Comparative Analyses of Multilingual Drug Entity Recognition Systems for Clinical Case Reports In Cardiology

Notebook for the ICUE@MultiCardioNER submission at CLEF 2024

Chaeeun Lee[1], T. Ian Simpson[1], Joram M. Posma[2] and Antoine D. Lain[2,*]

[1] *School of Informatics, University of Edinburgh, 10 Crichton Street, EH8 9AB, Edinburgh, UK*

[2] *Section of Bioinformatics, Division of Systems Medicine, Department of Metabolism, Digestion, and Reproduction, Faculty of Medicine, Imperial College London, London W12 0NN, United Kingdom*

Abstract

Performance disparities exist in Named Entity Recognition (NER) systems across languages due to variations in available human-annotated data. We participated in the MultiDrug subtask of MultiCardioNER, a shared task focusing on multilingual NER for cardiology, to compare the effectiveness of fine-tuning BERT-based monolingual and multilingual language models, and prompting Large Language Models (LLMs) for drug entity recognition across multiple languages. Our findings demonstrate that monolingual BERT models pretrained on biomedical corpora generally outperform their multilingual counterparts. However, for languages lacking access to a broader range of pretrained models, combining the translation capability of LLM [1, 2, 3, 4] with the best-performing pretrained monolingual BERT model yielded superior results. This approach effectively reduces the resource disparity while leveraging domain-specific knowledge captured by the monolingual BERT model. Our best systems in the MultiCardioNER track yielded F1-scores of 0.9277 for Spanish, 0.9107 for English, and 0.8776 for Italian. We highlight the comparative advantages of domain-specific fine-tuning and LLM-powered language translation for multilingual drug NER.

Keywords

Natural Language Processing, Multilingual, Named Entity Recognition, Cardiology, BERT

1. Introduction

Named entity recognition (NER) is one of the central tasks in natural language processing (NLP), particularly in specialised domains such as biomedicine [5, 6, 7, 8]. Accurately identifying specific types of entities such as diseases and medications in text is crucial for extracting relevant information from healthcare-related text data. However, clinical NER systems often rely on the availability and quality of human-annotated resources, which can vary significantly across languages. Recognising the challenges in clinical entity recognition across different languages, the MultiCardioNER shared task [9] focuses on the recognition of disease and medication mentions in cardiological clinical case documents in English, Spanish, and Italian. As part of this initiative, we participated in the MultiDrug subtask, specifically targeting mentions of medications. Our participation was motivated by the goal of adapting and comparatively analysing machine-learning based NER systems in the cardiology domain. Our contributions are outlined as follows:

1. **Fine-tuning Monolingual Language Models:** We explored fine-tuning BERT-based monolingual language models for drug NER in each target language individually.
2. **Multilingual Model Capabilities:** We explored fine-tuning multilingual models on combined datasets of Spanish, English, and Italian and compared results with the monolingual approach.

CLEF 2024: Conference and Labs of the Evaluation Forum, September 09–12, 2024, Grenoble, France

[*] Corresponding author.

chaeeun.lee@ed.ac.uk (C. Lee); ian.simpson@ed.ac.uk (T. I. Simpson); jmp111@ic.ac.uk (J. M. Posma); a.lain@imperial.ac.uk (A. D. Lain)

3. **Large Language Model Integration:** We developed drug NER system using generative Large Language Model (LLM), providing insights on the potential use of LLM in tasks that conventionally rely on fine-tuning approaches.
4. **LLM as Translation Module:** We utilise translation capability of LLM to enhance cross-lingual applicability of best-performing monolingual models. This approach allowed us to test the effectiveness of LLM in bridging the gaps among different languages in clinical entity recognition, providing a way for monolingual systems to be applied to multilingual tasks.

2. Data Description

For the MultiDrug subtask [9], training and development datasets included two primary sources: the DisTEMIST [10, 11] and DrugTEMIST corpora, and the CardioCCC dataset.

