CLinkaRT at EVALITA 2023: Overview of the Task on Linking a Lab Result to its Test Event in the Clinical Domain

Begoña Altuna[1,*], Goutham Karunakaran[2], Alberto Lavelli[3], Bernardo Magnini[3], Manuela Speranza[3] and Roberto Zanoli[3]

[1]HiTZ Center - Ixa, University of the Basque Country UPV/EHU, Manuel Lardizabal 1, 20018 Donostia, Spain
[2]Università di Trento, Trento, Italy
[3]Fondazione Bruno Kessler, Via Sommarive 18, 38123 Povo, Italy

Abstract

CLinkaRT at EVALITA 2023 is a relation extraction task based on clinical cases taken from the E3C corpus, i.e. Italian written documents reporting statements of a clinical practice. The task consists in identifying clinical results and measures and linking them to the laboratory tests and measurements from which they were obtained. Three teams participated in the task and various supervised machine learning models, both traditional and based on deep learning, were evaluated. In this evaluation, the deep learning models outperformed the traditional ones. Interestingly, none of the teams explored the use of few-shot language modeling. However, the fact that the supervised models significantly outperformed the task baselines implementing few-shot learning shows the crucial role still played by the availability of annotated training data.

Keywords

Relation Extraction, Clinical NLP, Named Entity Recogniton, Supervised Learning

1. Introduction and Motivation

There is a growing interest in processing clinical data for tasks of public interest, such as clinical decision making [1] or monitoring of the health status of a country [2]. 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 [3], 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 given priority to the identification of diseases [4] or events [5]. As far as the extraction of relations from clinical texts is concerned, previous work has focused on concept normalization [6] and temporal relations [7], among others. Laboratory tests and measurements and their results have been given little attention [8], although they provide interesting information on the patients' status at a certain time of the development of a disorder and are crucial to choose the right diagnosis. 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 task [9]. In this context, the CLinkaRT task (LINKing A Result to its Test in the CLINnical domain) in EVALITA 2023 [10] provides an opportunity to evaluate different Natural Language Processing approaches and does this with a focus on Italian, a less explored language than English.

2. Task Description

The CLinkaRT task consists in identifying textual mentions of both laboratory tests and measurements in a clinical narrative, and then linking these to their respective results. 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, the assessment of the patient's situation and the diagnosis, as well as the following treatments. Laboratory tests and measurements are commonly done as part of this process and are typically documented in clinical narratives.

Figure 1 presents an excerpt of a clinical case where laboratory tests have been marked in bold[1] and their results in italics.

[1]Note that the head of the mention is capitalized.

EVALITA 2023: 8th Evaluation Campaign of Natural Language Processing and Speech Tools for Italian, Sep 7 – 8, Parma, IT
*Corresponding author.
†These authors contributed equally.
begona.altuna@ehu.eus (B. Altuna); goutham.karunakaran@studenti.unitn.it (G. Karunakaran); lavelli@fbk.eu (A. Lavelli); magnini@fbk.eu (B. Magnini); manspera@fbk.eu (M. Speranza); zanoli@fbk.eu (R. Zanoli)
0000-0002-4027-2014 (B. Altuna)

Osvaldo, anni 52, ha una storia di diarrea e calo ponderale che si può far riferire a due anni prima. Non c'è storia di sanguinamento gastroenterico ed **una RICERCA di sangue occulto fecale** è risultata *negativa* su tre campioni. Ammette di averci dato dentro con l'alcol in passato, ma da diversi anni è assolutamente astinente. Ha un diabete, controllato con insulina. Sei anni prima è stato colecistectomizzato. **Gli ESAMI di laboratorio** sono *normali*, se si fa eccezione per una lieve anemia, così come *normali* sono **lo STUDIO radiologico del piccolo e del grosso intestino**.

Figure 1: A sample clinical case

In this example, we have the following Pertains-To relations that participants needed to identify between results and tests:

- negative / negativa −> the research for fecal blood / una RICERCA di sangue occulto fecale
- normal / normali −> the laboratory exams / gli ESAMI di laboratorio
- normal / normali −> the radiological study of the bowels / lo STUDIO radiologico del piccolo e del grosso intestino

3. Dataset

The CLinkaRT task is based on the Italian part of E3C, the multilingual European Clinical Case Corpus [11], a collection of clinical cases derived from different sources, such as published articles available from PubMed, and existing corpora. As such, the dataset encompasses a variety of clinical disciplines in different hospitals and a wide range of laboratory tests.

