Boosting Transformers for Job Expression Extraction and Classification in a Low-Resource Setting

Lukas Lange[1,2,3], Heike Adel[1], and Jannik Strötgen[1]

[1] Bosch Center for Artificial Intelligence
Robert-Bosch-Campus 1, 71272 Renningen, Germany
{Lukas.Lange,Heike.Adel,Jannik.Stroetgen}@de.bosch.com
[2] Spoken Language Systems (LSV),
[3] Saarbrücken Graduate School of Computer Science
Saarland Informatics Campus, Saarland University, Saarbrücken, Germany

Abstract. In this paper, we explore possible improvements of transformer models in a low-resource setting. In particular, we present our approaches to tackle the first two of three subtasks of the MEDDOPROF competition, i.e., the extraction and classification of job expressions in Spanish clinical texts. As neither language nor domain experts, we experiment with the multilingual XLM-R transformer model and tackle these low-resource information extraction tasks as sequence-labeling problems. We explore domain- and language-adaptive pretraining, transfer learning and strategic datasplits to boost the transformer model. Our results show strong improvements using these methods by up to 5.3 F_1 points compared to a fine-tuned XLM-R model. Our best models achieve 83.2 and 79.3 F_1 for the first two tasks, respectively.

Keywords: Named Entity Recognition · Neural Sequence Tagging · Domain- and Language-adapted Language Models · Strategic Datasplits

1 Introduction

Information extraction in non-standard domains is a challenging problem due to the large number of complex terms and unusual document structures [4]. Despite this, pretrained transformer models demonstrated robustness across languages and domains. However, these models still show their best performance when applied to targets similar to their pretraining corpora which can limit their applicability in many situations [7]. One example for this is the Spanish clinical domain, where both, language and domain, can be considered a non-standard setting in the English-centric NLP community [15].

In this paper, we explore possible enhancements of transformer models to overcome this domain and language gap in the context of the MEDDOPROF shared task [14]. In particular, we participate in the first two tasks of MEDDOPROF [14], a challenge concerned with the extraction, classification and normalization of job-related expressions in Spanish clinical texts. The first task *NER* requires the extraction of three different kinds of occupation and the second task *CLASS* demands to classify each of the previously extracted occupations into four classes reflecting the holder of that job.

We approach this challenge as NEITHER LANGUAGE NOR DOMAIN EXPERTS (NLNDE) and model them as sequence labeling tasks. Our solution for these tasks is a neural sequence tagger based on multilingual transformer models. In particular, we experiment with continuing the masked language modeling pretraining of the multilingual XLM-R model [3] on Spanish texts, transferring trained models between the two tasks [13] and using strategic datasplits [18].

Our results highlight the importance of domain- and language-adapted transformer models, as well as the advantages of combining several models trained on challenging datasplits with ensembling techniques. Using these methods, our best models achieve F_1-scores of 83.2 and 79.3 for the two tasks and outperform a fine-tuned XLM-R model by 4.2 and 5.3 F_1 points, respectively.

2 Related Work

The MEDODPROF challenge follows a series of shared tasks on Spanish clinical information extraction, including the MEDDOCAN shared task on medical document anonymization [15] and the PharmaCoNER shared task on concept extraction [5, 10]. Main findings of all of these challenges were that transformer models become more commonly used [5, 15] as they begun to dominate the field of information extraction due to their general applicability across languages and domains. For an overview of recent approaches to low-resource NLP, we refer the refer to [8].

As the inclusion of domain knowledge via domain-specific embeddings in these special settings is often beneficial [4, 11], we explore domain- and language-adaptive pretraining of transformer models in this paper. Several recent works have shown that this kind of adaptation boosts performance for downstream tasks in non-standard domains by, e.g., pretraining with masked language modeling (MLM) objectives on documents from the target domain [1, 7].

In addition, we analyze the effects of model transfer between the first tasks of the challenge, as model transfer between related tasks in similar domains can result in significant performance gains [13].