2.1. DisTEMIST and DrugTEMIST Corpora

The DisTEMIST [10] and newly introduced DrugTEMIST [9] corpora comprised training datasets, including 1,000 Spanish clinical case documents across various medical fields such as oncology, pediatrics, and psychiatry, among others. The source documents were drawn from the SPACCC [12] corpus. The original Spanish documents were annotated manually by clinical experts [13] for medication mentions and transferred into English and Italian, ultimately comprising a multilingual dataset.

2.2. CardioCCC

As a domain-specific dataset for drug NER, CardioCCC dataset is comprised of 508 cardiology clinical case reports. The original Spanish documents were annotated under the same guidelines used for DisTEMIST[10]/DrugTEMIST datasets, and likewise transferred to English and Italian. The dataset was divided into development and test splits, consisting of 258 and 250 documents respectively.

2.3. Conversion to BIO2 Format for Fine-Tuning

For systems based on the fine-tuning BERT-based language models, we transformed the datasets into BIO2 (Begin, Inside, Outside) format [14]. After tokenisation, original annotations indicating the start and end offsets of drug mentions were converted so that each token was labeled with one of the following tags:

- **`B (Begin)`:** start of a named entity.
- **`I (Inside)`:** tokens within the entity.
- **`O (Outside)`:** tokens that are not part of any entity.

BIO2 format is particularly effective for encoder-based language models like BERT [15], which predict the classification for each token individually [16]. This allows for accurate identification of the boundaries of each drug entity.

3. System Description

Our strategy for the MultiDrug subtask was structured around two key dimensions: assessing the comparative effectiveness of fine-tuning encoder-based language models versus prompting generative LLM for NER, and exploring different multilingual processing approaches. For the latter, we evaluated the results from multilingual BERT model compared to those achieved by translating NER results from one language to others.

Table 1
Overview of submitted systems. Post-processing rules were applied to systems marked with "+ pp".

System	fine-tuned Models			LLM Translation
	Spanish (ES)	English (EN)	Italian (IT)	
System 1	bert-base-multilingual-cased [15] + pp	bert-base-multilingual-cased + pp	bert-base-multilingual-cased + pp	-
System 2	bsc-bio-ehr-es [17] + pp	scibert_scivocab_cased [18] + pp	bert-base-italian-xxl-cased [19] + pp	-
System 3	bsc-bio-ehr-es	scibert_scivocab_cased	bert-base-italian-xxl-cased	-
System 4	bsc-bio-ehr-es	scibert_scivocab_cased	-	ES → IT
System 5	bsc-bio-ehr-es	-	-	ES → EN, IT

3.1. Fine-tuning BERT-based models

Our main strategy involved fine-tuning BERT-based language models [15, 18, 20] specifically for drug NER. We explored a range of models, testing both general-purpose and domain-specific pretrained language models of various parameter sizes. Only the training set from DrugTEMIST corpora was used for fine-tuning. We then selected the best-performing models based on their performance on the development set derived from the CardioCCC corpus.

As a baseline approach, we utilised pretrained monolingual BERT-based models in Spanish, English, and Italian. This method ensures that each model is attuned to the linguistic nuances specific to a language, but has limitations in addressing resource disparities across different languages [20]. Each of these models was trained on the respective monolingual training set. As an alternative approach, we fine-tuned a multilingual BERT model on a combined training set of all three languages.

3.1.1. Multilingual BERT

In System 1, in view of the multilingual nature of the given dataset, we opted for using a pretrained multilingual BERT-based language model. Specifically, we employed “google-bert/bert-base-multilingual-cased” [15]. To enhance its performance for drug entity recognition, we further fine-tuned the model on the BIO2-formatted data in English, Italian, and Spanish provided by the organiser. Each document was split into individual sentences and a batch size of 32, learning rate of 8e-5, weight decay of 1e-5, and 5 epochs were employed during fine-tuning. Hyperparameter optimisation was not performed.