One of the three sections which make up the E3C corpus has been manually annotated with different types of information, such as:

- events (which include laboratory tests, among others), temporal expressions and temporal relations, annotated according to THYME [12], an adaptation of the TimeML framework [13];
- results of laboratory tests and measurements, marked through the RML tag (defined within the E3C project), and Pertains-To relations holding between an RML and the event it refers to;
- clinical entities (in particular diseases, syndromes, findings, signs, symptoms, etc.) listed in medical taxonomies, which is useful for tasks focusing on clinical entity recognition and analysis [14].

More specifically, the CLinkaRT task is based on two sets of data:

1. Training and development data: 83 clinical cases, for a total of 28,856 tokens and 658 Pertains-To relations. These documents correspond directly with the manually annotated section of the E3C corpus and have then been revised for the task;
2. Test data: 80 clinical cases taken from E3C and specifically annotated for the task, for a total of 26,437 tokens and 612 Pertains-To relations.

3.1. Annotation

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

- Laboratory test and measurement EVENTS: they include medical procedures in which parts of the body or bodily substances (blood, urine, etc.) are analyzed, as well as different acts of measuring, such as measuring patients' physical features (e.g. height and weight) or the size of a lesion or mass.
- RMLs are the results of lab tests and measurements; they can consist of a text string (e.g. normal / normali) but more often contain numerical values, typically followed by a unit of measure (e.g. 7,5 g/dl);
- 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.

In the example below we have two Pertains-To relations between two EVENTS, i.e. a laboratory test (*protidemia totale*) and a measurement (*peso*), and their results (RMLs), i.e. 4,5 g/dl and 19 Kg respectively.

Peso corporeo di 19 Kg, protidemia totale 4,5 g/dl / Body weight of 19 Kg, total protidemia 4,5 g/dl
19 Kg −> Peso
4,5 g/dl −> protidemia

Both RMLs and test and measurements EVENTS are marked as strings of text; notice, however, that tests and measurements belong to the TimeML category EVENT and are therefore marked by their syntactic head only (i.e. strictly one token only) while RMLs, as defined within the E3C project, are marked by a whole syntactic chunk (one or more tokens).

3.2. Inter Annotator Agreement

All the data used for the task have been (manually) annotated by expert computational linguists and inter-annotator agreement has been assessed on ten documents, which have been annotated by two annotators independently. On average, each annotator has identified 111 relations.

100001|t|Osvaldo, anni 52, ha una storia di diarrea e calo ponderale che si può far riferire a due anni prima. Non c'è storia di sanguinamento gastroenterico ed **una RICERCA di sangue occulto fecale** è risultata *negativa* su tre campioni. Ammette di averci dato dentro con l'alcol in passato, ma da diversi anni è assolutamente astinente. Ha un diabete, controllato con insulina. Sei anni prima è stato colecistectomizzato. **Gli ESAMI di laboratorio** sono *normali*, se si fa eccezione per una lieve anemia, così come *normali* sono **lo STUDIO radiologico del piccolo e del grosso intestino.**

100001	REL	201-209	negativa	156-163	ricerca
100001	REL	442-449	normali	416-421	esami
100001	REL	502-509	normali	518-524	studio

Figure 2: Sample of an annotated clinical case

The resulting Dice's coefficient [15] is 0.87, which is quite high given that agreement between annotators is only considered as such when there is a complete overlap in the spans of the source and the target (exact match). The high agreement between annotators ensures that annotations throughout the whole dataset are consistent. More specifically, the inter-annotator agreement is particularly high when numerical values are present in the RMLs (it reaches 0.92 in terms of Dice's coefficient), while it is slightly lower (Dice=0.84) in the case of RMLs without numerical values.

3.3. Data Distribution Format

The annotated data have been provided to the participants in a format that is in 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 unsupervised baselines have been used for comparative analysis. These baselines have been made available through the GitLab repository.[2]

The supervised baseline was assessed using two different approaches.

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 and measurements present in the test set using the entities found in the training set. Additionally, regular expressions derived from the training data are used to recognize various result entities that pertain to measurements, typically represented by values. To establish relationships between the recognized entities, a relation is created for each pair of laboratory test/measurement and result entities that co-occur together within the same sentence.

