Procesamiento del Lenguaje Natural, Revista nº 71, septiembre de 2023, pp. 313-320 recibido 30-06-2023 revisado 07-07-2023 aceptado 19-07-2023

# Overview of TESTLINK at IberLEF 2023: Linking Results to Clinical Laboratory Tests and Measurements

## *Resumen de TESTLINK en IberLEF 2023: Creación de relaciones entre análisis de laboratorio y mediciones clínicas y sus resultados*

**Begoña Altuna,**[1] **Rodrigo Agerri,**[1] **Lidia Salas-Espejo,**[2] **José Javier Saiz,**[3] **Alberto Lavelli,**[4] **Bernardo Magnini,**[4] **Manuela Speranza,**[4] **Roberto Zanoli,**[4] **Goutham Karunakaran**[5]
[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU
[2]University of the Basque Country UPV/EHU
[3]Barcelona Supercomputing Center (BSC)
[4]Fondazione Bruno Kessler, [5]University of Trento
{begona.altuna, rodrigo.agerri}@ehu.eus, lsalas007@ikasle.ehu.eus
josejavier.saiz.anton@gmail.com, {lavelli, magnini, manspera, zanoli}@fbk.eu
goutham.karunakaran@studenti.unitn.it

**Abstract:** The TESTLINK task conducted at IberLEF2023 focuses on relation extraction from clinical cases in Spanish and Basque. The task consists in identifying clinical results and measures and linking them to the tests and measurements from which they were obtained. Three teams took part in the task and various (supervised) deep learning models were evaluated; interestingly, none of the teams explored the use of few-shot learning. The evaluation shows that in-domain fine-tuning and larger training datasets improve the results. In fact, the fact that supervised models significantly outperformed the baseline based on few-shot learning shows the crucial role still played by the availability of annotated training data.
**Keywords:** Named Entity Recognition, Information Extraction, Clinical NLP, Supervised Learning.

**Resumen:** La tarea TESTLINK de IberLEF2023 se centra en la extracción de relaciones de casos clínicos en español y euskera. La tarea consiste en identificar resultados y medidas clínicas y relacionarlos con las pruebas y mediciones de las que se obtuvieron. Tres equipos han participado en la tarea y se han evaluado varios modelos (supervisados) de aprendizaje profundo. Curiosamente, ninguno de los equipos exploró el uso del aprendizaje few-shot. La evaluación muestra que el fine-tuning en el dominio y conjuntos de datos de entrenamiento más grandes mejoran los resultados. De hecho, el hecho de que los modelos supervisados superaran significativamente la baseline basada en el aprendizaje few-shot muestra el papel crucial que aún desempeña la disponibilidad de datos de entrenamiento anotados.
**Palabras clave:** Reconocimiento de Entidades Nombradas, Extracción de Información, PLN Clínico, Aprendizaje Supervisado.

### *1 Introduction*

There is a growing interest in processing clinical data for tasks of public interest, such as clinical decision making (Jain and Prajapati, 2021) or monitoring of the health status of a country (Sankoh and Byass, 2014). While for this purpose large amounts of structured data are needed, the reality is that most clinical data are stored as free unstructured clinical texts. Hence, the ability of extracting information directly from natural language texts and to increase the volume of databases and structured datasets, such as MIMIC-III (Johnson et al., 2016), is crucial.

Having these goals into account, scholars have developed a series of resources for information extraction from clinical texts. Clinical information extraction efforts have often

ISSN 1135-5948 DOI 10.26342/2023-71-24 

given priority to the identification of diseases (Trigueros et al., 2022) or events (Santiso, Pérez, and Casillas, 2021). As far as the extraction of relations from clinical texts is concerned, previous work has focused on concept normalization (Newman-Griffis et al., 2020) and temporal relations (Alfattni, Peek, and Nenadic, 2020), among others. Laboratory tests and measurements and their results have been given little attention (Hao, Liu, and Weng, 2016), although they provide interesting information on the patients' status in a certain time of the development of a disorder and are crucial to choose the right diagnose. From a more technical point of view, processing laboratory tests and their results also brings up a new perspective on the treatment of data, since it requires interpreting numeric values and ranges and therefore can not be handled as a common named entity recognition (NER) task (Percha, 2021).