3.1.2. Monolingual BERT

Given the limited range of available multilingual BERT models, particularly the lack of biomedical domain-specific pretrained models, we investigated the effectiveness of fine-tuning independent monolingual models in Systems 2 and 3. This approach aimed to improve performance by leveraging language-specific pretrained biomedical models. However, availability of pretrained models differed across languages. While a wider range of selection was available for English and Spanish, a limited range of pretrained models available for Italian. Consequently, for Italian, we opted for “dbmdz/bert-base-italian-xxl-cased” [19]. To evaluate the efficacy of both general and domain-specific models available for the other two languages, we used the SeqEval library. This resulted in using “allenai/scibert_scivocab_cased” [18] for English and “PlanTL-GOB-ES/bsc-bio-ehr-es” [17] for Spanish. Hyperparameter search was not performed. Following the approach used for multilingual BERT, we applied sentence-level split of the data, a batch size of 32, learning rate of 8e-5, weight decay of 1e-5, and 5 epochs for all monolingual models.

3.2. Large Language Model Integration

In light of recent advancements in general-purpose LLMs, we explored their potential in biomedical multilingual NER. We employed two distinct strategies. The first strategy involved directly using the LLM for NER, bypassing the fine-tuning process. This approach aimed to assess the effectiveness of prompting LLM for multilingual NER tasks without domain-specific adaptation. The second strategy leveraged the best-performing monolingual BERT-based model for NER, while utilising LLM as a translation module to convert the NER results in a given language into the target language. This approach investigated the potential use of LLM to bridge the resource gaps among different languages while capitalising on the domain-specific knowledge captured by pretrained monolingual BERT models.

3.2.1. Drug NER with LLMs

In addition to fine-tuning BERT-based models for drug NER, we also experimented with directly using generative LLM for NER task. This approach involved prompting the LLM to produce drug annotations from clinical case reports without the conventional fine-tuning process. While this method takes advantage of the contextual understanding and embedded domain knowledge of LLMs [21, 22], it is recognised in existing literature that LLMs often fall short in achieving comparable performance on specific non-generative tasks such as NER [23].

Figure 1 illustrates an example of prompts used for LLM drug entity recognition. Our experiments were conducted in a zero-shot setting on the English development set, where we supplied the complete clinical text and tasked the LLM with identifying all drug mentions. We included an instruction in the prompt to output the predictions in JSON format for efficient post-processing. We experimented with three LLMs, `Meta-Llama-2-7B`, `Meta-Llama-3-8B`, and `gpt-3.5-turbo`. Llama-2-7B produced results that were significantly lower than the average scores of fine-tuned models, with a precision of 0.6689, recall of 0.1964, and F1 score of 0.3037. This outcome was expected due to the limited parameter size and the zero-shot setting. On the other hand, GPT-3.5 and Llama-3-8B demonstrated performances comparable to the average results from fine-tuned models (avg. precision: 0.8373; avg. recall: 0.8779; avg. F1: 0.8564). Comparing the two LLMs, GPT-3.5 achieved better recall (precision: 0.8236; recall: 0.8538; F1: 0.8384), while Llama-8B showed higher precision and F1 values (precision: 0.8767; recall: 0.8303; F1: 0.8529). LLaMA-3-8B achieved higher precision (0.8767) than the average precision of fine-tuned models (0.8373), which is noteworthy given that the LLaMA-3-8B model has significantly fewer parameters compared to GPT-3.5 and was used in a zero-shot setting. Further studies with a broader range of LLMs, as well as exploring few-shot settings and additional prompting methods, will help better understand the potential use of LLM in drug NER.

```
[
  {"role": "system", "content": "Your job is to review a clinical note that
  potentially contains mentions of drug names."},

  {"role": "user", "content":
  "Find all mentions of drug names in the following clinical note. Output your
  response in JSON format with keys 'drug 1', 'drug 2', and so on.

  Clinical Note:
  ANAMNESIS
  46-year-old Spanish woman.
  Married, with an 18-year-old daughter.
  Good family support.
  ..."}
]
```

