Vicomtech at MEDDOPROF: Automatic Information Extraction and Disambiguation in Clinical Text

Elena Zotova[1,2][0000−0002−8350−1331], Aitor García-Pablos[1][0000−0001−9882−7521], and Montse Cuadros[1][0000−0002−3620−1053]

[1] SNLT group at Vicomtech Foundation,
Basque Research and Technology Alliance (BRTA),
Mikeletegi Pasealekua 57, Donostia/San-Sebastián, 20009, Spain
{ezotova,agarciap,mcuadros}@vicomtech.org
[2] Department of Languages and Computer Systems. University of the Basque Country (UPV-EHU), Leioa, Spain

Abstract. This paper describes the participation of the Vicomtech NLP team in the MEDDOPROF shared task. The challenge consists in automatic detection of occupations and employment status, as well as their normalization or entity mapping, within medical documents in Spanish language. The competition is split into three tasks, NER, CLASS and NORM. We have participated using a multitask joint model based on Transformers, which tries to solve all the three tasks at once. However, the NORM task, which consists on disambiguation of the detected entities against thousands of different possible codes, can be solved more effectively using other approaches. Because of that, we have submitted an additional sequence-to-sequence based approach and a semantic-search based approach to deal with the NORM task. We achieve a 77% of F1-score for the NER task, and 70% of F1-score for the CLASS task, and a 48% of F1-score for the NORM task.

Keywords: Clinical Text · Information Extraction · Automatic Indexing

1 Introduction

This article presents the participation of the Vicomtech NLP team in the MEDDOPROF Shared Task: Medical Documents Profession Recognition shared task [7]. The shared task consists in developing systems for automatic detection of occupations and employment status, as well as their normalization or entity mapping, within medical documents in Spanish language. The target data consists in a corpus of clinical case reports from heterogeneous medical specialities.

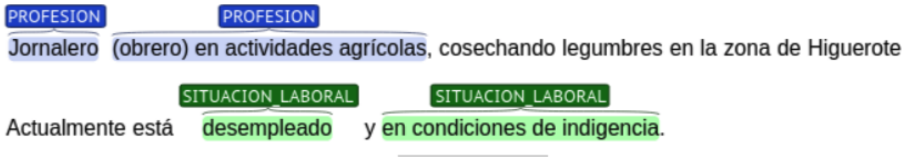
IberLEF 2021, September 2021, Málaga, Spain.

The competition is divided into three tasks. The first task, NER, requires automatically finding mentions of occupations and classifying each of them as a profession, an employment status or an activity. The second task, CLASS, requires classifying mentions of occupations to determine whether they are related to the patient, to a family member, to a health professional or to someone else. Finally, the third task, NORM, requires mapping the task 1 predictions to one of the codes in a list of unique concept identifiers. We refer the reader to the shared task overview article [7] for more detailed information about MEDDOPROF.

The rest of the document is structured as follows. Section 2 introduces the data provided by the organizers of the challenge. Sections 3 and 4 describe our submitted systems and the training setup, respectively. Section 5 presents the official results. In section 6, we discuss some decisions taken during the development and training phases, inherent flaws of our systems, and potential improvements. Finally, section 7 provides some concluding remarks and future work hints.

2 Data description

The provided corpus is a collection of 1844 clinical cases from over 20 different specialties annotated with professions and employment statuses. The gold annotations for NER and CLASS are provided in Brat format [12] (see Figure 1), while the codes for the NORM task are provided as a .tsv file with codes assigned to each profession/activity in the corpus (see Table 1). It must be noted that, in this regard, the NER and NORM tasks are related because the input for the NORM task are the entities detected in the NER task.



Jornalero (obrero) en actividades agrícolas, cosechando legumbres en la zona de Higuerote

Actualmente está desempleado y en condiciones de indigencia.

Fig. 1. Example of annotated clinical text for NER task.