In this context, we have organised TESTLINK at IberLEF 2023 (Jiménez-Zafra, Rangel, and Montes-y Gómez, 2023). The TESTLINK task provides an opportunity to evaluate different Natural Language Processing approaches and does this with a focus on languages other than English, specifically Spanish and Basque, that are less resourced and are typologically diverse.

## 2 Task Description

The TESTLINK task consists in identifying textual mentions of both laboratory tests and their results in a clinical narrative, and then linking tests to their respective results. In essence, the TESTLINK task primarily focuses on the extraction of relations (RE) between entities. Clinical narratives (or clinical cases) are documents reporting statements of a clinical practice, presenting the reason for a clinical visit, the description of physical exams, and the assessment of the patient's situation. Laboratory tests and measurements are commonly done as part of the diagnosis process and are documented in clinical narratives. In the sample clinical case in Figure 1, laboratory tests have been marked in bold and their reported results in italics.

The following relations have been hold between tests and results:

- 6 ng/ml -> the amounts of PSA / las cifras de PSA
- 12 ng/ml -> the amounts of PSA / las cifras de PSA
- negativa / negative -> a transrectal prostate BIOPSY guided by sextants / una BIOPSIA transrectal de próstata ecodirigida por sextantes
- 0,80 -> a Ch-Cr/Ci INDEX / un ÍNDICE de Ch-Cr/Ci

Paciente de 65 a. de edad, que presentaba una elevación progresiva de **las cifras de PSA** desde *6 ng/ml* a *12 ng/ml* en el último año. Dicho paciente había sido sometido un año antes a **una BIOPSIA transrectal de próstata ecodirigida por sextantes** que fue *negativa*. Se decide, ante la elevación del PSA, realizar una E-RME previa a la 2ª biopsia transrectal, en la que se objetiva una lesión hipointensa que abarca zona central i periférica del ápex del lóbulo D prostático. El estudio espectroscópico de ésta lesión mostró una curva de colina discretamente más elevada que la curva de citrato, con **un ÍNDICE de Ch-Cr/Ci >** *0,80*, que sugería la presencia de lesión neoplásica, por lo que se biopsia dicha zona por ecografía transrectal. La AP de la biopsia confirmó la presencia de un ADK próstata Gleason 6.

Figure 1: A sample clinical case.

Two different tracks have been defined in TESTLINK, one for each of the languages in the task. The tracks are complementary and can be addressed individually or conjunctly.

## 3 Dataset

The TESTLINK task is based on the Spanish and Basque parts of E3C, the multilingual European Clinical Case Corpus (Magnini et al., 2022), which consists of three sections of clinical cases published in medical journals and other medical resources. One of these sections has been manually annotated with events (which include laboratory tests, among others), temporal expressions and temporal relations according to THYME (Styler et al., 2014), an adaptation of the TimeML framework (Pustejovsky et al., 2003), and also with results of laboratory test and measurements (both marked

```
Paciente de 65 a.  de edad, que presentaba una elevación progresiva de las
cifras de PSA desde 6 ng/ml a 12 ng/ml en el último año.  Dicho paciente había
sido sometido un año antes a una BIOPSIA transrectal de próstata ecodirigida
por sextantes que fue negativa.  Se decide, ante la elevación del PSA, realizar
una E-RME previa a la 2ª biopsia transrectal, en la que se objetiva una
lesión hipointensa que abarca zona central i periférica del ápex del lóbulo
D prostático.  El estudio espectroscópico de ésta lesión mostró una curva de
colina discretamente más elevada que la curva de citrato, con un índice de
Ch-Cr/Ci > 0,80, que sugería la presencia de lesión neoplásica, por lo que se
biopsia dicha zona por ecografía transrectal.  La AP de la biopsia confirmó la
presencia de un ADK próstata Gleason 6.

100001      REL       94-101      6 ng/ml      84-87      PSA
100001      REL      104-112      12 ng/ml     84-87      PSA
100001      REL      251-259      negativa     185-192    biopsia
100001      REL      619-623      0,80         598-604    índice
```