Figure 1 – Example LLM Prompt for NER

3.2.2. Entity Translation

```
[
  {"role": "system", "content": "Your job is to review a clinical note that
  potentially contains mentions of drug names."},

  {"role": "user", "content":
  "I have Spanish drug names 'fluoxetina', 'clonazepam'. Find the corresponding
  drugs in English in the following clinical note. Output your response in JSON
  format, where the keys are the given Spanish drug names ('fluoxetina',
  'clonazepam'), and the values are the corresponding drug names in English found in
  the clinical note. Note that for every Spanish drug name, there is always at least
  one mention of the corresponding drug in English.

  Clinical Note:
  Francisca Valero, a 33-year-old stock market analyst, married with two children,
  was brought to the emergency department (ED) after 10 days of what her husband
  described as ..."}
]
```

Figure 2 – Example LLM Prompt for Translation

In addition to comparing systems that used three monolingual models and a single multilingual model, we explored the use of an LLM as a translation module to enable monolingual models to perform multilingual NER tasks. Figure 2 shows an example of prompts used for the LLM-powered translation. In System 4, we generated predictions for the Italian test set by translating the predictions from the best-performing Spanish monolingual model into Italian using GPT-3.5. For the English test set, we retained predictions from the English monolingual model. In System 5, predictions for both the English and Italian test sets were derived from the Spanish predictions and translated into the respective languages using GPT-3.5. This approach tests the feasibility of using an LLM-based translation system for multilingual NER tasks, especially when there is a disparity in available resources for each language, as an alternative to directly applying LLM for NER. [1, 2, 3, 4].

4. Results

```
filename                        label    start_span  end_span    text

casos_clinicos_cardiologia132    FARMACO   333    350    hidroclorotiazida
casos_clinicos_cardiologia132    FARMACO   359    369    olmersatan
casos_clinicos_cardiologia132    FARMACO   377    390    atorvastatina
casos_clinicos_cardiologia132    FARMACO   399    408    omeprazol
casos_clinicos_cardiologia132    FARMACO   2862    2875    noradrenalina
casos_clinicos_cardiologia132    FARMACO   2894    2904    dobutamina
```

Figure 3 – Example of the predictions on the testset

We submitted five drug entity recognition systems in all three languages (English, Italian, and Spanish). Details of these systems and the underlying BERT models are provided in Table 1. An example of the prediction format is presented in Figure 3, adhering to the formatting guidelines set by the organisers. For each system, three distinct files were generated, each corresponding to a different language. These files included the following details: filename, label, start span, end span, and text. However, during the system evaluation, only the filename, start span, and end span are utilised for comparison with the ground truth labels. System selection for final submission was based on performance on the development set (Table 2). The precision, recall, and F1 metrics are calculated based on a strict matching with the ground truth. In this approach, an exact match between the predicted and ground truth intervals is required to consider a true positive. The relaxed F1 score is computed based on a more lenient criterion. A true positive is counted if either the start of the prediction falls between or coincides with one of the start and end intervals present in the ground truth, or if a ground truth start falls between or coincides with one of the start and end intervals present in the prediction. This approach allows for some degree of imprecision in the predicted intervals while still considering them as correct. The relaxed F1 aims to understand the model predictions and if post-processing rules can be found to reduce the difference between the strict and relaxed scores.

- **Spanish:** Our monolingual BERT model achieved the top recall and F1 score (0.9277) among all participant submissions (mean F1: 0.6373; median F1: 0.8502).
- **English:** While our monolingual BERT model showed the highest precision, its F1 score (0.9107) was marginally lower than the best overall system (best F1: 0.9223; mean F1: 0.7101; median F1: 0.8768).
- **Italian:** In Italian, where model selection may be limited, we combined LLM translation with our Spanish monolingual system. This combined approach achieved the highest precision, while our monolingual Italian system showed the highest recall. Ultimately, the F1 score of our best Italian system (0.8776) was somewhat lower than the top-ranked submission (best F1: 0.8842; mean F1: 0.6506; median F1: 0.8421).

These results highlight the effectiveness of fine-tuning monolingual BERT models, particularly for high-resource languages with large amount of available domain-specific training data. The combination of LLM translation and monolingual models from other languages shows promise for low-resource languages (i.e., Italian) and requires further exploration.The results of our five submissions on the test set can be found in Table 3 alongside the best score for Precision, Recall and F1, the mean F1 and the median F1.

