

Simple Ideas@TESTLINK: Relying On FiNER Models

Marius Micluța-Câmpeanu^{1,*}, Liviu Petrișor Dinu^{1,2}

¹Faculty of Mathematics and Computer Science, University of Bucharest, Romania

²Human Language Technologies Research Center, University of Bucharest, Romania

Abstract

In this paper, we introduce an intuitive solution to extract relations using solely Named Entity Recognition (NER) models. Our approach achieves the best scores in the TESTLINK task among all other participants by a large margin, with 68.38% F1-score for Spanish and 72.65% F1-score for Basque.

General purpose relation extraction methods typically require a classifier to identify the relationship type or leverage generative models. When we are only interested in extracting a single relation type, such approaches might be overly complex or costly. We show that simple methods are capable of performing relation extraction by using NER models, data augmentation and basic text processing, without an additional relation classifier.

Our proposed solution consists of two NER models that predict each side of a relation at the sentence level, followed by a post-processing step to construct the final relations and to perform error corrections.

Keywords

named entity recognition, relation extraction, transformers, data augmentation, IberLEF, TESTLINK

1. Introduction

Medical and clinical texts have been widely studied for information extraction purposes on English corpora [1], while other languages received little to no attention regarding dedicated medical corpora and models. In recent years, there has been an effort to develop these kinds of resources for languages other than English [2], but not much research has been preoccupied with lower resource languages.

1.1. Task description

The shared TESTLINK task [3] at IberLEF 2023 [4] is concerned with correlating laboratory measurements and tests with textual mentions in clinical narratives. These relations can be later used as building blocks for other downstream data mining tasks, for example to build knowledge graphs [5], document treatment methods or uncover interactions between chemicals, drugs, proteins and diseases [6].


The TESTLINK task is based on a subset of the E3C Corpus [7], a collection of clinical cases in multiple European languages. The annotated subset for this task contains Spanish and Basque texts stored in a variant of the PubTator format to represent relationship information, along

IberLEF 2023, September 2023, Jaén, Spain

*Corresponding author.

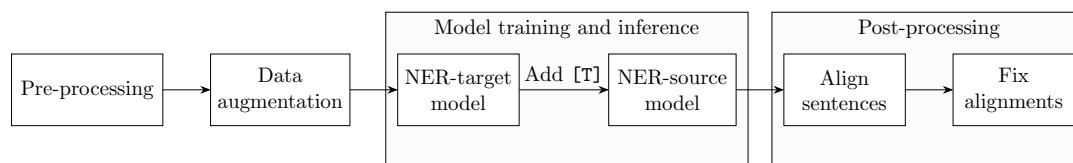
✉ marius.micluta-campeanu@unibuc.ro (M. Micluța-Câmpeanu); liviu.p.dinu@gmail.com (L. P. Dinu)

ORCID [0000-0002-7559-6756](https://orcid.org/0000-0002-7559-6756) (L. P. Dinu)

 © 2023 Copyright for this paper by its authors. Use permitted under Creative Commons License Attribution 4.0 International (CC BY 4.0).

 CEUR Workshop Proceedings (CEUR-WS.org)

Figure 1: Overview of our system architecture. Data augmentation is performed for both models. Each model is trained on the original dataset concatenated with the augmented set.



with tokenized versions for each document. Relations are pairs of entity mentions between sources and targets, where a source is a RML entity (tag for test results and measurements) and a target is an event describing relevant aspects to the clinical history of patients, such as symptoms. Following THYME annotations, all targets are comprised of a single token, while sources consist of one or more tokens.

In this paper, we present our team’s contribution to the TESTLINK task, introducing a straightforward methodology for extracting basic relations. First, we transform the input into an adequate internal representation and apply data augmentation. Then, we train one NER model to predict targets, followed by training a second NER model to predict sources with the aid of a special token for targets. Finally, we determine the final relations and apply minor corrections in a post-processing step. This pipeline is summarized in Figure 1.