Figure 2: Annotated clinical case sample.

| | Training set | | Test set | |
|---|---|---|---|---|
| | Tokens | Rel. | Tokens | Rel. |
| Spanish | 28,815 | 597 | 29,668 | 668 |
| Basque | 34,052 | 1,291 | 12,756 | 345 |

Table 1: Dataset statistics.

through the RML tag, specifically created within the E3C project).

More specifically, the TESTLINK task is based on the following sets of data (as reported in Table 1):

1. Training and development data: 81 and 90 clinical cases, for Spanish and Basque respectively (for a total of 28,815 and 34,052 tokens) that had been manually annotated during the E3C project and then specifically revised for the task by expert computational linguists (inter-annotator agreement has been assessed so as to ensure consistent annotations throughout the whole dataset); the total number of annotated pertains-to relations amounts to 597 and 1,291 for Spanish and Basque, respectively;

2. Test data: for each language, 80 clinical cases taken from E3C (for a total of 30,820 tokens for Spanish and 13,759 for Basque) with manual annotations specifically performed for the task (in total 668 and 345 relations have been annotated respectively).

### 3.1 Annotation

Among all the annotations foreseen by the E3C project, the data used for TESTLINK contain the following annotations:

- Laboratory tests and measurements; these belong to the TimeML category EVENT and are therefore marked by their syntactic head only (one token).
- Lab test, analytics and measurement results; these belong to the RML category, as defined within the E3C project, and are marked by a whole syntactic chunk (one or more tokens).
- Pertains-to relations connecting an RML (the source) to the relevant EVENT (the target); Pertains-to relations can be one-to-one, one-to-many and many-to-one.

The annotated data have been provided to the participants in a format that is an adaptation of the PubTator format (see an example in Figure 2). It consists of a straightforward tab-delimited text file, where every document in the dataset is in a new line preceded by the DOCID and the `|t|` marker. A space line is used as an indicator of the end of the document, followed by the annotated relations: every relation is in a separate line and is represented as an ordered pair, as in (RML `->` EVENT), and each string is represented by its start and end character offsets.

## *4 Baselines*

To improve the assessment of participant systems' performance, supervised and unsuper-

vised baselines have been used for comparative analysis. These baselines have been made available through the GitLab repository.[1]

Two different supervised baselines have been defined.

The first approach is based on vocabulary-transfer from training to testing (voc. tran.). In this approach, a system is used to recognize textual references to laboratory tests/measurements present in the test set by leveraging the laboratory tests/measurements identified in the training set. Additionally, regular expressions derived from the training data are used to identify a wide range of textual references to test results, commonly represented by values. Subsequently, a relationship is established between each pair of laboratory test/measurement and test result mentioned within the same sentence.

The second approach is based on multilingual BERT model[2] fine-tuned on textual mentions involved in relations within the training data. The implementation of this model has been carried out using the SimpleTransformer library.[3] The model recognizes textual references to laboratory tests/measurements and test results.