Further, there is a line of work now questioning traditional train-dev splits [6] as well as random splits [16]. More challenging datasplits can be created by clustering the documents based on their similarity, where each split encodes unique information to a certain degree [18]. We use this method to train ensembles of models on these splits in a cross-validation format, such that each model has observed slightly different training instances.
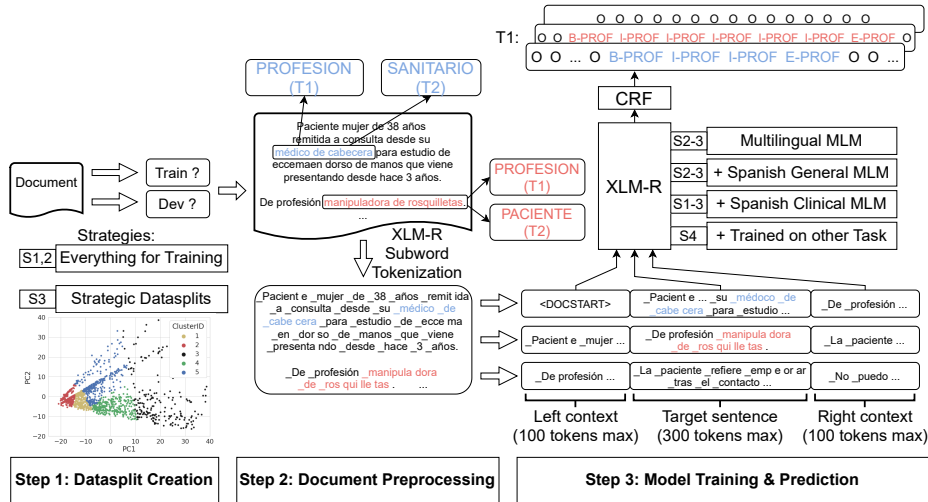
Fig. 1. Overview of the NLNDE system architecture. We mark the system variants S1-S4 referring to our different submissions. T1 and T2 refer to task 1 and task 2 of the shared task, respectively.

3 Approach

This section provides an overview of the different methods we used for the two tasks. The complete overview of our system is given in Figure 1.

3.1 Document Preprocessing

Tokenization can be challenging in non-standard domains, including the clinical domain [12]. We thus use the XLM-R subword tokenizer and perform sequence labeling on the subtoken level with spacy for sentence segmentation. Initial experiments showed possible improvements of up to 2 F_1 points compared to NER on token level.

3.2 Domain- and Language-specific Masked Language Modeling

We use XLM-R [3] as the main component of our models. XLM-R is a pretrained multilingual transformer model for 100 languages, including Spanish. It shows superior performance in different tasks across languages, and can even outperform monolingual models in certain settings. It was pretrained on a large-scale corpus, and Spanish documents made up only 2% of this data, as provided in Table 1.

Thus, we explore further pretraining of this model and tune it towards Spanish documents by pretraining either on (1) a medium-size Spanish corpus with general domain documents [2] or (2) a smaller Spanish clinical corpus consisting of the MeSpEN resources [17] and publicly available Scielo articles. Note that the clinical corpus was not part of the general-domain corpus.

Table 1. Sizes of Pretraining Corpora.

Corpus	Size
Original XLM-R corpus	2.4 TB
– Spanish Subcorpus	53.3 GB
Spanish Corpus	17.3 GB
Spanish Clinical Corpus	789.2 MB

Table 2. Hyperparameters.

XLM-R	
Embedding size	1024
Max. sentence length	300
Context to left/right	100
Optimizer	
Batch size	16
Learning rate	$2e-5$
β_1, β_2	0.9, 0.999
Weight decay	0.0
Training	
Epochs	20
Early stopping	training loss *or* on dev. set

We use masked language modeling for pretraining and trained for three epochs over the corpora, which roughly corresponds to 30k steps for the smaller clinical corpus and 685k steps for the general-domain corpus using a batch-size of 4. Finally, we have three XLM-R variants that we compare in this paper:

1. Standard **XLM-R** pretrained on 100 languages by [3].
2. **Spanish XLM-R** based on standard XLM-R with further pretraining using Spanish documents from the general domain.
3. Spanish **Clinical XLM-R** based on standard XLM-R with further pretraining using Spanish documents from the clinical domain.

