LSI_UNED at #SMM4H–HeaRD 2026: Grid-Based Biomedical Named Entity Recognition Across Languages and Entity Types

Alicia Ramirez-Arrabe[1*] aramirez@lsi.uned.es
Juan Martinez-Romo[1,2] juaner@lsi.uned.es
Andres Duque[1,2] aduque@lsi.uned.es

[1]NLP & IR Group, Dpto. Lenguajes y Sistemas Informáticos, Universidad Nacional de Educación a Distancia (UNED), Madrid, 28040, Spain

[2]Instituto Mixto de Investigación - Escuela Nacional de Sanidad (IMIENS), Madrid, 28029, Spain

Abstract

This paper describes the participation of the LSI_UNED team in the firt sub-task of MultiClinAI at the #SMM4H-HeaRD 2026 Workshop, which focuses on multilingual clinical named entity recognition in seven languages. The task requires identifying mentions of diseases, procedures, and symptoms in clinical case reports. We propose a set of systems based on the W^{2}NER architecture, with a separate model trained for each language and entity type. For Spanish, we use a RoBERTa-based model with data augmentation from additional NER resources, while English and Italian systems are based on different biomedical BERT variants. Results show consistent performance across languages, with the best overall results obtained for Spanish. Data augmentation improves recall and F1, while English and Italian models achieve competitive but slightly lower scores. Symptom recognition remains the most challenging entity type across all languages.

1 Introduction

The Social Media Mining for Health Applications and Health Real-World Data (#SMM4H-HeaRD) 2026 Workshop (Lopez-Garcia et al., 2026) encompasses eight shared tasks that focus on natural language processing (NLP), machine learning, and artificial intelligence using health-related data sources. Task 8, called MultiClinAI (Gallego-Donoso et al., 2026), addresses the growing need for multilingual corpora by evaluating systems that recognize medical concepts across seven languages. Its first sub-task, MultiClinNER, focuses on identifying diseases, procedures, and symptoms in clinical case reports written in Spanish, English, Dutch, Italian, Romanian, Swedish, and Czech. Participants must detect entity mentions and return their exact character offsets.

This paper describes the participation of our team, LSI_UNED, in MultiClinNER for the Spanish, English, and Italian languages. Because of time constraints, we could not apply the system to all seven languages, so we selected the three with which we are most familiar. We used the W^{2}NER (Li et al., 2022) architecture, adapted for the task. The fine-tuned systems were configured for a single-label and monolingual task. For the Spanish texts, we used a RoBERTa-based pre-trained model and also augmented the training set with additional NER resources. For English and Italian, two different BERT-based models were fine-tuned.

2 Related work

Named Entity Recognition (NER) is a widely addressed NLP task that converts unstructured text into structured information by detecting and classifying entities. In the medical domain, it is essential for identifying healthcare-related concepts and supports many downstream applications (Goyal and Singh, 2025; Ahmad et al., 2023).

In Spanish, several benchmarks address clinical entity extraction, including DisTEMIST (Miranda-Escalada et al., 2022) for diseases, MedProcNER (Lima-López et al., 2023) for procedures, SympTEMIST (Lima-López et al., 2023) for symptoms, and a sub-task of Cantemist (Miranda-Escalada et al., 2020) for tumor morphology. In English, GENIA (Kim et al., 2003) focuses on biological terms, MedMentions (Mohan and Li, 2019) on a broad range of biomedical entities, and CHEMDNER (Krallinger et al., 2015) on chemical and drug-related mentions. Italian has fewer resources, such as CLARO (Paolo et al., 2023), which covers 25 clinical entity types. Furthermore, as previously mentioned, the MultiClinAI shared task covers a total of seven languages, making it a valuable setting in the multilingual biomedical field. Other multilingual clinical NER corpora include MANTRA (Kors et al., 2015), which aligns biomedical concepts across English, Spanish, French, German,

*Corresponding author.