Table 1. Example of training set for NORM task

filename	**text**	**span**	**code**
caso_clinico_psiquiatria95	haber dejado el ejército	2562 2586	SCTID: 73438004
caso_clinico_psiquiatria23	le ha llevado a despedirse	2127 2153	SCTID: 73438004
caso_clinico_urologia302	médico de Atención Primaria	185 212	2211.1

In addition, the organizers provided an extension of the dataset with an extra set of labels for different entities: symptoms, diseases, procedures, negation markers and negation spans, etc. The entities do not count toward the competition evaluation, but the organizers encourage the participants to make use of them to develop more interesting and complete systems.

The organizers also provide a list of valid codes related to professions, labour activities and occupations from SNOMED Clinical Terms (www.snomed.org) (50 codes) and European Skills, Competences, Qualifications and Occupations (ESCO) (ec.europa.eu/esco) classifications (3508 codes). Both SNOMED CT and ESCO are described as a machine-readable multilingual thesaurus with an ontological foundation.

The core concepts of the ontology are: concept codes—numerical codes that identify clinical terms, organized in hierarchies; descriptions—textual descriptions of concept codes; and relationships between concepts. SNOMED CT comprehensive coverage includes a large variety of concepts such as symptoms, diagnoses, procedures, body structures etc. The use of SNOMED CT within this competition is restricted to classifying activities and employment status. European Skills, Competences, Qualifications and Occupations (ESCO) is a multilingual classification of skills, competences, qualifications and occupations relevant for the EU labor market and education.

Table 2. Labels distribution of the official training set

Track	**Label**	**Count**
NER	PROFESION	2528
	SITUACION_LABORAL	1011
	ACTIVIDAD	119
CLASS	PACIENTE	1735
	SANITARIO	1231
	OTROS	485
	FAMILIAR	207
NORM	PROFESION (ESCO)	2501
	SITUACION_LABORAL (SNOMED CT)	1157
	Total unique SNOMED CT and ESCO codes	297

Table 2 shows the distribution of labels for each of the tasks. The competition does not provide an official training and development set partitions, so we have created them. We have split the official training data into a 90%/10% subsets for training and validating our systems respectively.

3 Systems description

We have approached the challenge as a joint multitask end-to-end model based on Transformers, trying to solve the three tasks at the same time. However, the

NORM task can be solved more effectively using other techniques and separated models, so we have competed with several different approaches for this third task.

3.1 Multitask joint model

The multitask joint model tries to solve all the tasks, including the detection of the extended entity set, using a single model based on transformers.

Except for the NORM task, the other tasks are treated as regular sequence-labelling tasks. At the core, there is a pre-trained BERT model that encodes the texts, converting each token into a contextual word-embedding. These embeddings are the base for several classification heads that perform a IOB tagging [10].

This regular sequence labelling approach solves the NER task and the CLASS task. However, the joint model also tries to deal with the NORM task treating with a hierarchical classification approach.