1.2. Related work

The idea of using special tokens in discriminative transformer models to aid learning certain patterns dates to at least the original BERT paper [8]. SciBERT [1] appears to use a similar approach using entity markers. They predict the whole relation in one step, using start and end markers, while in our system we do not insert any end marker. BioGPT [9] formulates the relation extraction problem as a text generation task and experiments with different output label strategies. Like many generative models, BioGPT requires more compute resources, compared to our system which uses two lightweight token classifiers.

2. System description

We use two NER models to predict sources and targets at the sentence level, then link the predictions via post-processing. The first NER model called NER-target predicts all targets in a sentence. Sentences are determined using the provided tokenizations. For each target identified by the first model, we create a new example containing a single labeled target using a special marker token [T].

Example: albúmina sérica de 2,5 g/dl y proteinuria de 6 g/24 h.

NER-target predicts the targets albúmina and proteinuria.

We generate two examples for our next NER model:

- [T] albúmina sérica de 2,5 g/dl y proteinuria de 6 g/24 h
- albúmina sérica de 2,5 g/dl y [T] proteinuria de 6 g/24 h

The second model named `NER-source` is trained on examples containing target markers, with labels only for source tokens. Since the added marker is a simple pattern from the model’s perspective, the model is capable of inferring whether a sentence contains relations based on the presence of a target marker.

After sources are predicted for each sentence, we align the inferred sources and targets with the initial raw text in order to output appropriate text spans. Predictions are ordered by the start mention of sources since we train and evaluate at the sentence level.

This simple approach turns out to be very effective for determining relations of a single type. In the following subsections, we provide additional details regarding our system.

2.1. Pre-processing

We parse the provided annotations into an internal representation, then we serialize the results as JSON for easier dataset loading by the training script. Given that most examples have a lot more tokens than the limit of transformer models, we consider each sentence as an example. To ease post-processing, we serialize the start offset of every sentence for all examples, along with the example id and the sentence number. For the `NER-source` model, we also store the target span if that example contains a relation.

First, we create IOB2 tags (inside, outside, beginning) for target entities in order to train a NER model to predict targets from raw text. We choose to learn predicting only the targets at this stage because they consist of a single labeled token. While typical relation extraction methods use just one NER model to identify both sources and targets as an initial step, our approach does not require labeling sources because they would be discarded anyway in a later stage. Moreover, having a lower number of possible labels helps the model achieve better results.

Since targets are single tokens in the training set, `NER-target` only needs to predict two labels: “B-T” (begin target) and “O” (other). An example for `NER-target` has the following labeling:

(1) [O]Resto [O]de [B-T]parámetros [O]dentro [O]de [O]la [O]normalidad [O].

The relation extraction part of the system is divided into two pipelines: one for training and one for inference. This model receives target spans for each sentence from the train set or from the previous model. When training, source spans are also available. All relations are represented as source-target tuples of spans (ranges). We convert relations to IOB2 tags containing both sources and targets. Before each target, we insert a special marker token [T]. Example 1 is relabeled for the `NER-source` model in a similar fashion:

(2) [O]Resto [O]de [O][T] [O]parámetros [B-S]dentro [I-S]de [I-S]la [I-S]normalidad [O].

In order to distinguish between multiple relations in the same sentence, source and target tokens are labeled with their respective targets, ignoring tags from other relations. If multiple sources reference one target, modeling a many-to-one relation, we do not want to add the same target several times to our set of tags, because it would be redundant. In this case, we only add source tokens. However, if we have a one-to-many relation between sources and targets, the same source token will be added once for each corresponding target (see example 3).

This strategy allows us to unambiguously determine all separate relationships. It can be seen as a sliding window approach, asking the `NER-source` model to focus only on one target at a

time. Moreover, creating additional examples for every target in sentences with multiple targets acts as a form of data augmentation.

- (3) The example “estudios negativos para bacilo” has a one-to-many relation with the source “negativos” and the targets “estudios” and “bacilo”.

The NER-source model will receive two samples:

- a. [O][T] [O]estudios [B-S]negativos [O]para [O]bacilo
- b. [O]estudios [B-S]negativos [O]para [O][T] [O]bacilo

For the second NER model, we keep IOB2 tags only for source tokens and convert target IOB2 tags to O (outside) tags. We choose this strategy because we already have the target embedded in every example using the special marker token [T]. Similar to the previous NER model, having fewer labels increases model performance.