3.3 Sequence Tagger

For the sequence tagger, we use one of the XLM-R models, either the standard XLM-R or one of our adapted models, and apply a CRF layer on top [9]. We add this CRF layer to address the problem of longer multi-word annotations, as job descriptions often span several tokens or, in our case, subtokens (as explained in Section 3.1). In addition, a CRF prevents inconsistencies in the labels.

We split all sentences to a maximum length of 300 subtokens and add the context of up to 100 subtokens to the left/right to get cross-sentence information. The labels are in BIOSE encoding, which is an extended BIO encoding with additional labels for the last token of an annotation (E-) and single-token annotations (S-)

Our model architecture is basically the same across all runs. We only exchange the transformer model. The models are trained using an AdamW optimizer for a maximum of 20 epochs. Our hyperparameters are given in Table 2.

3.4 Strategic Datasplits

We test two options to train the sequence taggers:

(1) Using all of the available training data and stop training according to the training loss. This method provides the model with the most input instances. However, the stopping criterion is not as meaningful as using the task's metric on a held-out validation set.

(2) Thus, as our second method, we split the data into train and validation sets. Then, we train the model using only the train-fraction of all the data and use the held-out validation data to determine the best model, which is then used to annotate the test data. As an alternative to random splits, we follow [18] and create strategic datasplits by clustering the documents according to their similarity. This creates more challenging splits, as more distant documents are left out for validation.

For this, each document is represented as the average vector of the XLM-R embeddings for each token. This document representation is reduced to five dimensions using PCA. Finally, the documents are clustered into five equally-sized splits using k-Means clustering. We train five models for each task and embedding with each having a different validation split. Our splits are visualized in Figure 2. We see that clusters 1, 2 and 4 are densely populated with highly similar documents, while clusters 3 and 5 contain more distinct documents.

To better understand the strategic datasplits, we analyzed whether the different medical topics included in the corpus correlate with the splits. However, we found that the strategic clusters incorporate more diverse information than just topic similarity as there is no substantial overlap between topics and datasplits.
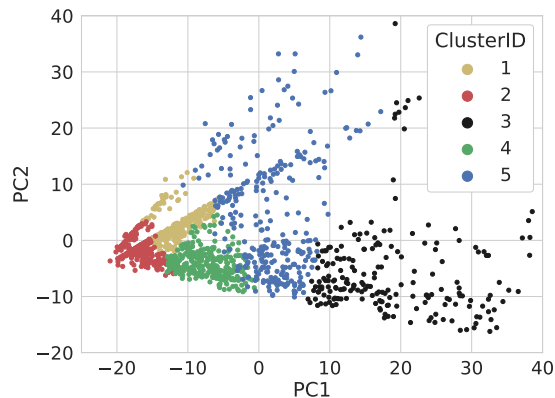

Fig. 2. Our strategic datasplits in the two-dimensional space with PCA.

3.5 Ensembling of Model Predictions

In order to capture the different advantages of multiple models, we combine them using ensembling. This is particularly helpful when models carry different types of information. For example, the models trained using our strategic datasplits all have seen a slightly different training set, and, thus, combining them using ensembling should further improve performance. We apply ensembling by majority voting. For this, we use hard voting that counts the labels by each model and does not consider the CRF probabilities. We convert the BIOSE labels to BIO labels for the ensembling process as the simpler BIO encoding leads to fewer conflicts. Further, we apply postprocessing of the label sequence to correct inconsistencies in the resulting label sequence and enforce for 0.23% of predictions for the test set that the first token of each annotations begins with B-.

3.6 Transfer Learning

As the first two tasks of the challenge are related and can possibly benefit from each other, we explore the potential of model transfer between them. For example, having basic knowledge of what and which occupations (task 1) are mentioned in a text can be useful to determine whether occupations are related to the patient or to someone else (task 2). For this, we first train models on the auxiliary task and then transfer the resulting model to the targeted main task. In our case, the auxiliary task is either task 1 or 2 and the other task is the main task.

3.7 Submissions

The following five runs are the NLNDE submissions to the MEDDOPROF shared task. We use the same model architectures for both tasks. Note that all submissions, except for $S1$ are ensembles of multiple models based on the three different embeddings. In Section 4, we compare these submissions with further model variations.