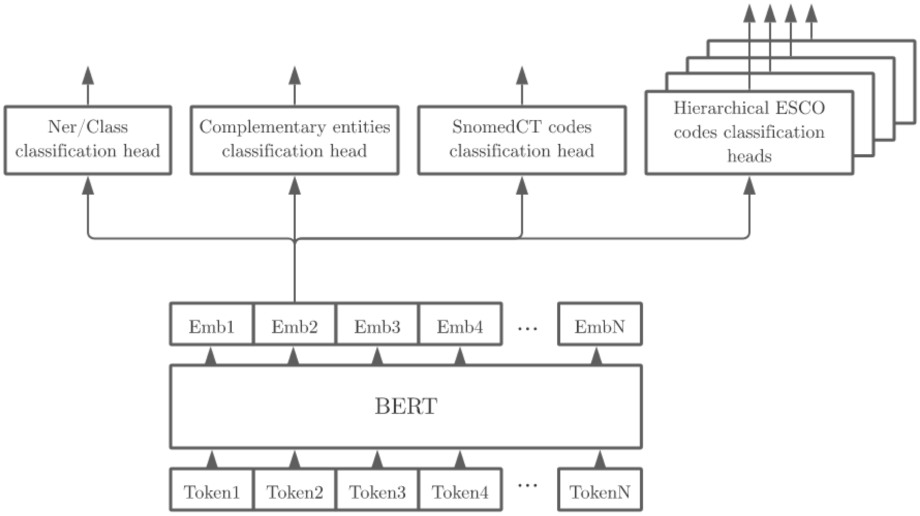

Ner/Class classification head
Complementary entities classification head
SnomedCT codes classification head
Hierarchical ESCO codes classification heads
Emb1
Emb2
Emb3
Emb4
...
EmbN
BERT
Token1
Token2
Token3
Token4
...
TokenN

Fig. 2. Architecture of the joint multitask model. The different entities from the different tasks are detected using a single classification head of top of each token. For the hierarchical ESCO codes of the NORM task a stack of classifiers, one per node of the resulting hierarchy are applied.

The hierarchical classification consists of a bunch of classification heads, one per each non-terminal node of the hierarchy, that are trained at the same time. The ESCO codes, used in this task to identify the professions, follow an hierarchical structure being the first digit the most coarse grained category. Each following digit adds a more fine-grained definition of the profession the code is describing. The key difference with a flat classifier is that, instead of trying to select a code from a flat list of potentially thousand of codes, a tree is built with the codes, level by level. Each node of the tree has only a limited amount of children nodes according to the actual ESCO hierarchy. These non-terminal nodes are turned into classifiers, and their children are the output size of each of those classifiers.

For the training, each ESCO code is decoded so only the appropriate classifiers have something to predict. The rest of the nodes are forced to predict a special "OUT" value. At inference time, the final code is reconstructed from the root of the hierarchy, classifier after classifier, following the hierarchy structure. The resulting code is emitted when the current classifier predicts the special value "OUT", or a leaf node (with no further children in the hierarchy) is reached.

This approach is suitable for this kind of task, but has several disadvantages. It is computationally expensive depending on the size of the hierarchy: for the ESCO codes involved in this competition it resulted in about 800 node-classifiers. Also, it is complex to implement, and each hierarchy may have subtleties that must be taken into account when modelling the tree structure and how the codes are encoded/decoded into a set of nodes. Finally, since there are a lot of nodes to train, the amount of training data available in the competition might not be enough.

Due to this reason, for the NORM task we have tried several additional approaches that are described in the next subsections.

3.2 Seq2Seq Translation System for NORM Task

The second approach to tackle NORM task is self-attention Transformer architecture [14]. We adapt sequence to sequence modelling (seq2seq) [13, 2] to the task of mapping terms from clinical texts to their codes in SNOMED CT and ESCO classifications. Term description is a source input for encoder and it's code in the corresponding ontology is a target input for decoder. High-level architecture of the system is depicted in Figure 3.

In order to train the mapping system, we prepare the training corpus as follows. The set of valid codes consists of 3.558 unique codes, some of them have various synonymous definitions, specifically, the number of synonyms varies from 1 to 38. We split all the multiple term definitions and assign a corresponding code to each synonym and get a dataset arranged as shown in Table 3. Here we can see that one code may be presented by various highly similar definitions.

Furthermore, we combine the training set of the NORM task and descriptions of all codes from the SNOMED CT and ESCO ontologies, which results in 297 unique codes with a distribution that ranges from 1 to 182 examples per code. Finally, we obtain 15.869 examples for train set and reserve 346 examples for development set (10% of original train set provided for the task). The dataset is highly unbalanced: only 10% of the codes has more than 10 examples per code.