2.2. Data augmentation

The E3C subset corpus annotated for this task contains 597 relations for Spanish and 1291 relations for Basque, excluding examples without relationships. The small number of labeled sentences appears to be a limiting factor for our system. With the help of n1paug library [10], we augment both positive and negative examples.

We perform data augmentation using contextual word embeddings by replacing random words and aligning the result with the original sentence in order to assign proper labels. While it is possible to perform word insertion or deletion in this scenario, label assignment would be more problematic due to larger offsets. The contextual word embeddings for augmentation are provided by pretrained models used for training.

Sources, targets and the target marker are copied from the initial sentence without changes because we cannot afford noisy labels in this task. For numeric values that are not part of a relation, we perform small perturbations of ± 2 for decimal numbers (usually age or large quantities) and ± 0.1 for real numbers (typically measurements). These numerical perturbations are applied together with the other changes. We chose not to create separate examples for word substitutions and numeric adjustments because we risk to add too many similar samples which would lead to overfitting.

In theory, these alterations could also be applied on labeled tokens since they should not change the relation or sentence meaning. They were left out in the final implementation due to corner cases involving intervals and sign changes. We intend to explore this area in future work because we believe it might help models generalize better in extracting this kind of information.

The special marker token is not included in this process and is added back afterwards. If a sentence contains less than 10 words, it is skipped. All augmented examples with a different word count than the original sentence are dropped since we cannot reliably apply any previous labels.

We experiment with values between 3 and 6 for the minimum word replacements. We choose 6 substitutions for Spanish and 4 for Basque to minimize the number of discarded samples. We repeat the augmentation process with different multipliers for positive examples containing relations (`in_multiplier`) and negative examples without any labels (`out_multiplier`). Based

Table 1

Data augmentation multipliers for each NER model.
Every sentence is augmented a number of times according to the multiplier.

Multiplier type	NER-target	NER-source
in_multiplier	4	2
out_multiplier	1	1

on our experiments, we pick the values in Table 1 for these multipliers. We need to augment with fewer source tokens used in relation extraction due to the fact that there are already duplicate examples for sentences with several targets.

2.3. Entity recognition model

Our first named entity recognition model extracts target tokens from raw text. We use the popular HuggingFace Transformers library [11] to train a token classification model. For the Spanish subtask, we use a RoBERTa model pretrained on biomedical and clinical texts [2], available as `PLanTL-GOB-ES/roberta-base-biomedical-clinical-es` on HuggingFace Hub. For the Basque subtask, there are very few models pretrained specifically for Basque if we exclude the multilingual ones. We use `ixa-ehu/berteus-base-cased` [12], a BERT model trained on news articles and Basque Wikipedia.

All models are trained for 4 epochs with default parameters: AdamW optimizer with a learning rate of $5e^{-5}$ using linear decay and no warmup, weight decay of $1e^{-2}$, batch size of 8 samples and 10% examples reserved for validation.

The inference step demands more post-processing work since transformer-based models provided by HuggingFace employ a different tokenization strategy than the one available in the TESTLINK dataset. We use spaCy [13] to align the examples prepared by HuggingFace Datasets with the initial texts. For each predicted target, we create one relation with a dummy source in order to apply the logic from our preprocessing step. The alignments are needed to output proper offsets for the target token.

2.4. Relation extraction model

The previous model is tasked to determine targets. Given a specific target in a sentence, the relation extraction model learns to predict the relevant source tokens. We process the input in a manner that allows this step to also become a named entity recognition problem instead of a relation classification problem.

When training, we take advantage of gold labels for both sources and targets, with the latter being converted to the special marker [T] as described earlier in section 2.1. Due to the similarities between the two models, we apply the same training procedure presented in the previous section with the same pretrained models.

At inference, this model expects sentences to have an embedded target marker [T] for examples that contain a relation, with the possibility to supply this marker through either gold labels or inference of the first model. Like before, we align the processed text required by this

Table 2

Results on Spanish for extracting relations between laboratory tests and their results from clinical texts.

