Table 3. MEDDOPROF-NER Micro-average metrics.

Metrics/Dataset	**Training**	**Test**
Precision	0.953	0.807
Recall	0.839	0.524
F-score	0.892	0.635

Table 4. MEDDOPROF-NORM Micro-average metrics.

Metrics/Dataset	**Training**	**Test**
Precision	0.956	0.720
Recall	0.840	0.467
F-score	0.894	0.566

4.2 Error Analysis

MEDDOPROF-NER Table 3 shows the comparison of metrics when model trained on training set is applied on the same set and unseen test set. Though its expected that performance will decrease when run on test set, but the drop in recall is almost double the drop in precision. The primary reason for low recall is in model's failure to detect the entity mentions itself. Table 6 shows the count of instances where model succeeded/failed in identifying entity mentions *irrespective* of the three entity classes.

Description of the match types:

- *Exact Match*: Predicted entity mentions text span matched exactly with ground truth.
- *Partial Match*: Predicted entity mentions text span matches partially with ground truth.
- *False Negative*: Ground truth entity mentions where model failed to generate any of the entity classes.
- *False Positive*: Model generated entity mentions where there's no entity class as per ground truth.

Around 37% of the ground truth entities falls under false negative. False negative/positive cases are the ones where there's not even partial match between ground truth and predicted entities.

The official scores shown in Table 3 is based on strict evaluation setting. This would show *Partial Match* entities as failed. Table 5 shows a granular analysis at token level. B-Entity stands for first token of the entity. I-Entity stands for rest of the entity tokens.

Table 5. MEDDOPROF-NER token level metrics on test set. Produced using seqeval library (https://github.com/chakki-works/seqeval)

	Precision	**Recall**	**F-score**	**Support**
B-ACTIVIDAD	0.818	0.321	0.462	28
I-ACTIVIDAD	0.867	0.433	0.578	60
B-PROFESION	0.923	0.657	0.767	693
I-PROFESION	0.825	0.616	0.705	1148
B-SITUACION_LABORAL	0.790	0.444	0.568	356
I-SITUACION_LABORAL	0.760	0.446	0.562	491
O	0.995	0.999	0.997	216660
accuracy			0.993	219436
macro avg	0.854	0.559	0.663	219436
weighted avg	0.993	0.993	0.993	219436

Table 6. Entity mentions detection in test set.

Match type	Count
Exact Match	579
Partial Match	102
False Negative	405
False Positive	38

MEDDOPROF-NORM As the output of MEDDOPROF-NER is fed as input for MEDDOPROF-NORM, the corresponding metrics for MEDDOPROF-NORM performs poorer. Improvement in MEDDOPROF-NER would get reflected in MEDDOPROF-NORM.

5 Conclusion

This paper proposes a CRF-based named entity extraction, and a vector embedding similarity based entity linking.

Future plans

- Extract global structured information features for the dependency parse tree [4].
- Develop LSTM-CRF model [6] which would automate the feature extraction.

Around 9.4% of the ground truth entities fall under *Partial Match*. [4] defines *valid span* as a word sequence that is covered by a chain of dependency parse arcs where no arc is covered by another. This enables extraction of dependency parse based global structured information rather than only local features as mentioned in 2.2. As per this definition, the entity *guardia de seguridad* shown in Fig 1 can be considered as a valid span. This can be seen in Fig 2. Using global structured information features would hopefully produce correct entity mention span.

The significant drop for both precision and recall on unseen test data compared to seen training data, shows that there's need for better features. Hence plan to develop LSTM model for improved feature extraction.

References

1. Fisher, K., Griffith, L., Gruneir, A., Upshur, R., Perez, R., Favotto, L., Nguyen, F., Markle-Reid, M., Ploeg, J.: Effect of socio-demographic and health factors on the association between multimorbidity and acute care service use: population-based survey linked to health administrative data. BMC Health Services Research **21**, 62 (01 2021). https://doi.org/10.1186/s12913-020-06032-5
2. Grave, E., Bojanowski, P., Gupta, P., Joulin, A., Mikolov, T.: Learning word vectors for 157 languages. In: Proceedings of the International Conference on Language Resources and Evaluation (LREC 2018) (2018)

3. Honnibal, M., Montani, I., Van Landeghem, S., Boyd, A.: spaCy: Industrial-strength Natural Language Processing in Python (2020). https://doi.org/10.5281/zenodo.1212303
4. Jie, Z., Muis, A.O., Lu, W.: Efficient dependency-guided named entity recognition. In: Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence. p. 3457–3465. AAAI'17, AAAI Press (2017)
5. Lafferty, J.D., McCallum, A., Pereira, F.C.N.: Conditional Random Fields: Probabilistic Models for Segmenting and Labeling Sequence Data. In: Proceedings of the Eighteenth International Conference on Machine Learning. p. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001)
6. Lample, G., Ballesteros, M., Subramanian, S., Kawakami, K., Dyer, C.: Neural Architectures for Named Entity Recognition. In: Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 260–270. Association for Computational Linguistics, San Diego, California (Jun 2016). https://doi.org/10.18653/v1/N16-1030
7. Lima-López, S., Farré-Maduell, E., Miranda-Escalada, A., Brivá-Iglesias, V., Krallinger, M.: Nlp applied to occupational health: Meddoprof shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. Procesamiento del Lenguaje Natural **67** (2021)
8. Nocedal, J.: Updating quasi-newton matrices with limited storage. Mathematics of Computation **35**(151), 773–782 (1980), http://www.jstor.org/stable/2006193
9. Okazaki, N.: Crfsuite: a fast implementation of Conditional Random Fields (CRFs) (2007), http://www.chokkan.org/software/crfsuite/
10. Park, S., Jeon, H.J., Kim, J.U., Kim, S., Roh, S.: Sociodemographic factors associated with the use of mental health services in depressed adults: Results from the korea national health and nutrition examination survey (knhanes). BMC health services research **14**, 645 (12 2014). https://doi.org/10.1186/s12913-014-0645-7