During the text preprocessing step the source terms are lower-cased, cleaned from punctuation and tokenized on word-level. The target codes are not preprocessed, just tokenized by space. Number of tokens in source examples varies from 1 to 22, and in target set it is 1 or 2 depending on code type. We train the

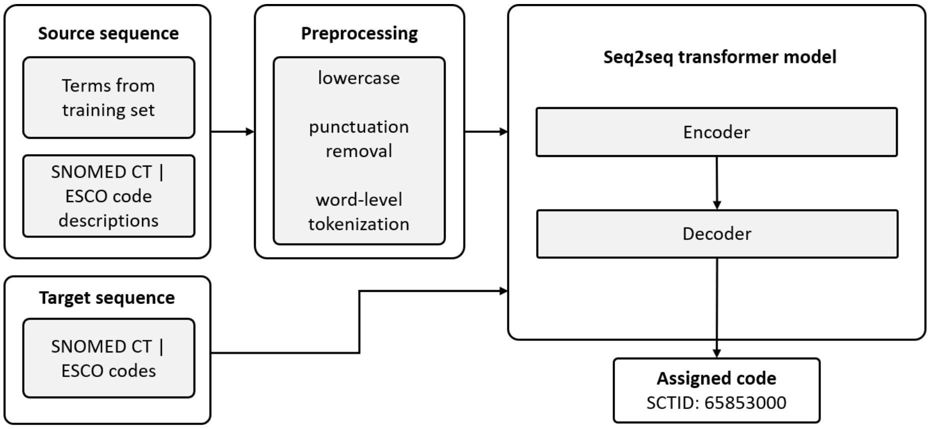


Fig. 3. Sequence to sequence architecture adapted to terms mapping.

Table 3. Examples of corpus from SNOMED ST and ESCO.

Description/Term	**Code**
jefe de producto de las TIC	1330.6
jefa de producto de las TIC	1330.6
encargado de la gestión de productos de las TIC	1330.6
encargada de la gestión de productos de las TIC	1330.6
product manager de las TIC	1330.6
jefa de producto de las TI	1330.6
jefa de producto de las TICs	1330.6
jefe de producto de las TI	1330.6
jefe de producto de las TICs	1330.6
desempleado	SCTID: 73438004
trabajador	SCTID: 106541005
refugiado	SCTID: 446654005

model with the parameters of the transformer architecture shown in the Table 4.

This approach has several advantages. First, it turns an extremely large multi-class classification problem into an straightforward sequence-to-sequence approach. Another advantage is that the pairs of codes and their descriptions, which are already defined in ontologies and vocabularies, can be leveraged as extra training instances complementing the actual training data.

3.3 Semantic similarity mapping for NORM Task

The following mapping system for NORM task is based on the idea of semantic search. Semantic search is an information retrieval method that leverages semantic similarity measure to retrieve semantically close documents. The main

Table 4. Transformer architecture parameters for code mapping.

Parameter	
encoder layers	6
attention heads	16
word vector size	512
gradient accumulation count	8
RNN size	512
batch size	4.096 tokens
training steps	10.000

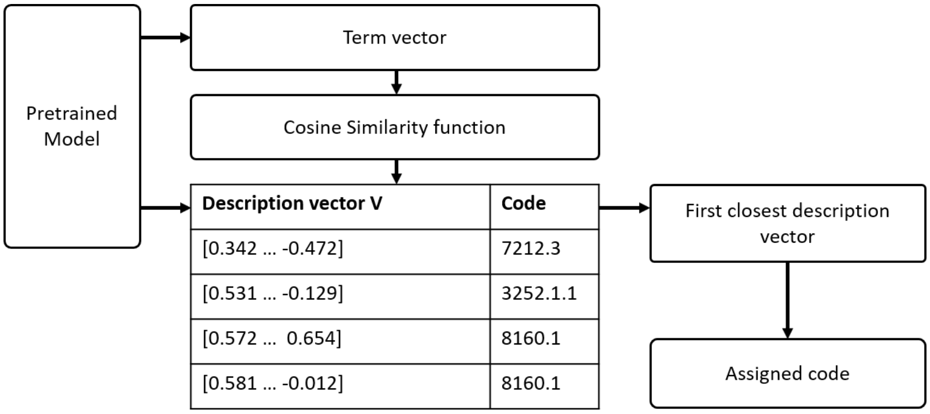
objective of semantic similarity is to measure the distance between the vectors that represent a pair of words, sentences, or documents. The key concepts of semantic search are the following: query, collection of documents, and degree of relevance between a query and retrieved documents.

