CLIN-X: pre-trained language models and a study on cross-task transfer for concept extraction in the clinical domain

Lukas Lange¹²∗, Heike Adel¹, Jannik Strötgen¹ and Dietrich Klakow²

¹Bosch Center for Artificial Intelligence, Robert Bosch GmbH, Renningen, 71272, Germany and
²Spoken Language Systems, Saarland University, Saarbrücken, 66111, Germany.

* To whom correspondence should be addressed.

Abstract

Motivation: The field of natural language processing (NLP) has recently seen a large change towards using pre-trained language models for solving almost any task. Despite showing great improvements in benchmark datasets for various tasks, these models often perform sub-optimal in non-standard domains like the clinical domain where a large gap between pre-training documents and target documents is observed. In this paper, we aim at closing this gap with domain-specific training of the language model and we investigate its effect on a diverse set of downstream tasks and settings.

Results: We introduce the pre-trained CLIN-X (Clinical XLM-R) language models and show how CLIN-X outperforms other pre-trained transformer models by a large margin for ten clinical concept extraction tasks from two languages. In addition, we demonstrate how the transformer model can be further improved with our proposed task- and language-agnostic model architecture based on ensembles over random splits and cross-sentence context. Our studies in low-resource and transfer settings reveal stable model performance despite a lack of annotated data with improvements of up to 47 F₁ points when only 250 labeled sentences are available. Our results highlight the importance of specialized language models as CLIN-X for concept extraction in non-standard domains, but also show that our task-agnostic model architecture is robust across the tested tasks and languages so that domain- or task-specific adaptations are not required.

Availability: The CLIN-X language models and source code for fine-tuning and transferring the model are publicly available at https://github.com/boschresearch/clin_x/ and the huggingface model hub.

Contact: Lukas.Lange@de.bosch.com

1 Introduction

Collecting and understanding key clinical information, such as disorders, symptoms, drugs, etc., from electronic health records (EHRs) has wide-ranging applications within clinical practice and research [Leaman et al. 2015] [Wang et al. 2013]. A better understanding of this information can, on the one hand, facilitate novel clinical studies, and, on the other hand, help practitioners to optimize clinical workflows. However, free text is ubiquitous in EHRs. This leads to great difficulties in harvesting knowledge from EHRs. Therefore, natural language processing (NLP) systems, especially information extraction components, play a critical role in extracting and encoding information of interest from clinical narratives, as this information can then be fed into downstream applications. For example, the extraction of structured information from clinical narratives can help in decision making or drug repurposing [Marimon et al. 2019].

However, information extraction in non-standard domains like the clinical domain is a challenging problem due to the large number of complex terms and unusual document structures [Lee et al. 2020]. In addition, pre-trained language models (PLM) such as BERT [Devlin et al. 2019] that demonstrated superior performance for many NLP tasks are typically trained on standard domains, such as web texts, news articles or Wikipedia. Despite showing some robustness across languages and domains [Conneau et al. 2020] these models still achieve their best performance when applied to targets similar to their pre-training corpora which can limit their applicability in many situations [Gururangan et al. 2020]. One way to overcome this domain-gap is the adaptation of existing language models to the new target domain or training a new domain-specific model from scratch [Effag et al. 2019] [Lee et al. 2020]. Several...
recent works have shown that this kind of adaptation boosts performance for downstream tasks in non-standard domains by e.g., pre-training with masked language modeling (MLM) objectives on documents from the target domain. (Webster et al., 2019; Nasernia et al., 2021).

While all the previous methods help to build high-performing model architectures, often there is also a lack of annotated data in the clinical domain which is usually needed for all deep-learning-based models. On the one hand, this domain has high requirements regarding the removal or masking of protected health information (PHI) of individuals (Trainor et al., 2007; Stubbs et al., 2015) which is particularly worthy of protection and can prevent data publication. On the other hand, information extraction tasks are often specific to their target domain and clinical concepts are only found very infrequently outside EHRs which limits reusability of existing resources. Possible solutions for the low-resource problem can be multi-task learning (Khan et al., 2020; Mulyar et al., 2021) or transfer learning (Lee et al., 2018; Peng et al., 2019) across similar corpora from the clinical domain. However, transferring knowledge is particularly challenging in the clinical domain as biomedical NLP models have problems generalizing to new entities (Kim and Kang, 2021). Therefore, one has to carefully select the transfer sources (Lange et al., 2021).