$S1$: The Spanish clinical XLM-R model trained on the complete training data.
$S2$: All three XLM-R language models combined in one ensemble (3 models).
$S3$: Ensemble of models trained using strategic datasplits (15 models).
$S4$: Ensemble of models based on transfer learning from the auxiliary task to the main task (3 models).
$S5$: The combination of all above models into one ensemble (21 models).

4 Results

Our official results for the first two tasks of the MEDDOPROF shared task are given in Table 3. In addition, we include several other models to analyze the performance of each embedding, because most of our submissions are ensembles combining all three XLM-R embeddings. The official evaluation metric is the F_1-score and our best models are highlighted.

We find that the Spanish XLM-R trained on general-domain Spanish data is often the best transformer compared to the standard XLM-R and the clinical one, probably because it was trained on the largest amount of data. In addition, the extraction of occupations is not unique to the clinical domain and general-domain Spanish knowledge seems to be beneficial for this as well.

We find that model transfer ($S4$) is only useful when transferring models from task 1 (the detection of occupations) to task 2 (the classification of occupations). Reusing models that already learned the detection of profession expressions as an auxiliary task improves the main task, e.g., by up to 2.9 F_1 points for XLM-R.

Training models on strategic datasplits ($S3$) provides the best results overall, and is even better than the ensemble of all models ($S5$). The strategic datasplits improved the ensemble model $S2$ by 0.4 and 2.3 F_1 points for task 1 and 2. Note that this submission unintentionally contained the Spanish general-domain XLM-R models with transfer learning in the ensemble. These were trained without our strategic datasplits. The corrected results are marked with "*".

The overall best model for task 1 is the ensemble of general-domain XLM-R trained using the strategic datasplits with an F_1-score of 83.2. This model was not submitted as a run to the shared task, but shows the importance of the language-adaptive pretraining and the usefulness of strategic datasplits.

Table 3. Results on the official test set. Our best submission is highlighted with underlines and the overall best model in bold.

	Task 1			Task 2		
	Precision	Recall	F1	Precision	Recall	F1
trained on all training instances						
XLM-R	83.9	75.7	79.6	76.9	71.3	74.0
clinical XLM-R (S1)	82.5	76.2	79.2	81.6	74.0	77.6
Spanish XLM-R	84.5	76.9	80.5	80.2	74.6	77.3
ensemble of all (S2)	85.1	77.7	81.2	80.6	73.7	77.0
transfer training						
XLM-R	82.5	75.2	78.7	80.8	73.4	76.9
clinical XLM-R	81.7	75.2	78.3	81.0	72.6	76.6
Spanish XLM-R (TSX)	81.6	75.0	78.2	81.5	74.9	78.1
ensemble of all (S4)	83.8	76.6	80.0	81.9	74.3	77.9
trained on strategic datasplits						
XLM-R	83.9	75.0	79.2	81.2	74.3	77.6
clinical XLM-R	83.8	76.6	80.0	80.7	75.4	77.9
Spanish XLM-R	**86.3**	**80.4**	**83.2**	82.5	**75.9**	79.1
ensemble of all *	**86.3**	78.7	82.3	82.1	75.5	78.7
ensemble of all + TSX (S3)	<u>85.5</u>	<u>78.3</u>	<u>81.8</u>	**83.0**	**75.9**	**79.3**
ensemble of all models (S5)	84.7	<u>78.3</u>	81.4	**83.0**	75.7	79.2

5 Conclusion

In this paper, we described our submissions for the first two tasks of the MED-DOPROF competition. By utilizing domain- and language-adaptive pretraining, strategic datasplits and ensembling methods, we were able to improve already high-performing transformer-based models by up to 5.3 F_1 points and achieved competitive results in the competition as neither language nor domain experts. Future work will include the exploration of different clinical corpora with our newly trained Spanish XLM-R models.

