Everything Transformers: Recognition, Classification and Normalisation of Professions and Family Relations

Salvador Medina Herrera[1][0000−0003−2473−8571] and Jordi Turmo Borràs[1][0000−0002−7521−1115]

Universitat Politècnica de Catalunya, Campus Nord, Carrer de Jordi Girona, 1, 3, 08034 Barcelona, Spain
{smedina,turmo}@cs.upc.edu

Abstract. This document describes the system submitted by TALP team for IberLEF 2021's MEDDOPROF Shared Task. The joint occupation mention identification and family relation classification model is composed of a pre-trained DistilBERT architecture followed by a Bidirectional LSTM layer. Occupation normalisation uses Sentence-BERT pre-trained for Semantic Text Similarity (STS) to map the ESCO and SNOMED-CT categories as well as the mentions of occupations from the documents to a vectorial space. K-nearest neighbours is then used to find the most likely category assignments.

Keywords: NER · SNOMED-CT Normalisation · DistilBERT · Sentence-BERT

1 Introduction

We present the work carried out by the TALP Team in the context of IberLEF 2021's MEDDOPROF Shared Task [3]. The system is composed of two independent sub-systems: a Named Entity Recognition and Classification (NERC) model that handles the occupation mention detection (Track 1 - MEDDOPROF-NER) as well as the family relation classification (Track 2 - MEDDOPROF-CLASS) tasks, and a sentence-embedding model that tackles the normalization task (Track 3 - MEDDOPROF-NORM). An in-depth description of these two models can be found in Section 2.

Due to the provided training dataset being relatively small for some occupation classes such as activities and also fairly unbalanced compared to professions, we also use a simple data augmentation algorithm. Similarly to the rare word substitution approach from Fadaee et Al. [2], we up-sample these entities by replacing other entities and hence adding more contexts. These new examples are then scored by using a general-purpose BERT Language Model to discard unlikely examples.

IberLEF 2021, September 2021, Málaga, Spain.

2 Systems Description

As already mentioned in Section 1, our system is comprised of two independent components. In this section, we give an in-depth look at them and describe the data augmentation approach that we followed.

We reserved a 10% of the documents from the shared tasks' training set for validation and the remaining 90% for training, using the same split for all three sub-tasks. The batch-wise evaluation for Tracks 1 and 2 was performed at the level of tokens rather than entities, since it is simpler and more efficient to compute in GPU. A full entity-wise evaluation was performed after the training process had ended.

2.1 Occupation Identification and Family Relation Classification

Even though the Shared Task splits the occupation mention identification and family relation classification into two different tracks, we approached these two tasks as a single joint task. We considered two alternatives: either using a single output whose set of labels is the result from the cross-concatenation of the occupation classes and the family relation classes or using two independent outputs for occupations and family relations. We used binary cross-entropy as the loss function for all outputs. In the two output case, the combined loss was computed as an average of the two independent output losses. After some preliminary evaluations, we discarded the single output model, since the increased number of classes led to a degradation of around 10% in token-wise F_1 score.

Output Encoding As for output encoding, we considered In-Out (IO) and Begin-In-Out (BIO) encodings. We didn't see any noticeable improvements in the two-output model when using BIO, so we opted for using IO encoding. Similarly, and due to the fact that BERT uses word-piece encodings, that is, a sub-word representation of the tokens; very few input sequences were only one word-piece in length, so we also discarded other widely-used output encodings such as Begin-In-Out-Unitary (BIOU).

Model Structure Our NERC model structure is shown in Figure 1. Input text is encoded using word-piece embeddings and then fed to the input transformer of a DistilBERT model [5]. The output of the last tranformer layer of the DistilBERT model is then fed to a Bidirectional LSTM layer. Up to this point, the architecture is common to the occupation mention identification and relation classification tasks. The outputs from the BiLSTM layer are then fed to two independent time-distributed fully connected layers, one for each output.