Proceedings of the 11th Social Media Mining for Health Research and Applications (SMM4H-HeaRD 2026) Workshop and Shared Tasks, pages 12–17

English			Spanish			Italian		
Label	**Total**	**Unique**	**Label**	**Total**	**Unique**	**Label**	**Total**	**Unique**
DISEASE	25,118	11,783	DISEASE	26,296	13,097	DISEASE	26,159	12,584
PROCEDURE	26,733	9,746	PROCEDURE	28,137	10,642	PROCEDURE	27,394	9,936
SYMPTOM	27,465	13,878	SYMPTOM	29,074	14,798	SYMPTOM	27,929	14,156
Total	79,316	35,407	**Total**	83,507	38,537	**Total**	81,482	36,676

Table 1: Statistics of the MultiClinNER corpus across the three evaluated languages (English, Spanish, and Italian), including the number of total and unique entity mentions for the three labels: DISEASE, PROCEDURE, and SYMPTOM.

and Dutch, and the E3C corpus (Magnini et al., 2020), a collection of clinical case reports in English, French, Italian, Spanish, and Basque.

Several architectures have been proposed for NER both in the general and biomedical domains. The most common approach is token-level sequence labeling, where each token is assigned a tag using schemes such as IOB2 or BILOU, with models based on CRF, BiLSTM-CRF, or Transformers (Zhang et al., 2018; Pooja and Jagadeesh, 2024). Another approach is span-based NER, which predicts entity boundaries directly instead of labeling tokens one by one. In this group, grid-based models such as W^{2}NER (Li et al., 2022) represent word-pair relations in a two-dimensional grid, which helps to detect nested or complex entities. More recent approaches include generative NER, which reformulates entity extraction as a text generation task (Yuan et al., 2022; Lv et al., 2025).

3 Methodology

3.1 Dataset

The MultiClinNER corpus (Lima López et al., 2026) contains 1,258 documents per language. For each language, it is annotated with three medical concepts: diseases, procedures, and symptoms. Table 1 presents the number of total and unique entities per label across the three languages for which our team submitted results. As can be observed, there are more than 25,000 mentions per entity type in each language. Moreover, the label with the fewest unique mentions is PROCEDURE, whereas SYMPTOM is the label with the highest number of both total and unique mentions, possibly indicating that it is the most challenging entity type.

3.2 System description

3.2.1 Grid-based NER method

The NER architecture used in this work is based on W^{2}NER, proposed in (Li et al., 2022). Unlike traditional sequential tagging approaches, W^{2}NER reformulates NER as a word-pair relation classification task using a two-dimensional grid. Each cell in the grid represents a pair of words, and the model predicts relations such as Next-Neighboring-Word (NNW), which links adjacent words within the same entity, and Tail-Head-Word (THW), which connects the last and first words of an entity while assigning its label. The architecture combines BERT embeddings, a BiLSTM for contextual encoding, and 2D convolution layers to capture interactions between word pairs. Entities are then reconstructed by linking words through the predicted relations. The original work reports state-of-the-art results on 14 NER benchmark datasets.

3.2.2 Pre-trained model selection and training

Different pre-trained models were used for each language. All of them are monolingual and pre-trained on biomedical texts. They are described below.

- English: For the English texts, two pre-trained models were used. The first was biobert-v1.1 (110M parameters) (Lee et al., 2020), which was pre-trained on large biomedical corpora, including PubMed abstracts and PubMed Central full-text articles. The second was Bio_ClinicalBERT (110M parameters) (Alsentzer et al., 2019), which was initialized from the former model and further pre-trained on clinical notes from the MIMIC-III database (Johnson et al., 2016).

- Spanish: The pre-trained model used for the Spanish texts was roberta-base-biomedical-clinical-es (125M parameters) (Carrino et al., 2021). It is based on the Spanish RoBERTa-base model and was further pre-trained on multiple clinical Spanish corpora collected from public sources and a dataset containing more than 278,000 documents.

- Italian: The pre-trained models used for the Italian texts were bioBIT (100M parameters)

(Buonocore et al., 2023), trained on Italian medical texts, and MedPsyNIT (100M parameters) (Crema et al., 2023), developed for medical and psychological text processing in Italian.

Due to time constraints, we were unable to perform hyperparameter optimization. The parameters were set to a batch size of 8, a learning rate of $1e^{-3}$, and a maximum sequence length of 512 tokens per chunk, with chunks segmented at the sentence level. Nevertheless, we conducted 5-fold cross-validation to select the optimal number of training epochs. With a maximum of 35 epochs, the best micro F1 score was consistently reached at approximately 20 epochs on average across folds. Experiments were conducted on a system equipped with an NVIDIA A30 GPU featuring 24 GB HBM2 ECC memory. Fine-tuning each model took 8 to 10 hours.