Over the last years, we have participated in a series of shared tasks on information extraction in the Spanish clinical domain (Marimon et al., 2019; Miranda-Lascalada et al., 2020; Aldea López et al., 2021). With our systems, we were able to outperform the other participants and won the competitions twice. The winning systems were task-agnostic and utilized domain-adapted language models and word embeddings (Lange et al., 2019), as well as improved training routines for transformer models (Lange et al., 2021). Based on our findings and lessons learned during the competitions, we propose in this paper a robust model architecture and training procedure for concept extraction in the clinical domain that is task- and language-agnostic. We introduce a new Spanish clinical language model CLIN-X$_{ES}$ (Clinical XLM-R) that outperforms existing transformer models on Spanish corpora and exemplifies the benefits of cross-language domain adaptation for English tasks as well. For this, we perform a broad evaluation of ten clinical information extraction tasks from two languages (English and Spanish), including low-resource and transfer settings. Finally, we perform cross-task transfer experiments and show that this can boost performance by more than $47$ F$_1$ points for few-shot training. Our results demonstrate great and consistent improvements compared to standard transformer models across all tasks in both languages. We release both, CLIN-X$_{ES}$ as well as its English counterpart CLIN-X$_{EN}$.

![Fig. 1: Overview of the concept extraction pipeline based on CLIN-X and our model components for subword-based extraction with cross-sentence context, BIOSE labels, CRFs and model transfer.](image)

2 Approach

In this paper, we introduce new pre-trained language models and propose a robust model architecture to perform concept extraction in the clinical domain for English and Spanish. The overall model architecture is shown in Figure 1 and our proposed model components are highlighted. First, the input is computed on subword-level instead of the usual word-level, which eliminates the need for external tokenization. In addition, the input is enriched with its cross-sentence context to capture a wider document context. Second, the input is processed by a transformer model that is adapted to our target domain. Third, the model output is computed using a conditional random field (CRF) output layer to address long annotations. Then, an ensemble over models trained on different training splits is computed that reduces variance and captures the complementary knowledge from all models. Finally, we experiment with cross-task model transfer to further improve the model in few-shot settings.

In summary, the contributions of this paper are as follows:

- We study the impact of domain-adaptive pre-training for clinical concept extraction for different embedding types and publish new language models that are adapted to the clinical domain. We show that this PLM outperforms other publicly available embeddings and models in our settings and we also show that cross-language domain adaptations works for English tasks as well.
- We perform a broad evaluation of ten clinical sequence labeling tasks across two languages, including low-resource and transfer settings. By this, we demonstrate how our methods can further boost already high-performing transformer models by using advanced training methods and effective changes in the architecture.
- Our models outperform the state-of-the-art methods for clinical and biomedical concept extraction, as well as various other transformer models for all ten tasks.
- We make our new domain-adapted CLIN-X language models and the source code for fine-tuning the concept extraction models using our methods publicly available.

3 Materials and Methods

In this section, we start with a brief description of the input representations. Then, we discuss our proposed architectural choices as well as the advanced training methods.

3.1 Input Representations for the Clinical Domain

State-of-the-art methods for concept extraction typically rely on word embeddings or language models as input representations. The standard approach is the pre-training of these models on large-scale unannotated datasets once and their reuse as powerful representations for many downstream applications (Collobert et al., 2011; Phan et al., 2019). These methods have shown that contextualized embeddings help in particular in the medical domain, e.g., due to the high number of synonyms. Thus, we focus on the usage of contextualized embeddings in this work, which are most often retrieved from transformer language models nowadays. This is either done with auto-regressive language modeling (Peters et al., 2018) or masked language modeling (Devlin et al., 2019), which we use in this paper.