We initialise the weights of the DistilBERT layers from a pre-trained general-purpose multi-lingual model from Huggingface (distilbert-base-multilingual-cased pre-trained model, a distilled version of bert-base-multilingual-cased with 6 layers, 768 dimensions per layer, 12 attention heads and 134M parameters.). We opted for DistilBERT instead of the original BERT [1] model because in our

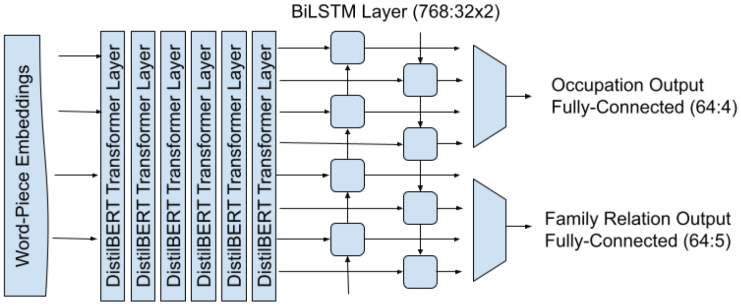
BiLSTM Layer (768:32x2)
Word-Piece Embeddings
DistilBERT Transformer Layer
DistilBERT Transformer Layer
DistilBERT Transformer Layer
DistilBERT Transformer Layer
DistilBERT Transformer Layer
DistilBERT Transformer Layer
Occupation Output
Fully-Connected (64:4)
Family Relation Output
Fully-Connected (64:5)

Fig. 1. Visual representation of the occupation and family relation identification system. The input and output sizes are shown in parenthesis.

tests there was no visible degradation in F_1 score yet it greatly reduced training time.

Training and Fine-tuning As for the training process, we explored several strategies that are worth mentioning.

First of all, we added configuration options to tweak the balance between positive, that is, sequences containing at least one valid entity; and negative, sequences with no entities. Note that even though all documents contain entities, due to computational limitations, we split the documents into overlapped sequences of up to 128 tokens. We saw that no matter the learning rate, training solely with positive examples led to very low precision while training with the raw documents led to low recall especially for under-represented classes. Our final training strategy was to interleave epochs limited to positive examples with full epochs while decreasing the learning rate each epoch.

As it was mentioned in Section 2.1, we initialise the weights from a pre-trained DistilBERT model. We also explored several alternatives for fine-tuning these weights to our task. More in particular, we froze a number of these layers during all or some training epochs. In our final submission, we included two options: either no fine-tuning or full fine-tuning of all DistilBERT weights.

Data Augmentation The labels in the training corpus for Track 1 of the MEDDOPROF Shared Task are fairly unbalanced: 2528 entities for *PROFESION*, 1011 for *SITUACION_LABORAL* status and just 122 for *ACTIVIDAD*. In order to improve F_1 score for the under-represented class, we decided to double that amount by taking examples in the training set and replacing their entities by other activities that are less common in the training set. In order not to add grammatically or semantically incorrect examples such as replacing *"Aunque **acudía con frecuencia al gimnasio**, y practicaba fundamentalmente pesas."* by *"Aunque **tocar la guitarra**, y practicaba fundamentalmente pesas"* (real example from the dataset), we used a general purpose BERT Language Model to compute a likelihood score so as to rank the synthetic examples (The default bert-base-multilingual-cased pre-trained model.).

2.2 Occupation Normalization

Our occupation normalization system is based on Semantic Text Similarity (STS) between sentences. The general idea is to put every occupation in a generic context and embed the whole sentence using a language model pre-trained for STS. These sentence embeddings are then associated with their respective ESCO or SNOMED-CT identifier. The inference is then performed by mapping the candidate occupation to the aforementioned vectorial space and then assigning the closest pre-mapped point's identifier. A visual representation of the inference process is shown in Figure 2.

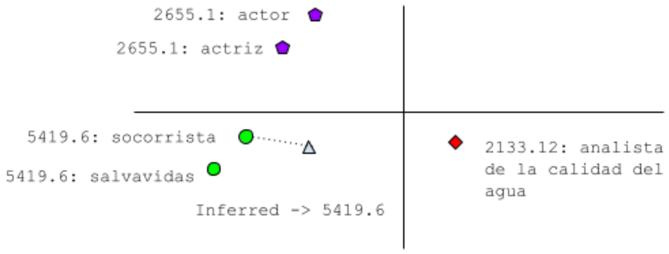

Fig. 2. Visual representation of the mapping and inference processes of the Occupation Normalization system. In this example, the target occupation (triangle) is assigned the identifier *5419.6* the same as its closest mapped definition (*socorrista*, circle).