We adapt the method to map terms written in natural language to the codes in SNOMED CT and ESCO classifications. In this case, a term previously detected by the NER system (see Subsection 3.1) as PROFESION, ACTIVIDAD o SITUACION_LABORAL is used as the query to search the closest document. The collection of documents is represented by SNOMED CT and ESCO ontologies provided by the organizers. The codes are separated by synonyms as shown in Table 3, so each code has various descriptions. The descriptions are the documents to search through. To compute a notion of similarity between a term and a code description we use the cosine distance.

We have experimented with different pretrained language models to create common vector space for terms and code descriptions. We have selected LaBSE model [5], because it obtained the best F-score during the experimentation. LaBSE is a BERT sentence embedding model supporting 109 languages. It is developed using masked language modelling [4] and translation language modelling [3] with a translation ranking task using bi-directional dual encoders.

Since the type of a term (i.e. whether it is a profession, and activity or an employment status) is detected in previous task, we execute the mapping process in two ways: 1) search in SNOMED CT and ESCO separately; 2) search in the database where SNOMED CT and ESCO codes are united. Figure 4 depicts the basic algorithm of semantic search applied to the NORM task independent from the database.

For the case of separate search, we select the closest description with the following condition: if assigned tag is SITUACION_LABORAL or ACTIVIDAD, the term is to be search in SNOMED CT database (50 codes), if the tag is PROFESION, the term is to be searched in ESCO database (3554 codes). In our experiments, the terms tagged as SITUACION_LABORAL, and thus mapped to the SNOMED CT codes, reached a micro-F1 score of 0.577, while and PROFESION terms mapped to ESCO obtained a micro-F1 score of 0.215. This suggests that, as could be expected, the performance of the semantic search method is



Description vector V	Code
[0.342 … -0.472]	7212.3
[0.531 … -0.129]	3252.1.1
[0.572 … 0.654]	8160.1
[0.581 … -0.012]	8160.1

Fig. 4. Algorithm of semantic search for the description closest to the term to map.

influenced by the number of target elements and to which extent they are semantically separable.

4 Training setup and submitted systems

We have participated in all the tasks proposed by the competition. The first two tasks, NER and CLASS, have been only dealt with the multitask joint model. For the NORM task we have submitted different runs using different approaches. The first approach is the same multitask model, since it aims to predict all the information requested in the competition in a single step.

In order to train the multitask joint model we used IXAmBERT [9] and BETO [1] as the pre-trained BERT models that form the core of the model. We have experimented with the two because both of them are pre-trained using Spanish data. After validation in the development set, BETO seemed to obtain a slight advantage, so finally we decided to make the submission using the BETO-based multitask model.

The multitask joint model has been implemented in Python 3.7 with HuggingFace's transformers library [15] (github.com/huggingface/transformers) and it has been trained on a Nvidia GeForce RTX 2080ti GPU with ∼11GB of memory. The learning rate was set to 2E-5 and the optimizer was AdamW [8]. During the training the micro-F1 score of the predictions on the development set was monitored, with 100 epochs of early stopping patience. That means that the model continued training until reaching 100 consecutive epochs without any improvement in the validation metric.

The NORM task Transformer model implemented in OpenNMT toolkit[6] and its PyTorch based framework OpenNMT-py (opennmt.net/OpenNMT-py). The model was trained on on a Nvidia GeForce RTX 2080ti GPU with ∼11GB of memory, with learning rate set to 2E-5 and the optimizer AdamW [8] during 10.000 steps. The best model was selected by the micro-F1 score.