*Pérdida de 12 Kg (peso actual 42 Kg) / 12 Kg weight loss (current weight 42 KG)*

In the example above, which includes two relations (12 Kg -> Pérdida and 42 Kg -> peso), the model identifies the following entity mentions using the IOB annotation. In this annotation, TST is used to label laboratory tests/measurements, while RML is employed to indicate test results:

*Pérdida [B-TST] de 12 [B-RML] Kg [I-RML] (peso [B-TST] actual 42 [B-RML] Kg [I-RML]*

Following that, another multilingual BERT model (configured similarly to the previous BERT model) was fine-tuned on the annotated relations within the training data to extract the relationships between the recognized laboratory tests/measurements and their results in the test data. Concerning the training data, both positive and negative examples were generated for sentences containing at least one laboratory test/measurement and one test result. For each generated example, the entities in the relationship were marked by adding "[TST]" as both the prefix and suffix to the laboratory tests/measurements, while "[RML]" was used to mark the test results. The number of examples produced for each sentence is calculated by multiplying the number of laboratory tests/measurements by the number of test results present in the sentence. As for the test data, the examples to be classified were generated similarly to the training data, but instead of using the entities in the gold standard, the predicted entities were used. In the case of the example provided above, the following examples are generated along with the corresponding model prediction (1=positive, 0=negative):

```
1 [TST]Pérdida[TST] de [RML]12 Kg[RML]
(peso actual 42 Kg)

0 [TST]Pérdida[TST] de 12 Kg (peso
actual [RML]42 Kg[RML])

0 Pérdida de [RML]12 Kg[RML]
([TST]peso[TST] actual 42 Kg)

1 Pérdida de 12 Kg ([TST]peso[TST]
actual [RML]42 Kg[RML])
```

The unsupervised baseline method uses GPT and OpenAI's API (text-davinci-003 model, with a temperature setting of 0 and default values for other parameters) for its implementation. By using a single annotated example instead of the complete set, this baseline approach is aligned more closely with unsupervised learning rather than supervised learning. For Spanish and Basque, the baseline evaluation was performed using the subsequent prompts.

Spanish: *Tengo una tarea que consiste en extraer menciones de pruebas de laboratorio y sus resultados de declaraciones clínicas. Aquí hay un ejemplo de texto y salida. Texto: {DOCID} Nota: El resultado se escribe primero y luego el nombre de la*

[1] https://gitlab.fbk.eu/zanoli/clinkart-baseline.git

[2] model=bert-base-multilingual-cased, epochs=5, learning_rate=4e-5, batch_size=16

[3] https://simpletransformers.ai

*prueba en la salida. Están separados por "|". Ahora dame la salida para el siguiente texto: {TEXT} Imprime solo la salida y si no la hay, no imprimas nada.*[4]

Basque: *Laborategiko proben aipamenak eta haiei dagozkien emaitzak adierazpen klinikoetatik ateratzeko zeregina daukat. Hona hemen testuaren eta irteeraren adibide bat. Testua: {DOCID} Orain eman iezadazu testu honen irteera: {TEXT} Inprimatu bakarrik irteera existitzen bada eta kito.*[4]

Within the prompts, {DOCID} represents the annotated document (docId:100320 and docId:100009 for Spanish and Basque, respectively) selected from the training dataset as the only example for GPT. {Text} represents the document to be annotated.

## 5 System Descriptions

Three teams submitted their annotated data, so we evaluated a total of four runs for the Spanish track (submitted by the three teams) and two runs for the Basque track (submitted by two of the teams).

**HiTZ Center**: a conventional two-step approach was used to perform relation extraction on the Spanish dataset. This approach is in line with the mBERT baseline presented in Section 4. This involves automated entity recognition followed by the extraction of relations between the recognized entities. Several models were tested, including BERT and its derivative models, which were pre-trained on both general and biomedical domains. Annotations from two of these models were submitted for evaluation. The first model, BETO (Cañete et al., 2020), is a BERT model pre-trained on a large Spanish corpus within the general domain. The second model, BSC-bio-ehr-es-Pharmaconer, is a fine-tuned version of the bsc-bio-ehr-es (Carrino et al., 2022) for the task of entity recognition in the biomedical domain. BETO and BSC-bio-ehr-es-Pharmaconer were employed for both entity recognition and relation extraction tasks.