Domain-specific embeddings. A popular way to approach the challenges of NLP in non-standard domains is the inclusion of domain knowledge via domain-specific embeddings (Friedrich et al., 2020). For this, word embeddings or language models are pre-trained or further specialized on documents of the target domain. These embeddings can be used in
In addition to the Spanish 1, information extraction tasks are typically performed on the token level, while most transformers work on finer subwords instead. Thus, the input representations from transformers for tokens are either retrieved from the first subword or the average 20. In contrasts, we perform concept extraction directly on the subword level. By doing this, there is no need for external tokenization besides the subword segmentation of the transformer. Note that the usage of domain-specific subwords is still often beneficial compared to the general domain segmentation 23. Cross-sentence context. Transformers are suited to incorporate information from a larger context. 24, 25 showed that context information from neighboring sentences has positive effects for named entity recognition on the general domain. 26 also showed the positive impact of context for clinical concept extraction. We follow these approaches and add context information to the input similar to 27. We incorporate the context of 100 subwords to the left and right and use the document boundaries to set the context limits as all corpora are clearly separated in documents.

Conditional Random Field Output. 28, 29 have shown, entity recognition models in the biomedical domain tend to memorize training instances and their labels. This can result in incorrect label encodings as the model fails to generalize. A conditional random field (CRF) 30 can constrain these incorrect sequences as the Viterbi algorithm is used for decoding. In addition, the CRF has advantages when it comes to long entities covering multiple tokens 31 that appear frequently in the clinical domain.

3.3 Training on Data Splits.

Many NLP tasks suffer from a lack of labeled data. This includes non-standard domains like the clinical domain in particular. One solution to improve performance in these domains is the usage of resources from a related task in a transfer process. For example, 32 have shown that few-shot NER in the biomedical domain can be improved by transferring trained weights from a similar task. We perform a similar kind of model transfer by transferring the transformer to the new target. However, not all transfer sources are actually useful as many can lead to negative transfer 33. Thus, we first have to predict a suitable transfer source. We follow 34 and compute...
In addition to the other methods, ensembling can be used to combine multiple model predictions into one. This ensemble, similar to the similarity between our datasets using their proposed model similarity measure. This has been shown to work well across different tasks and domains. The similarity between two models is computed based on the neural feature representations for the target datasets between two task-specific trained models. In our experiments, we study the effect of transfer from different sources in comparison to standard single-task training. Further, we will investigate this kind of transfer in low-resource settings, when the target task has only limited training resources.

Ensembles over models. In addition to the other methods, ensembling can be used to combine multiple model predictions into one. This ensemble is usually better than a single model – in particular if the models or their training data differ to some degree. We either create ensembles by majority voting ([Clark et al. 2019]) of training runs that vary by their random seed (standard splits) or their training data (random splits).

Table 2. Statistics of the English datasets.

| Corpus       | Size (#Sentences) | Train | Dev  | Test |
|--------------|------------------|-------|------|------|
| i2b2 2006    | Uzuner et al. 2007 | 51,429 | 18,770 |
| i2b2 2010    | Uzuner et al. 2011 | 16,487 | 27,882 |
| i2b2 2012    | Sum et al. 2013   | 7,636  | 5,785 |
| i2b2 2014    | Stabbers et al. 2015 | 52,026 | 33,317 |

Table 3. Overview of different models averaged for the two languages (P) for the CLIN-X models. The training takes less than 1 day with a batch size of 4 per device and a sequence length of up to 512 subwords. The models were trained with the huggingface trainer for MLM.

| Pre-training Domain | Model       | English | Spanish |
|---------------------|-------------|---------|---------|
| General             | word2vec    | 80.26   | 78.20   |
| (e.g., Web, News, ...) | flair       | 85.15   | 80.28   |
| Wikipedia, ...)     | BETO (En)   | 85.34   | 77.78   |
|                     | XLM-R       | 87.13   | 83.87   |

Combined models. The similarity between two models is computed based on the neural feature representations for the target datasets between two task-specific trained models. In our experiments, we study the effect of transfer from different sources in comparison to standard single-task training. Further, we will investigate this kind of transfer in low-resource settings, when the target task has only limited training resources.

4 Results

This section describes the experimental setup starting with tasks, datasets and implementation details, and discusses the results for our experiments.