Semantic similarity inference implemented with Sentence Transformers library [11] on a Nvidia GeForce RTX 2080ti GPU with ∼11GB of memory.

5 Results

Table 5 shows the results obtained in the official evaluation provided by the competition organizers. The metrics are micro-averaged precision, recall and F-score. The results are split by track and, for the case of the NORM task, by the different submitted runs. The table also shows the official baseline score for each track as they have been reported by the organizers. The baseline is based on a simple word lookup based on the training data annotations.

Table 5. Official evaluation results obtained in the test set for the different competition tracks and submitted systems

	System	Precision	Recall	F-score
NER	meddoprof-joint	**0.758**	**0.739**	**0.748**
CLASS	meddoprof-joint	**0.710**	**0.691**	**0.701**
NORM	meddoprof-joint	0.426	0.380	0.402
NORM	transformer-system	*0.488*	*0.474*	*0.481*
NORM	semantic-mapping	0.260	0.254	0.257
NORM	semantic-mapping-version2	0.254	0.247	0.250
NER	Official baseline	0.465	0.508	0.486
CLASS	Official baseline	0.391	0.377	0.384
NORM	Official baseline	**0.502**	**0.533**	**0.517**

The multitask-joint model performs reasonably well for NER and CLASS tasks. The F-scores scores for NER and CLASS tasks achieved by our multitask joint model are 25.6% and 31.7% above the baseline respectively. For NORM task the best performing system is the one that uses a sequence-to-sequence approach based on transformers.

The score for the NORM task, even for the best performing system, is below the baseline score. A possible explanation is that the most frequent codes are repeated a lot of times and the the baseline approach can easily find those common codes very straightforwardly.

Since the NORM task input is the output of the NER task, and the participants do not have access to any gold-labelled input for the NORM task, the competing systems need to rely on the imperfect outcomes of the corresponding NER system. This fact results in an error accumulation that lowers the final score. To clarify this point, it would be interesting to compare our results with other participants.

At the time of writing these working notes, the official ranking with the scores from all the participants has not been published yet, so we cannot assess to which extent our results are competitive.

6 Discussion

In order to better understand the behaviour and the result of some of our submitted systems, we have carried out some error analysis to pose some discussion points for future work.

In the NORM task we see the following challenging issues:

- The Seq2Seq Translation system (see Subsection 3.2) seems to be biased due to unbalanced dataset: some codes have only one description while the others have more than 130.
- Hierarchical structure of the ESCO classification and short descriptions lead to many semantically close terms that are labelled with different codes. This leads to codes that are "almost" correctly predicted, in the sense of that only the most fine-grained part of the code is incorrect. However this counts as an error regardless of how close the predicted code was from the correct one. For instance the term *"vendedora en un comercio pequeño"* ("salesperson in a small business") is manually labelled as code 5223 (Asistentes de venta de tiendas y almacenes - sales assistants of shops and warehouses) and the system predicts code 5223.7 (vendedor especializado/vendedora especializada - specialized salesperson).
- The Seq2Seq Translation system performance is highly influenced by hyper-parameters and other facts that deserve further experimentation.
- The Semantic mapping method presented in this article is straightforward and does not require previous training. However, the system fails mainly in mapping semantically close terms. It performs better when the search database is of moderate size and the documents are more semantically separable.

7 Conclusions

In these working notes we have presented Vicomtech's participation in MEDDOPROF shared task. We have participated with a multitask joint model based on Transformers, which solves the three tasks, NER, CLASS and NORM. In addition, we have presented another two systems to solve the NORM task. The multitask joint model works for the three tasks at the same time, although the NORM task can be better tackled using other approaches, such as using a sequence to sequence approach to map terms and codes. The quantitative results seem reasonable, but at the moment of this writing the official score ranking has not been published, so we cannot perform any comparison against other participants to conclude if our proposed systems are competitive or not.