3.2.3 Data augmentation

For the Spanish corpus, we decided to augment the training set because several biomedical NER resources in Spanish have been published in recent years. Data augmentation was not performed for the English and Italian corpora, as we did not find corpora annotated with the same entity types required for the task. We used the corpora from the DisTEMIST, MedProcNER, and SympTEMIST tasks (described in Section 2) to augment the training data for diseases, procedures, and symptoms, respectively. Furthermore, we employed the CARMEN-I resource (Farre Maduell et al., 2024), which also includes annotations for the three entity types. The CARMEN-I corpus contains documents in Spanish, Catalan, and bilingual texts. Therefore, we selected only the texts written in Spanish and incorporated their annotations to further augmenting the training set. After the evaluation phase, we noticed that a subset of the added documents overlapped with the original texts, resulting in only 33–35% of the added documents being unseen. Nonetheless, the results in Section 4 show that data augmentation still improves the performance of the system.

4 Experiments and results

The evaluation of the task considers six metrics in total. The *strict* metrics (strict_precision, strict_recall, strict_f1) measure exact match performance, where a prediction is only correct if it perfectly matches the gold annotation in entity type and span boundaries. The *char* metrics (char_precision, char_recall, char_f1) evaluate character-level overlap between predicted and gold spans, where precision is the proportion of the predicted span that overlaps with the gold span and recall is the proportion of the gold span covered by the prediction.

Table 2 shows the final results, reporting, for each language and label, the values obtained for the six evaluation metrics considered. Results are ranked by strict_f1, and the numbers in parentheses indicate the ranking achieved in the shared task. For each language, two different runs were submitted. Each system is single-label, meaning that the models were fine-tuned separately for each entity type. The main reason for this was that the English and Italian texts showed slight variations across labels, so the offset alignments were not identical for each entity type. For Spanish, the training texts were correctly aligned across labels, but experiments with a multi-label system in a 5-fold cross-validation setting produced lower results than the single-label approach, so we chose the simpler and more consistent method.

For the English documents, the system that employs the model biobert-v1.1 (conf1_biobert) achieved better results across the three labels, with an improvement of around 0.02 in strict_f1 compared with the system fine-tuned from Bio_ClinicalBERT (conf1_clinical). The system conf1_biobert achieved its best ranking in strict_recall for the label PROCEDURE, where it placed second.

With respect to the Spanish systems, those trained with the augmented training set (conf2) reached a higher strict_f1 than those using the original training set (conf1) across the three labels. In such cases, we can see that the strict_recall is higher than in conf1, especially in DISEASE and SYMPTOM, suggesting that the data augmentation allows the model to identify more entities. For the label DISEASE, the system conf2 ranked first in strict_recall, while for all three labels it achieved third place in strict_f1.

Regarding the Italian language, the system fine-tuned on the model bioBIT (conf1_biobit) achieved better strict_f1 in the labels DISEASE and SYMPTOM, while the system using the model MedPsyNIT (conf1_medpsynit) reached a better score in PROCEDURE. However, the differences in strict_f1 scores are lower than 0.01 in all cases. The system conf1_biobit achieved its best ranking