Table 2
Development Set Performance for model selection. Relaxed F1 is computed based on the overlap between the ground-truth and predicted spans. Post-processing rules were applied to systems marked with "+ pp". Values in bold represent the highest value for a given metric and a given language.

Language	Model	Precision	Recall	F1	Relaxed F1
English	BioBERT cased	0.8327	0.8486	0.8406	**0.9404**
	SciBERT cased	0.8617	0.8717	0.8667	**0.9404**
	SciBERT cased + pp	**0.9261**	0.8992	**0.9125**	-
	BERT base cased	0.8140	0.8371	0.8254	0.9156
	BERT multilingual base model cased	0.7624	0.8884	0.8206	0.9041
	BERT multilingual base model cased + pp	0.8271	**0.9227**	0.8723	-
Spanish	bsc-bio-ehr-es cased	0.7500	0.7665	0.7678	**0.9680**
	bsc-bio-ehr-es cased + pp	**0.9395**	**0.9406**	**0.9401**	-
	bert-base-spanish-wwm-uncased	0.7004	0.7422	0.7207	0.9275
	BERT multilingual base model cased	0.7785	0.9016	0.8355	0.9153
	BERT multilingual base model cased + pp	0.8370	0.9291	0.8807	-
Italian	bert-base-italian-xxl-cased	0.7969	0.8816	0.8371	**0.9227**
	bert-base-italian-xxl-cased + pp	**0.8665**	**0.9141**	**0.8897**	-
	BERT multilingual base model cased	0.7805	0.8778	0.8263	0.9067
	BERT multilingual base model cased + pp	0.8453	0.9091	0.8751	-
All	BERT multilingual base model cased	0.7742	0.8894	0.8278	**0.9242**
	BERT multilingual base model cased + pp	**0.8362**	**0.9203**	**0.8762**	-

Table 3
Test set results communicated by the challenge organisers. Values in bold represent the highest value obtained across all submissions. Values that are underlined represent the second-best performance according to our submissions against the best submission across all participants. The Mean and Median are calculated from all the participants of the challenge.

System	Spanish			English			Italian		
	Precision	Recall	F1	Precision	Recall	F1	Precision	Recall	F1
System 1	0.8287	0.9348	0.8786	0.8314	0.9343	0.8799	0.8139	0.9114	0.8487
System 2	0.9146	**0.9412**	**0.9277**	**0.9086**	0.9128	0.9107	0.8186	**0.9000**	0.8574
System 3	0.8777	0.9272	0.9018	0.8734	0.8977	0.8854	0.7879	0.8894	0.8356
System 4	0.9146	**0.9412**	**0.9277**	**0.9086**	0.9128	0.9107	**0.9114**	0.8461	0.8776
System 5	0.9146	**0.9412**	**0.9277**	0.8767	0.8635	0.8700	**0.9114**	0.8461	0.8776
Best submission	**0.9242**	**0.9412**	**0.9277**	**0.9086**	**0.9477**	**0.9223**	**0.9114**	**0.9000**	**0.8842**
Mean submission	-	-	0.6373	-	-	0.7101	-	-	0.6506
Median submission	-	-	0.8502	-	-	0.8768	-	-	0.8421

4.1. Error analysis

While the relaxed F1 scores in Table 2 are encouraging, we analysed the entities that overlapped partially with the gold standard annotations, suggesting a potential mismatch between system predictions and the human-annotated ground truth. We observed that the annotation guidelines provided to human annotators allowed for the inclusion of consecutive drug entities separated by specific delimiters like '/', 'and' (corresponding to 'y' in Spanish and 'e' in Italian), 'of' (corresponding to 'de' in Spanish and 'di' in Italian), '+', and '/ '. This specific format caused performance degradation for our system.