System	Precision	Recall	F1
vocabulary transfer baseline	17.41	30.24	22.10
mBERT baseline	61.13	60.03	60.57
Ours	71.45	65.57	68.38
Ours (dev set)	83.52	81.59	82.42

model with the original text. If the predicted source appears after its corresponding target, we also have to account for the additional [T] marker and subtract its length from the predicted text span. Finally, we perform minor corrections for off-by-one errors due to alignment mismatches and then sort relations by the start mention of the first source token in each relation.

3. Results

The results of our system along with provided baselines are shown in Table 2 and Table 3. In the Spanish subtask, we obtain an F1-score of 68.38, with a 13% improvement over the mBERT baseline with a score of 60.57. For the Basque subtask, our method achieves an F1-score of 72.65, while the mBERT baseline has a score of 78.56. Our results are significantly better than the other competing systems. For completeness, we also include the results obtained locally on a development set extracted from the provided training set.

Our Spanish submission has an underlying model pretrained on medical and clinical texts, while the Basque submission is pretrained on generic texts such as news articles. One explanation for the better results obtained on Basque is that we have more than twice as many annotated relations for the downstream task, which compensates the lack of a dedicated pretrained model.

We perform parameter selection for the best performing model for each language using 10-fold cross-validation. The models training employ a validation of their own, meaning that they are trained on 81% of the initial train set: 10% examples are held out for cross-validation and 10% of the remaining 90% are reserved for validation during training.

Throughout our experiments, we noticed a significant gain in performance after adding more train samples. Given the small number of examples in the training set, even after data augmentation, we try to use as many samples as possible. Therefore, we train the final system on the entire train set, leaving only 10% for validation used in training. This approach allows us to fully utilize the available examples, although it might also lead to some overfit since we do not perform more splits.

Regarding the types of errors in our system, we notice other situations besides obvious false positives (usually acronyms, quantities or numbers) and false negatives. The training data contains a few ambiguous instances where consecutive tokens form either one relationship or more. Our model seems to prefer predicting more relations in these scenarios. Example of a relation split:

- (4) There is one true relation: 29,5 cm x 27,5 cm x 16 cm masa

Table 3

Results on Basque for extracting relations between laboratory tests and their results from clinical texts.

System	Precision	Recall	F1
vocabulary transfer baseline	17.97	35.94	23.96
mBERT baseline	81.37	75.94	78.56
Ours	72.54	72.75	72.65
Ours (dev set)	83.04	82.21	82.54

The system predicts two relations: 29,5 cm x 27,5 cm masa and 16 cm masa

The reverse situation is also present, as seen in the following example:

- (5) There are two true relations: > 2 adenopatias and < 5cm adenopatias
 There is only one predicted relation: > 2 y <5 cm adenopatias

Sometimes, the text span stops in the middle of a word, which can be fixed relatively easy in post-processing by leveraging the existing tokenization. Examples of incomplete predictions are shown below:

- (6) True relation: 4340 g/l IgM, predicted relation: 4340 g IgM
 True target: alfabetoproteína, predicted target: alfafe

4. Conclusion and future work

In this paper, we presented an alternative method for relation extraction relying solely on NER models. We have successfully applied this method in the TESTLINK shared task at IberLEF 2023 for linking laboratory results to their respective tests and measurements.

Our team achieved the first place across all competing teams in both languages. On the Spanish dataset, we surpass the mBERT baseline score by 8 percentage points regarding F1-score, while we come close to the mBERT baseline on Basque with less than 6 percentage points below in F1-score, proving the effectiveness of our approach. One of the main differences between our solution and the provided baseline is that we used a monolingual model. We have further applied our method in the CLinkaRT twin task [14] where we also reached the first place and performed a few additional experiments [15].

We provide the source code of our system to promote the design and use of simpler, lightweight models (available at <https://gitlab.com/marius.micluta-campeanu/testlink-clinkart-2023>). For future work, we intend to systematically analyze the system errors more thoroughly and explore this method in a multilingual environment or with the help of machine translation. Another unexplored path is finding the minimum amount of context necessary for better results, since our model relied on sentence-level predictions. Finally, we are interested in further development of medical applications based on extracted relations.