**LinkMed**: the presented system is similar to the one presented by HiTZ. It consists of two consecutive components that first perform entity extraction and then entity linking. The NER model obtains mentions of events and their corresponding results. In the second module, the classification model takes each event-result pair of mentions and predicts whether there is a relation between them. For this, a domain-specific language model is fine-tuned to create contextual representations of the spans found in the NER module, which are fed into a linear layer. In order to include a larger token window to increase the textual context, FLERT (Schweter and Akbik, 2020) is used, which allows to fine-tune transformer-based models —in this case, biomedical and clinical versions of RoBERTa (Carrino et al., 2022) and the Spanish version of BERT, BETO (Cañete et al., 2020)— at document level.

**Simple Ideas**: The proposed approach uses a pipeline of two entity recognition modules in cascade. The first one identifies and marks the test events and the second one, which is trained on examples containing event markers, finds the result/measurement pertaining to each event. In both steps, sentences are fed one at a time. For both steps, a RoBERTa model pre-trained on biomedical and clinical texts (Carrino et al., 2021) is used in the case of Spanish, while BERTeus (Agerri et al., 2020) is employed for Basque. Data augmentation is exploited for training: synthetic sentences are generated by replacing words in the original sentences with other semantically similar words and by slightly varying numeric values. The availability of the implemented code contributes to the reproducibility of the presented results.

## 6 Results

The evaluation of the results obtained by the different systems was conducted using the standard Precision (Pr), Recall (Re), and $F_1$ measures. In this evaluation, a relation prediction is considered correct if the start and end character offsets of the source and target entities, as well as their order within the relation, are all accurately predicted. The BioCreative V CDR task[5] scorer was used for carrying out this evaluation. The results of the systems for both Spanish and Basque are

[4] I have a task to extract mentions of laboratory tests and their results from clinical statements. Here is an example of text and output. Text: {DOCID} Note: The result is written first and then the test name in the output. They are separated by "|". Now give me the output for the following text: {TEXT} Only print the output if there is any and nothing else.

[5] https://biocreative.bioinformatics.udel.edu/tasks/biocreative-v/track-3-cdr/

presented in Table 2, while the results of the baselines are in Table 3.

| Teams | Pr | Re | F1 |
|---|---|---|---|
| Spanish | | | |
| Simple Ideas | 71.45 | 65.57 | 68.38 |
| LinkMed[6] | 46.99 | 43.26 | 45.05 |
| HiTZ_BETO | 43.27 | 11.08 | 17.64 |
| HiTZ_Pharmaconer | 34.84 | 11.53 | 17.32 |
| Basque | | | |
| Simple Ideas | 72.54 | 72.75 | 72.65 |
| LinkMed[7] | 15.71 | 15.94 | 15.83 |

Table 2: Precision, Recall and $F_1$ measure of the participating systems.

| Baseline | Type | Pr | Re | F1 |
|---|---|---|---|---|
| Spanish | | | | |
| mBERT | S | 61.13 | 60.03 | 60.57 |
| GPT | U | 25.24 | 38.29 | 30.43 |
| voc. tran. | S | 17.41 | 30.24 | 22.10 |
| Basque | | | | |
| mBERT | S | 81.37 | 75.94 | 78.56 |
| GPT | U | 21.96 | 21.10 | 21.52 |
| voc. tran. | S | 17.97 | 35.94 | 23.96 |

Table 3: Precision, Recall and $F_1$ measure of the supervised (S) and unsupervised (U) baselines calculated on the Spanish and Basque datasets.

For the sake of comparison, we report that the $F_1$ measure obtained by the best system in the parallel task for Italian, CLinkaRT at EVALITA 2023[8] (Altuna et al., 2023), is 62.99.

## 7 Discussion

Both traditional NER and RE pipelines and more innovative approaches have been presented to the task.

HiTZ and LinkMed opted for the more traditional two-step approach of first solving NER and then performing RE among the annotated entities.