4.1 Tasks and Datasets

Many datasets for natural language processing in specialized domains are published in the context of shared tasks – competitions to evaluate different systems and approaches. Besides English, the clinical domain is well addressed for Spanish, and there exists an active community of researchers for natural language processing of Spanish clinical texts. Thus, in the context of the IberLEF workshop series (Iberian Language Evaluation Forum), several shared tasks have been proposed by the Barcelona Supercomputing Center concerning concept extraction in the clinical domain (Marimon et al. 2019; Gonzalez-Agirre et al. 2019; Miranda Escalada et al. 2020b; Lima-Lopez et al. 2021). In addition to datasets of these shared tasks for Spanish, we consider four English datasets published during a series of shared tasks of the i2b2 project (Uzuner et al. 2007, 2011; Sun et al. 2013; Stabbers et al. 2015). Information on the dataset sizes are given in Table 2 and 3 for Spanish and English, respectively. Note that the Meddoprof and i2b2 2012 corpora consist of two different extraction tasks each. Thus, we consider both tracks as separated tasks in this work resulting in a total of ten tasks. Following the evaluations in the shared tasks, we use the strict micro F1 for all datasets as evaluation metric.

4.2 Experimental Setup and Implementation Details

| Pre-training Domain | Model       | English | Spanish |
|---------------------|-------------|---------|---------|
|                     | word2vec    | 80.26   | 78.20   |
|                     | flair       | 85.15   | 80.28   |
|                     | BETO (En)   | 85.34   | 77.78   |
|                     | XLM-R       | 87.13   | 83.87   |

4.3 Evaluation of Embeddings

The choice of input embeddings has a large impact on downstream performance and may be the most important factor. Table 3 shows the average performance of several different embeddings and transformer models for the two languages. As expected, the monolingual transformers (BERT, BETO) excel at their target language, but cannot compete with multilingual models (mBERT, XLM-R) when applied to an unseen language. The lower part of Table 3 lists domain-specific variants of the embeddings which are generally more powerful in our domain-specific setting. We see that our CLIN-X models perform best for their respective languages. Furthermore, the CLIN-X$_{EN}$ performs almost as well as the CLIN-X$_{ES}$ model on the English datasets, for which it was not explicitly trained. This shows, that the domain adaptation of multilingual models can also help for texts from other languages of the same domain. Due to CLIN-X$_{ES}$ stable performance across all tasks and languages, we will use this model for the following ablations and transfer experiments.

4.4 Evaluation of Training Methods

The foundation for all following concept extraction models is the CLIN-X$_{EN}$, as it has shown robust results across all tasks. For comparison to fixed standard splits, we train the models on different random splits. We see in Table 3 that in particular ensembles over random
splits are a lot better than the standard splits and also all training instances. While the median performance is roughly similar for all methods, the random splits offer a lot more variety in training instances and allow for better maximum performance models. Thus, the ensemble based on random splits achieves also much higher numbers.

Table 5. Cross-task transfer results for few-shot settings for the English corpora (F1). The predicted transfer source and the best models are highlighted.

| Method          | English | Spanish |
|-----------------|---------|---------|
| All             | 87.83   | 86.46   |
| Median model    | 87.63   | 85.16   |
| Best model      | 87.85   | 85.99   |
| Ensemble        | 87.95   | 86.06   |
| Median model    | 87.69   | 86.17   |
| Best model      | 88.31   | 86.85   |
| Ensemble        | 88.78   | 88.15   |

- BIOSE Labels 88.52 87.13
- CRF 88.38 85.95
- Context 87.83 86.84
- Subword NER 87.38 86.81

4.5 Evaluation of Concept Extraction Models

The lower part of Table 1 lists an ablation study of our individual model components. For example, adding cross-sentence context to the transformers boosts performance across all tasks by 0.5 F1 on average. Performing concept extraction on the subword level helps even further. This is particularly helpful considering that no external tokenization is needed, which can be challenging in the clinical domain (Lange et al., 2020). The CRF helps for both languages, though the differences are larger for Spanish, as the two MEDDOPROF tasks have particularly long sequences. Performing concept extraction on the subword level helps even further. The ensemble based on random splits achieves also much higher numbers.