Interestingly, combining entities separated by '/' and 'of' (with its translation in Spanish and Italian) resulted in improved performance, while the impact of other delimiters was negligible. Additionally, we implemented a rule to include the entire span of a word entity even if our model only predicted a subset of the words. This adjustment aimed to further reduce the discrepancy between predicted and gold standard annotations.

It is worth noting that our strategy did not involve training on the validation set. Only the training split of the dataset, which is derived from the DrugTEMIST dataset, was used for fine-tuning, while the development and test sets were obtained from CardioCCC. There was no significant difference between the models' performances on the validation and test sets, suggesting that our systems likely did not exhibit signs of overfitting. Our analysis highlights the importance of considering relaxed F1 scores when evaluating NER systems. It also suggests that incorporating post-processing rules specific to the annotation guidelines employed can improve performance, particularly when dealing with specific entity formatting conventions.

5. Conclusion

In this study, we conducted a comparative analysis of multilingual drug NER in English, Italian, and Spanish. We evaluated the effectiveness of four distinct approaches: fine-tuning multilingual BERT, fine-tuning monolingual BERT models for each language, zero-shot LLM prompting, and a hybrid method that combines monolingual BERT and LLM as a translation module. Our findings revealed that fine-tuning monolingual BERT models generally outperformed other approaches. Specifically, our system for the Spanish test set, based on fine-tuning a BERT-based monolingual Spanish language model pretrained on a biomedical corpus, achieved the top ranking among all participant submissions.

On the other hand, for Italian, where there is limited availability of domain-specific human-annotated data and consequently a smaller range of pretrained models, a hybrid approach of combining best-performing monolingual model's prediction and LLM translation demonstrated greater efficacy compared to results achieved by a monolingual Italian BERT model alone. This suggests that LLM can be used as a translation module to enhance the cross-lingual applicability of fine-tuned monolingual models.

Promising results were obtained with zero-shot LLM prompting, where LLaMA-3-8B achieved higher precision (0.8767) than the average precision of fine-tuned models (0.8373) on English development set. Future work will involve experimenting with a broader range of LLMs and prompting methods in few-shot settings.

Funding

C.L. was supported by the United Kingdom Research and Innovation (grant EP/S02431X/1), UKRI Centre for Doctoral Training in Biomedical AI at the University of Edinburgh, School of Informatics. For the purpose of open access, the author has applied a creative commons attribution (CC BY) licence to any author accepted manuscript version arising. J.M.P. and A.D.L. are supported by the CoDiet project. The CoDiet project is funded by the European Union under Horizon Europe grant number 101084642 and supported by UK Research and Innovation (UKRI) under the UK government's Horizon Europe funding guarantee [grant number 101084642].

References

[1] S. Kumar, A. Anastasopoulos, S. Wintner, Y. Tsvetkov, Machine translation into low-resource language varieties, arXiv preprint arXiv:2106.06797 (2021).

[2] S. Ranathunga, E.-S. A. Lee, M. Prifti Skenduli, R. Shekhar, M. Alam, R. Kaur, Neural machine translation for low-resource languages: A survey, ACM Computing Surveys 55 (2023) 1–37.

[3] R. Koshkin, K. Sudoh, S. Nakamura, Transllama: Llm-based simultaneous translation system, arXiv preprint arXiv:2402.04636 (2024).

[4] H. Huang, S. Wu, X. Liang, B. Wang, Y. Shi, P. Wu, M. Yang, T. Zhao, Towards making the most of llm for translation quality estimation, in: CCF International Conference on Natural Language Processing and Chinese Computing, Springer, 2023, pp. 375–386.

[5] B. Song, F. Li, Y. Liu, X. Zeng, Deep learning methods for biomedical named entity recognition: a survey and qualitative comparison, Briefings in Bioinformatics 22 (2021) bbab282.

[6] P. N. Ahmad, A. M. Shah, K. Lee, A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain, in: Healthcare, volume 11, MDPI, 2023, p. 1268.

[7] D. F. Navarro, K. Ijaz, D. Rezazadegan, H. Rahimi-Ardabili, M. Dras, E. Coiera, S. Berkovsky, Clinical named entity recognition and relation extraction using natural language processing of medical free text: A systematic review, International Journal of Medical Informatics (2023) 105122.