English							
Entity	**System**	**strict_precision**	**strict_recall**	**strict_f1**	**char_precision**	**char_recall**	**char_f1**
DISEASE	conf1_biobert	0.7633 (7)	0.7914 (3)	**0.7771** (5)	0.8421 (8)	0.8523 (8)	**0.8472** (7)
DISEASE	conf1_clinical	0.7574 (8)	0.7572 (9)	0.7573 (8)	0.8395 (9)	0.8226 (13)	0.8309 (12)
PROCEDURE	conf1_biobert	0.7584 (4)	0.7293 (2)	**0.7436** (3)	0.8595 (7)	0.8074 (7)	**0.8326** (6)
PROCEDURE	conf1_clinical	0.7531 (5)	0.6979 (9)	0.7244 (6)	0.8516 (8)	0.7739 (10)	0.8109 (10)
SYMPTOM	conf1_biobert	0.7433 (6)	0.6802 (7)	**0.7104** (6)	0.8308 (6)	0.7409 (10)	**0.7833** (6)
SYMPTOM	conf1_clinical	0.7397 (7)	0.6609 (10)	0.6981 (7)	0.8272 (7)	0.7220 (12)	0.7710 (9)
Spanish							
Entity	**System**	**strict_precision**	**strict_recall**	**strict_f1**	**char_precision**	**char_recall**	**char_f1**
DISEASE	conf2	0.7852 (5)	0.8432 (1)	**0.8132** (3)	0.8495 (5)	0.8973 (3)	**0.8728** (4)
DISEASE	conf1	0.7925 (4)	0.8231 (5)	0.8075 (5)	0.8545 (4)	0.8758 (8)	0.8650 (7)
PROCEDURE	conf2	0.8018 (2)	0.8049 (4)	**0.8034** (3)	0.8730 (2)	0.8662 (10)	**0.8696** (4)
PROCEDURE	conf1	0.8011 (3)	0.8004 (6)	0.8008 (4)	0.8699 (4)	0.8559 (11)	0.8628 (7)
SYMPTOM	conf2	0.7472 (4)	0.7667 (4)	**0.7568** (3)	0.8226 (4)	0.8335 (5)	**0.8280** (4)
SYMPTOM	conf1	0.7626 (2)	0.7488 (5)	0.7556 (4)	0.8365 (2)	0.8066 (9)	0.8213 (5)
Italian							
Entity	**System**	**strict_precision**	**strict_recall**	**strict_f1**	**char_precision**	**char_recall**	**char_f1**
DISEASE	conf1_biobit	0.7058 (8)	0.7245 (2)	**0.7150** (5)	0.8202 (9)	0.8188 (4)	**0.8195** (5)
DISEASE	conf1_medpsynit	0.7163 (7)	0.7102 (4)	0.7132 (6)	0.8250 (7)	0.8014 (6)	0.8130 (7)
PROCEDURE	conf1_medpsynit	0.7124 (5)	0.6320 (9)	**0.6698** (7)	0.8502 (3)	0.7436 (10)	0.7933 (9)
PROCEDURE	conf1_biobit	0.6888 (7)	0.6377 (8)	0.6623 (9)	0.8428 (6)	0.7636 (9)	**0.8013** (8)
SYMPTOM	conf1_biobit	0.6898 (5)	0.6167 (5)	**0.6512** (4)	0.8170 (4)	0.7076 (9)	0.7584 (7)
SYMPTOM	conf1_medpsynit	0.6802 (6)	0.6241 (4)	0.6509 (6)	0.8119 (6)	0.7187 (8)	**0.7624** (6)

Table 2: Final results of the system for each label and language, ranked by strict_f1. The numbers in parentheses indicate the ranking achieved for each metric with respect to the other participating teams. For each language and label, the best strict_f1 and char_f1 results are highlighted in bold.

in strict_recall for the label DISEASE, where it placed second.

To conclude, the overall results seem consistent and robust across languages and labels. In general, the obtained results in the Spanish corpus were the best ones. Moreover, the last three metrics are higher than the first three in all cases, as they consider the character-overlap of the predictions instead of counting non-exact matchings as errors. Furthermore, it can be observed that in most cases, the models make more mistakes and obtain lower results when detecting entities corresponding to the label SYMPTOM. This aligns with the hypothesis made in Subsection 3.1, referring to the high number of unique mentions presented by this label.

5 Conclusions

This work presents the participation of the LSI_UNED team in the MultiClinNER sub-task for Spanish, English, and Italian using the W^{2}NER architecture. The submitted systems followed a monolingual and single-label strategy. Overall, the results were consistent across languages, with the best performance obtained for Spanish, where data augmentation with external NER resources improved the system's performance and obtained a third place for the three labels in strict_f1 (including the first place in strict_recall for the entity DISEASE). English and Italian also produced competitive results, although symptom detection remained the most challenging category.

One of the main limitations of this work is that no hyperparameter optimization was performed due to time constraints. As a result, the selected configurations may not represent the optimal performance of the proposed approach.

Future work will explore more complex modeling strategies, including multi-label and multilingual configurations. Another direction would be to compare W^{2}NER with alternative NER approaches, such as traditional token-level sequence labeling using IOB2 or BILOU tagging schemes, as well as more recent architectures like GLiNER (Zaratiana et al., 2024). In addition, it would be interesting to carry out hyperparameter optimization using cross-validation to identify the optimal combination of hyperparameters.

Acknowledgments

This work has been partially supported by the Spanish Ministry of Science and Innovation within the EDHER-MED Project under grant PID2022-136522OB-C21 as well as by the Universidad Nacional de Educación a Distancia (UNED), Spain within project SICAMESP (2023-VICE-0029).