All in all, the objective of the proposed tasks in relevant and interesting, and it is still far from being solved. In order to keep improving the results, apart from trying new approaches, more experimentation will be needed to improve some design decisions and chose better hyper-parameter settings that seem to highly influence the performance of the systems.

Acknowledgments

This work has been partially funded by the projects DeepText (KK-2020-00088, SPRI, Basque Government) and DeepReading (RTI2018-096846-B-C21, MCIU/AEI/FEDER, UE).

References

1. Cañete, J., Chaperon, G., Fuentes, R., Pérez, J.: Spanish Pre-Trained BERT Model and Evaluation Data. In: Proceedings of the Practical ML for Developing Countries Workshop at the Eighth International Conference on Learning Representations (ICLR 2020). pp. 1–9 (2020)
2. Cho, K., van Merriënboer, B., Gulcehre, C., Bahdanau, D., Bougares, F., Schwenk, H., Bengio, Y.: Learning Phrase Representations using RNN Encoder–Decoder for Statistical Machine Translation. In: Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP). pp. 1724–1734. Association for Computational Linguistics, Doha, Qatar (2014)
3. Conneau, A., Lample, G.: Cross-lingual Language Model Pretraining. In: Wallach, H., Larochelle, H., Beygelzimer, A., d'Alché-Buc, F., Fox, E., Garnett, R. (eds.) Advances in Neural Information Processing Systems. vol. 32. Curran Associates, Inc. (2019)
4. Devlin, J., Chang, M.W., Lee, K., Toutanova, K.: BERT: Pre-training of Deep Bidirectional Transformers for Language Understanding. In: Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers). pp. 4171–4186 (2019)
5. Feng, F., Yang, Y., Cer, D., Arivazhagan, N., Wang, W.: Language-agnostic BERT Sentence Embedding (2020)
6. Klein, G., Kim, Y., Deng, Y., Senellart, J., Rush, A.: OpenNMT: Open-Source Toolkit for Neural Machine Translation. In: Proceedings of ACL 2017, System Demonstrations. pp. 67–72. Association for Computational Linguistics, Vancouver, Canada (2017)
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. Loshchilov, I., Hutter, F.: Decoupled Weight Decay Regularization. In: Proceedings of the Seventh International Conference on Learning Representations (ICLR 2019). pp. 1–18 (2019)
9. Otegi, A., Agirre, A., Campos, J.A., Soroa, A., Agirre, E.: Conversational Question Answering in Low Resource Scenarios: A Dataset and Case Study for Basque. In: Proceedings of The 12th Language Resources and Evaluation Conference. pp. 436–442 (2020)
10. Ramshaw, L.A., Marcus, M.P.: Text Chunking Using Transformation-based Learning. In: Natural language processing using very large corpora, pp. 157–176. Springer (1999)
11. Reimers, N., Gurevych, I.: Making Monolingual Sentence Embeddings Multilingual Using Knowledge Distillation. In: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing. Association for Computational Linguistics (11 2020)

12. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: A Web-based Tool for NLP-assisted Text Annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics (EACL '12). pp. 102–107 (2012)
13. Sutskever, I., Vinyals, O., Le, Q.V.: Sequence to Sequence Learning with Neural Networks. In: Proceedings of the 27th International Conference on Neural Information Processing Systems - Volume 2. p. 3104–3112. NIPS'14, MIT Press, Cambridge, MA, USA (2014)
14. Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A.N., Kaiser, Ł., Polosukhin, I.: Attention Is All You Need. In: Proceedings of the Thirty-first Conference on Advances in Neural Information Processing Systems (NeurIPS 2017). pp. 5998–6008 (2017)
15. Wolf, T., Debut, L., Sanh, V., Chaumond, J., Delangue, C., Moi, A., Cistac, P., Rault, T., Louf, R., Funtowicz, M., Brew, J.: HuggingFace's Transformers: State-of-the-art Natural Language Processing. arXiv:1910.03771 pp. 1–11 (2019)