The second approach relies on a fine-tuned multilingual BERT model[3] trained on textual mentions involved in relations within the training data. The implementation of this model has been carried out using the SimpleTransformer library.[4] The model is capable of recognizing both textual references to laboratory tests and measurements and their results.

Peso corporeo di 19 Kg, protidemia totale 4,5 g/dl / Body weight of 19 Kg, total protidemia 4,5 g/dl
19 Kg −> Peso
4,5 g/dl −> protidemia

In the example above the implemented model identifies the following mentions, using the IOB annotation where test events are represented as TST and results as RML:

```
Peso [B-TST] corporeo di 19 [B-RML] Kg
[I-RML], protidemia [B-TST] totale 4,5
[B-RML] g/dl [I-RML])
```

Subsequently, an additional 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 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 result entity. For each generated example, the entities in the relationship were

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

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

[4] https://simpletransformers.ai

marked by adding "[TST]" as both the prefix and suffix to the laboratory tests and measurements, while "[RML]" was used to denote the results. The number of examples generated per sentence was determined by multiplying the number of laboratory tests by the number of result entities present in the sentence.

For the test data, the examples to be classified were generated following a similar process, with the difference that instead of using the entities from the gold standard we used the predicted entities. In the case of the sentence reported above, the following examples were generated along with their corresponding model predictions (1=positive, 0=negative):

```
1 [TST]Peso[TST] corporeo di [RML]19
Kg[RML], protidemia totale 4,5 g/dl

0 [TST]Peso[TST] corporeo di 19 Kg,
protidemia totale [RML]4,5 g/dl[RML]

0 Peso corporeo di [RML]19 Kg[RML],
[TST]protidemia[TST] totale 4,5 g/dl

1 Peso corporeo di 19 Kg,
[TST]protidemia[TST] totale [RML]4,5
g/dl[RML]
```

The unsupervised baseline uses GPT and OpenAI's API (text-davinci-003). It focuses on one-shot learning, where the model receives a single example during inference through the prompt. This makes one-shot learning more similar to unsupervised learning than supervised learning. The prompt used for performing this evaluation is: *Ho un compito che è quello di estrarre menzioni di test di laboratorio e dei loro risultati da casi clinici. Ecco un esempio di testo e output: docId:100998. Nota: nell'output viene scritto prima il risultato e poi il nome del test. Sono separati da "|". Ora dammi l'output per il seguente testo.*[5] Within the prompt, docId:100998 represents the annotated document selected from the training dataset as the only example for GPT.

5. System Descriptions

Eight teams expressed their interest to participate in the task. Eventually, four teams submitted their annotated data, resulting in a total of six runs. After the evaluation phase, one team decided to withdraw so we now present the results of four runs submitted by three different teams.

[5]My task is to extract laboratory test mentions and their results from clinical cases. Here you have an example of a text and its output: docId:100998. Notice: in the output you first write the result and then the name of the test. They are separated by "|". Now give me the output for the following text.

Participants explored various (supervised) approaches, including traditional machine learning methods, as well as using BERT [16] and its derivative models, and top Large Language Models (LLMs) such as LLaMA [17]. A brief overview of each team's approach is reported below, while the corresponding results are reported in Table 1.

Simple Ideas: Unlike conventional methods that extract entities and relations separately in a pipeline, the proposed approach uses a pipeline in which first EVENTS are identified and then the Pertains-to relations are created from those. Several BERT-based models were assessed, including Italian BERT [18] and DistilBERT [19], which were pre-trained on general topics. Additionally, BioBIT and MedBIT-R3-plus [20] were evaluated as they were specifically pre-trained for the medical field. Among these models, MedBIT-R3-plus resulted as the best model. To optimize their performance, the models were fine-tuned on an augmented version of the original dataset. This augmentation involved the addition of new sentences derived from the original ones, wherein random words were substituted with similar words in the embedding space. This approach achieved the best results in the task and it also obtained the highest ranking in the parallel TESTLINK task at IberLEF 2023 [21]. The availability of the implemented code contributes to the reproducibility of the presented results.