Acknowledgments

References

1. Beltagy, I., Lo, K., Cohan, A.: SciBERT: A pretrained language model for scientific text. In: Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP). pp. 3615–3620. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-1371, https://www.aclweb.org/anthology/D19-1371
2. Cañete, J.: Compilation of large spanish unannotated corpora (May 2019). https://doi.org/10.5281/zenodo.3247731, https://doi.org/10.5281/zenodo.3247731
3. Conneau, A., et al.: Unsupervised cross-lingual representation learning at scale. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8440–8451. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.747, https://www.aclweb.org/anthology/2020.acl-main.747
4. Friedrich, A., et al.: The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 1255–1268. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.116, https://www.aclweb.org/anthology/2020.acl-main.116
5. Gonzalez-Agirre, A., Marimon, M., Intxaurrondo, A., Rabal, O., Villegas, M., Krallinger, M.: PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. pp. 1–10. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-5701, https://www.aclweb.org/anthology/D19-5701
6. Gorman, K., Bedrick, S.: We need to talk about standard splits. In: Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics. pp. 2786–2791. Association for Computational Linguistics, Florence, Italy (Jul 2019). https://doi.org/10.18653/v1/P19-1267, https://www.aclweb.org/anthology/P19-1267

7. Gururangan, S., et al.: Don't stop pretraining: Adapt language models to domains and tasks. In: Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics. pp. 8342–8360. Association for Computational Linguistics, Online (Jul 2020). https://doi.org/10.18653/v1/2020.acl-main.740, https://www.aclweb.org/anthology/2020.acl-main.740

8. Hedderich, M.A., Lange, L., Adel, H., Strötgen, J., Klakow, D.: A survey on recent approaches for natural language processing in low-resource scenarios. In: Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies. pp. 2545–2568. Association for Computational Linguistics, Online (Jun 2021). https://doi.org/10.18653/v1/2021.naacl-main.201, https://www.aclweb.org/anthology/2021.naacl-main.201

9. 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. pp. 282–289. ICML '01, Morgan Kaufmann Publishers Inc., San Francisco, CA, USA (2001), http://dl.acm.org/citation.cfm?id=645530.655813

10. Lange, L., Adel, H., Strötgen, J.: NLNDE: Enhancing neural sequence taggers with attention and noisy channel for robust pharmacological entity detection. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks. pp. 26–32. Association for Computational Linguistics, Hong Kong, China (Nov 2019). https://doi.org/10.18653/v1/D19-5705, https://www.aclweb.org/anthology/D19-5705

11. Lange, L., Adel, H., Strötgen, J.: NLNDE: The neither-language-nor-domain-experts' way of spanish medical document de-identification. In: Proceedings of The Iberian Languages Evaluation Forum (IberLEF 2019). CEUR Workshop Proceedings (2019), http://ceur-ws.org/Vol-2421/MEDDOCAN_paper_5.pdf

12. Lange, L., Dai, X., Adel, H., Strötgen, J.: NLNDE at CANTEMIST: neural sequence labeling and parsing approaches for clinical concept extraction (2020), https://arxiv.org/abs/2010.12322

13. Lange, L., Strötgen, J., Adel, H., Klakow, D.: To share or not to share: Predicting sets of sources for model transfer learning. arXiv preprint arXiv:2104.08078 (2021)

14. 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)

15. Marimon, M., et al.: Automatic de-identification of medical texts in spanish: the meddocan track, corpus, guidelines, methods and evaluation of results. In: Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2019), CEUR Workshop Proceedings, 2019. pp. 618–638 (2019)

16. Søgaard, A., Ebert, S., Bastings, J., Filippova, K.: We need to talk about random splits. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume. pp. 1823–1832. Association for Computational Linguistics, Online (Apr 2021), https://www.aclweb.org/anthology/2021.eacl-main.156

17. Villegas, M., Intxaurrondo, A., Gonzalez-Agirre, A., Marimon, M., Krallinger, M.: The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. LREC

MultilingualBIO: Multilingual Biomedical Text Processing (Malero M, Krallinger M, Gonzalez-Agirre A, eds.) (2018)

18. Wecker, H., Friedrich, A., Adel, H.: ClusterDataSplit: Exploring challenging clustering-based data splits for model performance evaluation. In: Proceedings of the First Workshop on Evaluation and Comparison of NLP Systems. pp. 155–163. Association for Computational Linguistics, Online (Nov 2020). https://doi.org/10.18653/v1/2020.eval4nlp-1.15, https://www.aclweb.org/anthology/2020.eval4nlp-1.15