References

- [1] I. Beltagy, K. Lo, A. Cohan, 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), Association for Computational Linguistics, Hong Kong, China, 2019, pp. 3615–3620. URL: <https://aclanthology.org/D19-1371>. doi:10.18653/v1/D19-1371.
- [2] C. P. Carrino, J. Armengol-Estapé, A. Gutiérrez-Fandiño, J. Llop-Palao, M. Pàmies, A. Gonzalez-Agirre, M. Villegas, Biomedical and clinical language models for Spanish: On the benefits of domain-specific pretraining in a mid-resource scenario, 2021. arXiv:2109.03570.
- [3] B. Altuna, R. Agerri, L. Salas-Espejo, J. J. Saiz, R. Zanoli, M. Speranza, B. Magnini, A. Lavelli, G. Karunakaran, Overview of TESTLINK at IberLEF 2023: Linking Results to Clinical Laboratory Tests and Measurements, *Procesamiento del Lenguaje Natural* 71 (2023).
- [4] S. M. Jiménez-Zafra, F. Rangel, M. Montes-y Gómez, 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, 2023.
- [5] D. N. Nicholson, C. S. Greene, Constructing knowledge graphs and their biomedical applications, *Computational and structural biotechnology journal* 18 (2020) 1414–1428.
- [6] L. Luo, P.-T. Lai, C.-H. Wei, C. N. Arighi, Z. Lu, BioRED: a rich biomedical relation extraction dataset, *Briefings in Bioinformatics* 23 (2022) bbac282.
- [7] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C project: Collection and annotation of a multilingual corpus of clinical cases, in: J. Monti, F. Dell’Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, volume 2769 of *CLiC-It*, CEUR-WS, Milan Italy, 2020, pp. 422–431.
- [8] J. Devlin, M.-W. Chang, K. Lee, K. Toutanova, 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, Volume 1 (Long and Short Papers), Association for Computational Linguistics, Minneapolis, Minnesota, 2019, pp. 4171–4186. URL: <https://aclanthology.org/N19-1423>. doi:10.18653/v1/N19-1423.
- [9] R. Luo, L. Sun, Y. Xia, T. Qin, S. Zhang, H. Poon, T.-Y. Liu, Biogpt: generative pre-trained transformer for biomedical text generation and mining, *Briefings in Bioinformatics* 23 (2022).
- [10] E. Ma, NLP Augmentation, <https://github.com/makcedward/nlpaug>, 2019.
- [11] T. Wolf, L. Debut, V. Sanh, J. Chaumond, C. Delangue, A. Moi, P. Cistac, T. Rault, R. Louf, M. Funtowicz, J. Davison, S. Shleifer, P. von Platen, C. Ma, Y. Jernite, J. Plu, C. Xu, T. L. Scao, S. Gugger, M. Drame, Q. Lhoest, A. M. Rush, Transformers: State-of-the-art natural language processing, in: Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations, Association for Computational Linguistics, Online, 2020, pp. 38–45. URL: <https://www.aclweb.org/anthology/2020.emnlp-demos.6>.
- [12] R. Agerri, I. San Vicente, J. A. Campos, A. Barrena, X. Saralegi, A. Soroa, E. Agirre, Give

your text representation models some love: the case for Basque, in: Proceedings of the Twelfth Language Resources and Evaluation Conference, European Language Resources Association, Marseille, France, 2020, pp. 4781–4788. URL: <https://aclanthology.org/2020.lrec-1.588>.

- [13] M. Honnibal, I. Montani, S. Van Landeghem, A. Boyd, spaCy: Industrial-strength Natural Language Processing in Python (2020). doi:10.5281/zenodo.7715077.
- [14] B. Altuna, G. Karunakaran, A. Lavelli, B. Magnini, M. Speranza, R. Zanoli, CLinkaRT task overview: building links between clinical tests and their results (2023).
- [15] M. Micluța-Câmpeanu, L. P. Dinu, Simple Ideas@CLinkaRT: LeaNER and MeaNER Relation Extraction (2023). (to appear).