References

Pir Noman Ahmad, Adnan Muhammad Shah, and KangYoon Lee. 2023. A review on electronic health record text-mining for biomedical name entity recognition in healthcare domain. In *Healthcare*, volume 11, page 1268. MDPI.

Emily Alsentzer, John R. Murphy, Willie Boag, Wei-Hung Weng, Di Jin, Tristan Naumann, and Matthew B. A. McDermott. 2019. Publicly available clinical BERT embeddings. *CoRR*, abs/1904.03323.

Tommaso Mario Buonocore, Claudio Crema, Alberto Redolfi, Riccardo Bellazzi, and Enea Parimbelli. 2023. Localizing in-domain adaptation of transformer-based biomedical language models. *Journal of Biomedical Informatics*, 144:104431.

Casimiro Pio Carrino, Jordi Armengol-Estapé, Asier Gutiérrez-Fandiño, Joan Llop-Palao, Marc Pàmies, Aitor Gonzalez-Agirre, and Marta Villegas. 2021. Biomedical and clinical language models for spanish: On the benefits of domain-specific pretraining in a mid-resource scenario. *CoRR*, abs/2109.03570.

Claudio Crema, Tommaso Mario Buonocore, Silvia Fostinelli, Enea Parimbelli, Federico Verde, Cira Fundarò, Marina Manera, Matteo Cotta Ramusino, Marco Capelli, Alfredo Costa, and 1 others. 2023. Advancing italian biomedical information extraction with transformers-based models: Methodological insights and multicenter practical application. *Journal of Biomedical Informatics*, 148:104557.

Eulalia Farre Maduell, Salvador Lima-Lopez, Santiago Andres Frid, Artur Conesa, Elisa Asensio, Antonio Lopez-Rueda, Helena Arino, Elena Calvo, Maria Jesús Bertran, Maria Angeles Marcos, Montserrat Nofre Maiz, Laura Tañá Velasco, Antonia Marti, Ricardo Farreres, Xavier Pastor, Xavier Borrat Frigola, and Martin Krallinger. 2024. CARMEN-I: A resource of anonymized electronic health records in Spanish and Catalan for training and testing NLP tools. *PhysioNet*. Version 1.0.1.

Fernando Gallego-Donoso, Salvador Lima-López, Judith Rosell, Eulàlia Farré-Maduell, and Martin Krallinger. 2026. The MultiClinAI Shared Task on Multilingual Clinical Corpus Construction and Concept Extraction: Systems, Evaluation, and Datasets. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Nandita Goyal and Navdeep Singh. 2025. Named entity recognition and relationship extraction for biomedical text: A comprehensive survey, recent advancements, and future research directions. *Neurocomputing*, 618:129171.

Alistair EW Johnson, Tom J Pollard, Lu Shen, Li-wei H Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G Mark. 2016. Mimic-iii, a freely accessible critical care database. *Scientific data*, 3(1):1–9.

J-D Kim, Tomoko Ohta, Yuka Tateisi, and Jun'ichi Tsujii. 2003. Genia corpus—a semantically annotated corpus for bio-textmining. *Bioinformatics*, 19(suppl_1):i180–i182.

Jan A Kors, Simon Clematide, Saber A Akhondi, Erik M Van Mulligen, and Dietrich Rebholz-Schuhmann. 2015. A multilingual gold-standard corpus for biomedical concept recognition: the mantra gsc. *Journal of the American Medical Informatics Association*, 22(5):948–956.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, and 1 others. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of cheminformatics*, 7(Suppl 1):S2.

Jinhyuk Lee, Wonjin Yoon, Sungdong Kim, Donghyeon Kim, Sunkyu Kim, Chan Ho So, and Jaewoo Kang. 2020. Biobert: a pre-trained biomedical language representation model for biomedical text mining. *Bioinform.*, 36(4):1234–1240.

Jingye Li, Hao Fei, Jiang Liu, Shengqiong Wu, Meishan Zhang, Chong Teng, Donghong Ji, and Fei Li. 2022. Unified named entity recognition as word-word relation classification. In *Thirty-Sixth AAAI Conference on Artificial Intelligence, AAAI 2022, Thirty-Fourth Conference on Innovative Applications of Artificial Intelligence, IAAI 2022, The Twelveth Symposium on Educational Advances in Artificial Intelligence, EAAI 2022 Virtual Event, February 22 - March 1, 2022*, pages 10965–10973. AAAI Press.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gascó, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2023. Overview of medprocner task on medical procedure detection and entity linking at bioasq 2023. In *Working Notes of the Conference and Labs of the Evaluation Forum (CLEF 2023), Thessaloniki, Greece, September 18th to 21st, 2023*, CEUR Workshop Proceedings, pages 1–18. CEUR-WS.org.