We used both the provided *meddoprof_valid_codes.tsv* file and examples from the gold-standard training corpus as pre-mapped points in the vectorial space. We also added additional points computed as the centroid of subsets of occupation descriptions that were mapped to the same identifier. For example, *"operador de centrifugadora", "operaria de centrifugado"* and *"responsable de centrifugación"* are all associated to the identifier *8160.14* so in addition to all three points, their geometric center is also added.

We defined two generic contexts in which the occupation descriptions are introduced: *"Trabaja de OCUPACION."* y *"Se dedica a OCUPACION."*, where OCUPACION is replaced by the occupation's description.

Among all the publicly available pre-trained STS models, we opted for DistilUSE (We used the pre-trained distiluse-base-multilingual-cased-v2 multilingual model) [4], a lightweight distilled version of the Universal Sentence Encoder [6].

3 Task Results

We submitted three models for evaluation at the MEDDOPROF Shared Task: *default*, with full fine-tuning of DistilBERT's weights but no data augmentation; *extended*, with the same parameters as the former but adding examples from data augmentation; and *no_fine_tune*, with data augmentation but no fine-tuning. We used the same occupation normalization system for all three submissions.

In our tests with the training and validation split described in Section 2, the best performing model of the three was *extended*, which achieved 0.72 in F_1 score for Track 1, compared to *default*'s 0.705. This advantage funnels to Tracks 2 and 3. The results in test corpus show a different picture though: *default* achieves 0.698, outperforming *extended* by 0.027 (0.671).

If we look at precision and recall, we can see that *extended* balances precision and recall (0.671 and 0.671 respectively) whereas *default* favours precision over performance (0.761 and 0.645). This was to be expected, as adding synthetic examples often leads to the inclusion of noisy examples.

Sadly, at the time of writing this system description paper, the competition's final results have not yet been made public and no competitive comparisons can be made. We will then explore those results at a later date.

4 Conclusions

This paper describes the participation of the TALP team in IberLEF 2021's MEDDOPROF Shared Task. Our system makes extensive use of BERT-like language models for all three tracks of the challenge. We present effective strategies to deal with the dataset's data imbalance either with training scheduling or data augmentation, although this might not be reflected in the final score as the Shared Task's evaluation framework does not provide independent evaluation of the classes. This data augmentation strategy not only benefits under-represented occupation classes but also makes our system more balanced with regards to precision and recall.

It has not been possible to do a qualitative analysis of our system compared to the rest of the participants, nor an in-depth error analysis of our system due to the tight schedule. However, we strongly believe that the simple approach presented in this document has clear potential and could be improved and extended upon for the next iterations of this Shared Task.

5 Acknowledgments

This contribution has been funded by the Spanish Ministry of Science and Innovation (MICINN) (PID2019-106942RB-C33).

References

1. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: Bert: Pre-training of deep bidirectional transformers for language understanding. arXiv preprint arXiv:1810.04805 (2018)
2. Fadaee, M., Bisazza, A., Monz, C.: Data augmentation for low-resource neural machine translation. arXiv preprint arXiv:1705.00440 (2017)
3. 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)
4. Reimers, N., Gurevych, I.: Making monolingual sentence embeddings multilingual using knowledge distillation. arXiv preprint arXiv:2004.09813 (2020)
5. Sanh, V., Debut, L., Chaumond, J., Wolf, T.: Distilbert, a distilled version of bert: smaller, faster, cheaper and lighter. arXiv preprint arXiv:1910.01108 (2019)
6. Yang, Y., Cer, D., Ahmad, A., Guo, M., Law, J., Constant, N., Abrego, G.H., Yuan, S., Tar, C., Sung, Y.H., et al.: Multilingual universal sentence encoder for semantic retrieval. arXiv preprint arXiv:1907.04307 (2019)