Table 6. Cross-task transfer results for few-shot settings for the Spanish corpora (F1). The predicted transfer source and the best models are highlighted.

| Tgt. Src. / Setting 250 500 1000 2500 7500 All | # training sentences |
|-----------------------------------------------|----------------------|
| No Transfer                                   | 71.24 81.06 84.15 95.49 96.89 98.34 |
| Ablation                                      | 71.58 80.31 82.99 95.87 97.97 97.41 |
| 2012-C                                        | 87.52 90.86 91.87 97.11 97.95 98.50 |
| 2012-T                                        | 65.38 74.96 82.59 85.54 88.88 89.10 |
| 2012-C                                        | 83.99 86.25 86.88 89.46 89.34 89.74 |
| 2012-T                                        | 69.49 74.92 82.11 83.32 85.25 85.65 |
| 2014                                          | 72.05 79.11 82.49 85.25 87.69 88.80 |
| No Transfer                                   | 69.09 73.21 75.70 78.03 80.36 80.42 |
| Ablation                                      | 68.50 71.50 74.64 77.86 79.25 80.15 |
| 2012-C                                        | 76.39 77.98 79.44 80.90 81.65 80.93 |
| 2012-T                                        | 65.30 69.61 73.30 75.88 80.25 80.12 |
| 2014                                          | 68.67 72.56 75.39 77.96 79.98 80.83 |
| No Transfer                                   | 67.49 72.67 75.44 78.00 78.33 78.48 |
| Ablation                                      | 68.57 72.49 74.34 77.73 78.43 78.34 |
| 2012-C                                        | 70.17 75.04 76.36 78.12 78.54 80.03 |
| 2012-T                                        | 69.44 72.66 75.04 77.88 88.76 79.36 |
| No Transfer                                   | 64.96 81.61 85.74 92.70 96.08 97.62 |
| Ablation                                      | 81.50 85.76 88.96 93.51 96.04 97.46 |
| 2012-C                                        | 71.72 83.55 87.81 93.18 96.14 97.17 |
| 2012-T                                        | 71.24 82.97 87.09 93.15 96.13 97.33 |
| 2014                                          | 69.12 81.25 85.08 91.35 96.02 97.00 |

4.6 Evaluation of Transfer Learning

In addition to the training based on random splits, we explore the effects of transfer learning. For this, we simulate low-resource settings where we limit the annotated data of the target dataset between 250 labeled sentences to up to 7500 sentences, roughly the size of the smallest corpus. The results are given in Table 4 and Table 5 for English and Spanish, respectively.

Large positive transfer happens in most settings, particularly for the low-resource settings with up to (+47.3 F1 points) for Meddocan when only 250 labeled sentences are available. The improvements in the full data scenario are below 1 F1. However, there is also negative transfer, in 250 labeled sentences, which can be challenging in the clinical domain (Lange et al., 2020). The CRF helps for both languages, though the differences are larger for Spanish, as the two MEDDOPROF tasks have particularly long sequences. Performing concept extraction on the subword level helps even further. The ensemble based on random splits achieves also much higher numbers.

4.5 Evaluation of Concept Extraction Models

The lower part of Table 1 lists an ablation study of our individual model components. For example, adding cross-sentence context to the transformers boosts performance across all tasks by 0.5 F1 on average. Performing concept extraction on the subword level helps even further. This is particularly helpful considering that no external tokenization is needed, which can be challenging in the clinical domain (Lange et al., 2020). The CRF helps for both languages, though the differences are larger for Spanish, as the two MEDDOPROF tasks have particularly long lengths. In addition to the training based on random splits, we explore the effects of transfer learning. For this, we simulate low-resource settings where we limit the annotated data of the target dataset between 250 labeled sentences to up to 7500 sentences, roughly the size of the smallest corpus. The results are given in Table 4 and Table 5 for English and Spanish, respectively.