Salvador Lima-López, Eulàlia Farré-Maduell, Luis Gasco-Sánchez, Jan Rodríguez-Miret, and Martin Krallinger. 2023. Overview of symptemist at biocreative viii: corpus, guidelines and evaluation of systems for the detection and normalization of symptoms, signs and findings from text. In *Proceedings of the BioCreative VIII Challenge and Workshop: Curation and Evaluation in the era of Generative Models*, page 11.

Salvador Lima López, Judith Rosell, Jan Rodríguez Miret, Fernando Gallego-Donoso, and Martin Krallinger. 2026. Multiclinai shared task training data.

Guillermo Lopez-Garcia, Jose Miguel Acitores Cortina, Jacob Berkowitz, Joey Chan, Ganesh Chandrasekar, Sumon Kanti Dey, Ivan Flores Amaro, Fernando Gallego, Lauren Gryboski, Ari Z Klein, Martin Krallinger, Salvador Lima-López, Tomohiro

Nishiyama, Lisa Raithel, Ahmad Rezaie Mianroodi, Roland Roller, Judith Rosell, Frank Rudzicz, Abeed Sarker, and 8 others. 2026. Overview of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Shared Tasks at ACL 2026. In *Proceedings of the 11th Social Media Mining for Health (#SMM4H) and Health Real-World Data (HeaRD) Workshop and Shared Tasks*. Association for Computational Linguistics.

Tengxiao Lv, Ling Luo, Juntao Li, Yanhua Wang, Yuchen Pan, Chao Liu, Yanan Wang, Yan Jiang, Huiyi Lv, Yuanyuan Sun, and 1 others. 2025. A unified biomedical named entity recognition framework with large language models. *arXiv preprint arXiv:2510.08902*.

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Manuela Speranza, and Roberto Zanoli. 2020. The e3c project: Collection and annotation of a multilingual corpus of clinical cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics (CLiC-it 2020)*, pages 190–196.

Antonio Miranda-Escalada, Eulàlia Farré, and Martin Krallinger. 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. *IberLEF@ SEPLN*, pages 303–323.

Antonio Miranda-Escalada, Luis Gascó, Salvador Lima-López, Eulàlia Farré-Maduell, Darryl Estrada, Anastasios Nentidis, Anastasia Krithara, Georgios Katsimpras, Georgios Paliouras, and Martin Krallinger. 2022. Overview of distemist at bioasq: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *Proceedings of the Working Notes of CLEF 2022 - Conference and Labs of the Evaluation Forum, Bologna, Italy, September 5th - to - 8th, 2022*, CEUR Workshop Proceedings, pages 179–203. CEUR-WS.org.

Sunil Mohan and Donghui Li. 2019. Medmentions: A large biomedical corpus annotated with umls concepts. *arXiv preprint arXiv:1902.09476*.

Domenico Paolo, Alessandro Bria, Carlo Greco, Marco Russano, Sara Ramella, Paolo Soda, and Rosa Sicilia. 2023. Named entity recognition in italian lung cancer clinical reports using transformers. In *2023 IEEE International Conference on Bioinformatics and Biomedicine (BIBM)*, pages 4101–4107. IEEE.

H Pooja and MP Prabhudev Jagadeesh. 2024. A deep learning based approach for biomedical named entity recognition using multitasking transfer learning with bilstm, bert and crf. *SN Computer Science*, 5(5):482.

Hongyi Yuan, Zheng Yuan, Ruyi Gan, Jiaxing Zhang, Yutao Xie, and Sheng Yu. 2022. Biobart: Pretraining and evaluation of a biomedical generative language model. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 97–109.

Urchade Zaratiana, Nadi Tomeh, Pierre Holat, and Thierry Charnois. 2024. Gliner: Generalist model for named entity recognition using bidirectional transformer. In *Proceedings of the 2024 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies (Volume 1: Long Papers)*, pages 5364–5376.

Yuan Zhang, Hongshen Chen, Yihong Zhao, Qun Liu, and Dawei Yin. 2018. Learning tag dependencies for sequence tagging. In *IJCAI*, pages 4581–4587.