As regards HiTZ, it is interesting to note that the BSC-bio-ehr-es-Pharmaconer model pre-trained on the biomedical domain ($F_1$ 17.32) performed in line with the BETO model pre-trained on the more general domain ($F_1$ 17.64). However, considering that the task data belongs to the medical domain, it was expected that BSC-bio-ehr-es-Pharmacone would outperform BETO. The team's reduced performance in named entity recognition on the actual test set, as opposed to their evaluation test set, explains why they did not achieve results consistent with mBERT, which follows the same approach.

In the case of LinkMed, they also made use of a model pre-trained in the biomedical domain, still not achieving the results obtained by the baseline mBERT model. They performed a result analysis that showed that relations involving single-token entities were easier to obtain. In fact, considering only relations that had a one-token RML, the model achieved a $F_1$ of 51, compared to the overall result ($F_1$ 45.05). They also compared the performance of relations involving two-token and multiple-token RMLs and discovered that the performance plummeted in the two-token scenario ($F_1$ 36.0), but was coherent with the rest of the results in the multiple-token one ($F_1$ 45.0). They also measured the performance of the model by considering the distance between the entities in each relation. Precision decreased as the distance between the entities augmented, while recall increased. The highest F score ($F_1$ 53.0) was achieved between entities that were separated by between 9 and 25 characters.

Simple Ideas, on the other hand, implemented a two-step-pipeline procedure which first identifies test events and then their results. The system performed better than the more traditional approaches, but not better than the mBERT baseline for Basque. This approach obtained the best results in both tracks and it also obtained the highest ranking for the Italian language, in the parallel CLinkaRT task.

In regard to system architecture, all participants have opted for BERT-styled models for both NER and RE. For Spanish models that have been fine-tuned on biomedical and clinical data such as BSC-bio-ehr-es-Pharmaconer and a domain fine-tuned version of RoBERTa are available, while for Basque, only general domain BERT models, such as BERTeus or mBERT, are ready to use.

As expected, models that have seen biomedical and clinical data during their training processes perform better. Still, HiTZ's Phar-

[6]These results are the ones of a run submitted after the deadline.

[7]No description for this run has been provided by the participants.

[8]`https://www.evalita.it`

maconer approach seems not to perform so well and the general domain BERTeus obtains good results for Basque and the general-domain mBERT used for the baselines also ranks high in our task.

The results obtained did not allow us to determine whether the task being examined is inherently more difficult in one language compared to other languages due to language-specific traits. Within this framework, the vocabulary transfer baseline, which was expected to provide a preliminary indication of the task's difficulty, achieved better results on the Italian CLinkaRT task ($F_1$ 30.88) compared to the results for Basque ($F_1$ 23.96) and Spanish ($F_1$ 22.10) in TESTLINK. However, participant systems, such as Simple Idea, showed contrasting results. As this team hypothesizes, better results in the Basque track might be due to the larger amount of training examples in that language. The size of the dataset may also explain the better performance of the Simple Ideas team as they make use of augmented data in the model training step.

## 8 Conclusions

Extracting laboratory tests and measurements and their results from clinical narratives seems to be a challenging task in clinical information extraction. The great variety of tests and the fact that most results contain numerical values differentiate this task from most entity recognition and linking tasks.

Participant systems have achieved good results but there is still room for improvement. Performing NER in the first place and then performing RE is prone to error propagation and this affects the final results. In addition, the effect of the volume of training data should also be further studied.

As this was the first time that we were proposing this task, in parallel to the twin CLinkaRT task at EVALITA, we decided to keep it strictly focused on relations between tests and their results, but in the future, it might be interesting to integrate this task in a more complex information extraction effort that considers a wider range of clinical entities and relations.

## Acknowledgements

This work has been partially funded by the Basque Government postdoctoral grant POS 2022 2 0024.

## References

Agerri, R., I. San Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, and E. Agirre. 2020. Give your text representation models some love: the case for Basque. In *Proceedings of the Twelfth Language Resources and Evaluation Conference*, pages 4781–4788, Marseille, France, May. European Language Resources Association.