Large positive transfer happens in most settings, particularly for the low-resource settings with up to (+47.3 F1 points) for Meddocan when only 250 labeled sentences are available. The improvements in the full data scenario are below 1 F1. However, there is also negative transfer, in particular using i2b2 2012-T and Cantemist datasets as transfer sources.
often result in negative transfer. The source selection is also crucial in low-resource scenarios, as not every source is equally beneficial. Using the model similarity measure from Lange et al. (2021b) we are able to predict good transfer sources in all settings, often the best source is selected.

### 4.7 Comparison to State-of-the-Art Models

As our results demonstrate, we have proposed a robust model for the clinical domain that works well across the different tasks in both languages. Finally, we compare CLIN-X to various transformer models as introduced earlier. We also compare to HunFlair (Richter et al., 2021), the current state-of-the-art for concept extraction in the biomedical domain. We use their model architecture based on clinical flair and fasttext embeddings and train models accordingly on our datasets. In addition, we compare to our NLNDE submissions for the Spanish shared tasks and the ClinicalBERT (Altenzter et al., 2019) for the English datasets.

The results for each task are shown in Table 7. The CLIN-X language models in combination with our model architecture outperform the other transformers and HunFlair by a large margin. CLIN-X is able to utilize the domain knowledge obtained from the additional pre-training with further improvements from the ensemble over random splits. Even though CLIN-X works best in combination with our model architecture, CLIN-X based on the standard transformer architecture with a single classification layer already outperforms the existing models on 8 out of 10 tasks. We tested statistical significance between CLIN-XES with and without transfer learning – highlighted with asterisks in Table 7. We find that all differences for English are significant, while only one difference for Spanish is significant. This might indicate the complementary relationship of domain adaptation and model transfer learning. As CLIN-X was explicitly adapted to Spanish, additional transfer is not necessary in high-resource settings. In contrast, the cross-language domain adaptation for English can still be improved with transfer from related sources, where CLIN-XES +Transfer has also notably higher performances in 3 out of 5 settings compared to CLIN-XEN, which is adapted to English.

### 5 Conclusion

In this paper, we described the newly pre-trained CLIN-X language models for the clinical domain. We have shown that CLIN-X sets the new state of the art for ten clinical concept extraction tasks in two languages. We demonstrated the positive impact of other model components, such as ensembles over random splits and cross-sentence context and we have studied the effects of cross-task transfer learning from different clinical corpora. Using a model similarity measure, we found good transfer sources for almost all datasets in general and for low-resource scenarios in particular. We are convinced that the new CLIN-X language models will help boosting performance for various Spanish and English clinical information extraction tasks with our or other model architectures.

### References

Akbik, A. et al. (2018). Contextual string embeddings for sequence labeling. In Proceedings of the 27th International Conference on Computational Linguistics, pages 1638–1649, Santa Fe, New Mexico, USA. ACL.

Altenzter, E. et al. (2019). Publicly available clinical BERT embeddings. In Proceedings of the 2nd Clinical Natural Language Processing Workshop (Clin-NLP), pages 72–78, Minneapolis, Minnesota, USA. ACL.

Belnagy, I. et al. (2019). 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), pages 3615–3620, Hong Kong, China. ACL.

Clark, C. et al. (2019). Don’t take the easy way out: Ensemble based methods for avoiding known dataset biases. 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), pages 4069–4082, Hong Kong, China. ACL.

Collobert, R. et al. (2011). Natural language processing (almost) from scratch. J. Mach. Learn. Res., 12, 2493–2537.

Conneau, A. et al. (2020). Unsupervised cross-lingual representation learning at scale. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 8440–8451, Online. ACL.

Devlin, J. et al. (2019). 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 (NAACL-HLT), pages 4171–4186, Minneapolis, Minnesota. ACL.

Finkel, J. et al. (2004). Exploiting context for biomedical entity recognition: From syntax to the web. In Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications (NLPA/BioNLP), pages 91–94, Geneva, Switzerland.