[8] N. S. Pagad, N. Pradeep, Clinical named entity recognition methods: an overview, in: International Conference on Innovative Computing and Communications: Proceedings of ICICC 2021, Volume 2, Springer, 2022, pp. 151–165.

[9] S. Lima-López, E. Farré-Maduell, J. Rodríguez-Miret, M. Krallinger, MultiCardioNER Corpus: Multilingual Adaptation of Clinical NER Systems to the Cardiology Domain, 2024. URL: https://doi.org/10.5281/zenodo.11368861. doi:10.5281/zenodo.11368861.

[10] A. Miranda-Escalada, L. Gascó, S. Lima-López, E. Farré-Maduell, D. Estrada, A. Nentidis, A. Krithara, G. Katsimpras, G. Paliouras, M. Krallinger, Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources., in: CLEF (Working Notes), 2022, pp. 179–203.

[11] S. L. López, E. F. Maduell, L. G. Sánchez, M. Krallinger, MedProcNER Corpus: Gold Standard annotations for Clinical Procedures Information Extraction, 2023. URL: https://doi.org/10.5281/zenodo.8224056. doi:10.5281/zenodo.8224056.

[12] A. Intxaurrondo, M. Krallinger, Spaccc, 2019. URL: https://doi.org/10.5281/zenodo.2560316. doi:10.5281/zenodo.2560316.

[13] S. Lima-López, E. Farré-Maduell, M. Krallinger, DrugTEMIST Guidelines: Annotation of Medication in Medical Documents, 2024. URL: https://doi.org/10.5281/zenodo.11065433. doi:10.5281/zenodo.11065433.

[14] L. Ramshaw, M. Marcus, Text chunking using transformation-based learning, in: Third Workshop on Very Large Corpora, 1995. URL: https://aclanthology.org/W95-0107.

[15] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, Bert: Pre-training of deep bidirectional transformers for language understanding, arXiv preprint arXiv:1810.04805 (2018).

[16] R. Sharma, D. Chauhan, R. Sharma, Named entity recognition system for the biomedical domain, in: 2022 17th Conference on Computer Science and Intelligence Systems (FedCSIS), IEEE, 2022, pp. 837–840.

[17] C. P. Carrino, J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, M. Villegas, Pretrained biomedical language models for clinical NLP in Spanish, in: Proceedings of the 21st Workshop on Biomedical Language Processing, Association for Computational Linguistics, Dublin, Ireland, 2022, pp. 193–199. URL: https://aclanthology.org/2022.bionlp-1.19. doi:10.18653/v1/2022.bionlp-1.19.

[18] I. Beltagy, K. Lo, A. Cohan, Scibert: A pretrained language model for scientific text, in: EMNLP, Association for Computational Linguistics, 2019. URL: https://www.aclweb.org/anthology/D19-1371.

[19] S. Schweter, Italian bert and electra models, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[20] K. Hakala, S. Pyysalo, Biomedical named entity recognition with multilingual BERT, in: K. Jin-Dong, N. Claire, B. Robert, D. Louise (Eds.), Proceedings of the 5th Workshop on BioNLP Open Shared Tasks, Association for Computational Linguistics, Hong Kong, China, 2019, pp. 56–61. URL: https://aclanthology.org/D19-5709. doi:10.18653/v1/D19-5709.

[21] S. Wang, X. Sun, X. Li, R. Ouyang, F. Wu, T. Zhang, J. Li, G. Wang, Gpt-ner: Named entity recognition via large language models, arXiv preprint arXiv:2304.10428 (2023).

[22] D. Ashok, Z. C. Lipton, Promptner: Prompting for named entity recognition, arXiv preprint arXiv:2305.15444 (2023).

[23] I. Jahan, M. T. R. Laskar, C. Peng, J. X. Huang, A comprehensive evaluation of large language models on benchmark biomedical text processing tasks, Computers in Biology and Medicine (2024) 108189.