Alfattni, G., N. Peek, and G. Nenadic. 2020. Extraction of temporal relations from clinical free text: A systematic review of current approaches. *Journal of Biomedical Informatics*, 108:103488.

Altuna, B., G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, and R. Zanoli. 2023. CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain. In *Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023)*, Parma, Italy, September. CEUR.org.

Carrino, C. P., J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, and M. Villegas. 2021. Biomedical and Clinical Language Models for Spanish: On the Benefits of Domain-Specific Pretraining in a Mid-Resource Scenario.

Carrino, C. P., J. Llop, M. Pàmies, A. Gutiérrez-Fandiño, J. Armengol-Estapé, J. Silveira-Ocampo, A. Valencia, A. Gonzalez-Agirre, and M. Villegas. 2022. Pretrained Biomedical Language Models for Clinical NLP in Spanish. In *Proceedings of the 21st Workshop on Biomedical Language Processing*, pages 193–199, Dublin, Ireland, May. Association for Computational Linguistics.

Cañete, J., G. Chaperon, R. Fuentes, J.-H. Ho, H. Kang, and J. Pérez. 2020. Spanish Pre-Trained BERT Model and Evaluation Data. In *PML4DC at ICLR 2020*.

Hao, T., H. Liu, and C. Weng. 2016. Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text. *Methods of information in medicine*, 55:266–75.

Jain, K. and V. Prajapati. 2021. NLP/Deep Learning Techniques in Healthcare for Decision Making. *Primary Health Care*, 11.

Jiménez-Zafra, S. M., F. Rangel, and M. Montes-y Gómez. 2023. Overview of IberLEF 2023: Natural Language Processing Challenges for Spanish and other Iberian Languages. In *Proceedings of the Iberian Languages Evaluation Forum (IberLEF 2023), co-located with the 39th Conference of the Spanish Society for Natural Language Processing (SEPLN 2023), CEUR-WS.org*.

Johnson, A. E., T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, and R. G. Mark. 2016. MIMIC-III, a freely accessible critical care database. *Scientific Data*, 3.

Magnini, B., B. Altuna, A. Lavelli, A.-L. Minard, M. Speranza, and R. Zanoli. 2022. European Clinical Case Corpus. In G. Rehm, editor, *European Language Grid*. Springer, Cham, Switzerland, 1 edition, November, chapter 17, pages 283–288.

Newman-Griffis, D., G. Divita, B. Desmet, A. Zirikly, C. P. Rosé, and E. Fosler-Lussier. 2020. Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets. *Journal of the American Medical Informatics Association*, 28(3):516–532, 12.

Percha, B. 2021. Modern Clinical Text Mining: A Guide and Review. *Annual Review of Biomedical Data Science*, 4(1):165–187. PMID: 34465177.

Pustejovsky, J., J. M. Castaño, R. Ingria, R. Saurí, R. J. Gaizauskas, A. Setzer, G. Katz, and D. R. Radev. 2003. TimeML: Robust Specification of Event and Temporal Expressions in Text. *New directions in question answering*, 3:28–34.

Sankoh, O. and P. Byass. 2014. Cause-specific mortality at INDEPTH Health and Demographic Surveillance System Sites in Africa and Asia: concluding synthesis. *Global health action*, 7.

Santiso, S., A. Pérez, and A. Casillas. 2021. Adverse Drug Reaction extraction: Tolerance to entity recognition errors and sub-domain variants. *Computer Methods and Programs in Biomedicine*, 199:105891.

Schweter, S. and A. Akbik. 2020. FLERT: Document-Level Features for Named Entity Recognition. *CoRR*, abs/2011.06993.

Styler, W. F., S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, et al. 2014. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

Trigueros, O., A. Blanco, N. Lebeña, A. Casillas, and A. Pérez. 2022. Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention. *International Journal of Medical Informatics*, 157:104615.