---

**Table 7** Comparison to baseline systems and state-of-the-art results (*P*). We highlight statistically significant differences between CLIN-XES +OurArchitecture with and without transfer following the significant codes of R: ***p*-value ≤ 0.001; **p*-value < 0.01; *p*-value < 0.05; †highlights our ClinicalBERT results.

| Model                  | English (2b2) | 2006 | 2010 | 2012-C | 2012-T | 2014 | Spanish | 2014 | Medocan | M. prof-N | M. prof-C | Pharma. |
|------------------------|---------------|------|------|--------|--------|------|---------|------|---------|-----------|-----------|---------|
| BERT/BETO (monolingual) | 94.80         | 94.80| 94.80| 94.80  | 94.80  | 94.80| 94.80   | 94.80| 94.80   | 94.80     | 94.80     | 94.80   |
| BERT (multilingual)    | 94.79         | 94.79| 94.79| 94.79  | 94.79  | 94.79| 94.79   | 94.79| 94.79   | 94.79     | 94.79     | 94.79   |
| XLM-R (multilingual)   | 96.72         | 96.72| 96.72| 96.72  | 96.72  | 96.72| 96.72   | 96.72| 96.72   | 96.72     | 96.72     | 96.72   |
| HunFlair (monolingual) | 93.48         | 93.48| 93.48| 93.48  | 93.48  | 93.48| 93.48   | 93.48| 93.48   | 93.48     | 93.48     | 93.48   |
| ClinicalBERT           | 94.8          | 94.8 | 94.8 | 94.8   | 94.8   | 94.8 | 94.8    | 94.8 | 94.8    | 94.8      | 94.8      | 94.8    |
| NLNDE                  | -             | -    | -    | -      | -      | -    | -       | -    | -       | -         | -         | -       |
| **CLIN-X**             | 96.25         | 96.25| 96.25| 96.25  | 96.25  | 96.25| 96.25   | 96.25| 96.25   | 96.25     | 96.25     | 96.25   |
| **CLIN-X** +OurArchitecture | 98.49       | 98.49| 98.49| 98.49  | 98.49  | 98.49| 98.49   | 98.49| 98.49   | 98.49     | 98.49     | 98.49   |
| **CLIN-XES** +OurArchitecture +Transfer | 98.50  | 98.50| 98.50| 98.50 | 98.50 | 98.50| 98.50   | 98.50| 98.50   | 98.50     | 98.50     | 98.50   |
International Committee on Computational Linguistics.

Friedrich, A. et al. (2020). 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 (ACL), pages 1255–1268, Online. ACL.

Gonzalez-Agirre, A. et al. (2019). PharmaCoNER: Pharmacological substances, compounds and proteins named entity recognition track. In Proceedings of The 5th Workshop on BioNLP Open Shared Tasks (BioNLP-OST), pages 1–10, Hong Kong, China. ACL.

Gorman, K. and Bedrick, S. (2019). We need to talk about standard splits. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 2786–2791, Florence, Italy. ACL.

Gururangan, S. et al. (2020). Don’t stop pretraining: Adapt language models to domains and tasks. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL), pages 8342–8360, Online. ACL.

Haynes, R. B. et al. (2005). Optimal search strategies for retrieving scientifically strong studies of treatment from medline: analytical survey. Bmj, 330, 1179.

Hofer, M. et al. (2018). Few-shot learning for named entity recognition in medical text. arXiv preprint arXiv:1811.05468.

Khan, M. R. et al. (2020). Mi-bioner: Multi-task learning for biomedical named entity recognition using deep bidirectional transformers. arXiv preprint arXiv:2001.08904.

Kim, H. and Kang, J. (2021). How do your biomedical named entity models generalize to novel entities? arXiv preprint arXiv:2101.00160.

Lafferty, J. D. et al. (2001). Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In Proceedings of the Eighteenth International Conference on Machine Learning, ICML, ’01, pages 282–289, San Francisco, CA, USA.

Lange, L. et al. (2019). NLNDE: The neither-language-nor-domain-experts’ way of spanish medical document de-identification. In Proceedings of The Iberian Languages Evaluation Forum (IberLEF), CEUR Workshop Proceedings.

Lange, L. et al. (2020). NLNDE at CANTEMIST: neural sequence labeling and parsing approaches for clinical concept extraction. In Proceedings of The Iberian Languages Evaluation Forum (IberLEF), CEUR Workshop Proceedings.

Lange, L. et al. (2021a). Boosting transformers for job expression extraction and classification in a low-resource setting. In Proceedings of The Iberian Languages Evaluation Forum (IberLEF), CEUR Workshop Proceedings.

Lange, L. et al. (2021b). To share or not to share: Predicting sets of sources for model transfer learning. In Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing (EMNLP), pages 8744–8753, Online and Punta Cana, Dominican Republic. ACL.

Leaman, R. et al. (2015). Challenges in clinical natural language processing for automated disorder normalization. J. Biomed. Inform., 57, 28–37.

Lee, J. et al. (2020). Biobert: a pre-trained biomedical language representation model for biomedical text mining. Bioinformatics, 36, 1234–1240.

Lee, J. Y. et al. (2018). Transfer learning for named-entity recognition with neural networks. In Proceedings of the Eleventh International Conference on Language Resources and Evaluation (LREC), Miyazaki, Japan. European Language Resources Association.

Lima-López, S. et al. (2021). Nlp applied to occupational health: Meddoped shared task at iberlef 2021 on automatic recognition, classification and normalization of professions and occupations from medical texts. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF), CEUR Workshop Proceedings.

Luoma, J. and Pyysalo, S. (2020). Exploring cross-sentence contexts for named entity recognition with BERT. In Proceedings of the 28th International Conference on Computational Linguistics (COLING), pages 904–914, Barcelona, Spain (Online). International Committee on Computational Linguistics.

Marimon, M. et al. (2019). 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), CEUR Workshop Proceedings.

Miranda-Escalada, A. et al. (2020). Named entity recognition, concept normalization and clinical coding: Overview of the cantemist track for cancer text mining in spanish, corpus, guidelines, methods and results. In Proceedings of the Iberian Languages Evaluation Forum (IberLEF), CEUR Workshop Proceedings.

Mulyar, A. et al. (2021). Mc-clinical bert: scaling clinical information extraction with multitask learning. J. Am. Med. Inform. Assoc., 28, 2108–2115.

Naseem, U. et al. (2021). Bioalbert: A simple and effective pre-trained language model for biomedical named entity recognition. In 2021 International Joint Conference on Neural Networks (IJCNN), pages 1–7. IEEE.

Peng, Y. et al. (2019). Transfer learning in biomedical natural language processing: An evaluation of BERT and ELMo on ten benchmarking datasets. In Proceedings of the 18th BioNLP Workshop and Shared Task (BioNLP), pages 58–65, Florence, Italy. ACL.

Peters, M. E. et al. (2018). Deep contextualized word representations. In Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (NAACL-HLT), pages 2227–2237, New Orleans, Louisiana. ACL.

Phan, M. C. et al. (2019). Robust representation learning of biomedical names. In Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics (ACL), pages 3275–3285, Florence, Italy. ACL.

Schweter, S. and Akbik, A. (2020). Flert: Document-level features for named entity recognition. arXiv preprint arXiv:2011.06993.

Stubbins, A. et al. (2015). Automated systems for the de-identification of longitudinal clinical narratives: Overview of 2014 2b2/health shared task track 1. J. Biomed. Inform., 58, 11–19.

Sun, W. et al. (2013). Evaluating temporal relations in clinical text: 2012 2b2 challenge. J. Am. Med. Inform. Assoc., 20, 806–813.

Uzuner, Ö. et al. (2007). Evaluating the state-of-the-art in automatic de-identification. J. Am. Med. Inform. Assoc., 14, 550–563.

Uzuner, Ö. et al. (2011). 2010 2b2/nu challenge on concepts, assertions, and relations in clinical text. J. Am. Med. Inform. Assoc., 18, 552–556.

Villegas, M. et al. (2018). The mespen resource for english-spanish medical machine translation and terminologies: census of parallel corpora, glossaries and term translations. LREC MultilingualBIO.

Wang, Y. et al. (2018). Clinical information extraction applications: a literature review. J. Biomed. Inform., 77, 34–49.

Weber, L. et al. (2019). HUNER: improving biomedical NER with pretrained. Bioinformatics, 36, 295–302.

Weber, L. et al. (2021). HunFlair: an easy-to-use tool for state-of-the-art biomedical named entity recognition. Bioinformatics, 37, 2792–2794.