ExtremITA: The team employed a unified neural model to address all the EVALITA 2023 tasks. To achieve this, they experimented with two different approaches. One approach involved fine-tuning an encoder-decoder model, specifically T5 [22] pre-trained on Italian texts. The second approach is an instruction-tuned Decoder-only model based on the LLaMA [17] foundational models. This model was initially trained on Italian translations of Alpaca [23] instruction data. In both cases, the models were fine-tuned by using the complete set of datasets provided by the EVALITA 2023 tasks. Moreover, the CLinkaRT dataset was expanded with annotated documents derived from the Spanish dataset made available in the TESTLINK task. The model built upon the LLaMA model showed strong performance across multiple tasks at EVALITA 2023, including the CLinkaRT task, where it ranked second. The implemented code has been made available.

Polimi: The team used a traditional pipeline-based approach for relation extraction. The first module focused on recognizing entities related to laboratory tests and their corresponding measurements. The module was implemented using two diverse models: CRF [24] and UmBERTo [25]. For training the CRF, a range of lexical features were used, along with external sources of knowledge like UMLS [26]. Subsequently, the second module aimed at establishing relationships between exams and results by pairing them based on proximity within the same sentence. While the CRF method obtained quite sat-

isfactory results, tokenization issues prevented any results from being obtained using UmBERTo.

6. Results

We conducted the evaluation of systems' performance using the BioCreative V CDR task[6] scorer 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 results of the systems that participated in the task, computed using the standard Precision (Pr), Recall (Re), and F_1 measures, are presented in Table 1. As a comparative reference, we report that the F_1 measure obtained by the best system in the parallel TESTLINK task is 68.38 and 72.65 for Spanish and Basque, respectively. The results of the baselines described in Section 4 are reported in Table 2.

Team	Pr	Re	F_1
Simple Ideas-BERT	65.55	60.62	62.99
ExtremITA-LLaMA	71.10	50.65	59.16
Polimi-CRF	70.34	27.12	39.15
ExtremITA-T5	46.82	26.47	33.82

Table 1
Precision, Recall and F_1 measure obtained by the participating systems.

Baseline	Type	Pr	Re	F_1
mBERT	S	61.37	64.37	62.83
GPT	U	29.55	48.73	36.79
voc. tran.	S	29.95	31.86	30.88

Table 2
Precision, Recall and F_1 measure obtained by the supervised (S) and unsupervised (U) baselines.

We additionally evaluated systems' performance (in terms of F_1 measure) based on two different dimensions. Table 3 shows the results distinguishing two categories of relations, i.e. n-ary relations (one-to-many and many-to-one) and one-to-one relations. Table 4 presents separate results for relations involving numerical RMLs and non-numerical RMLs. Finally, Table 5 reports the accuracy of participant systems in the recognition of RMLs and EVENTs, i.e. the sources and targets of the relations.

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

Team	n-ary	one-to-one
Simple Ideas-BERT	37.50	70.77
ExtremITA-LLaMA	30.77	65.16
Polimi-CRF	ND	46.60
ExtremITA-T5	10.10	33.38

Table 3
F_1 measure of participating systems across n-ary and one-to-one relations.

Team	non-numerical	numerical
Simple Ideas-BERT	43.03	79.09
ExtremITA-LLaMA	47.35	66.36
Polimi-CRF	18.64	49.21
ExtremITA-T5	28.20	37.76

Table 4
F_1 measure of participating systems across relations involving numerical and non-numerical RMLs.

Team	EVENT	RML
Simple Ideas-BERT	74.89	75.26
ExtremITA-LLaMA	65.69	74.63
Polimi-CRF	48.85	49.80
ExtremITA-T5	48.80	58.39

Table 5
F_1 measure per entity type.

7. Discussion

Both traditional machine learning and more recent deep learning models were tested for relation extraction. It is worth noting that all participating systems were based on supervised approaches. Additionally, every system outperformed the vocabulary transfer baseline, which represents the threshold below which systems are not expected to perform.

Surprisingly, none of the teams attempted to evaluate few-shot learning with LLMs such as GPT [27] or LLaMA [17]. However, ExtremITA did evaluate LLaMA, but instead of employing few-shot learning, they opted for a fine-tuning approach, refining the model using the available training data.

The assessment of the GPT-based baseline highlights the present understanding that few-shot learning cannot be considered a viable alternative to fine-tuning in the context of the present task. Fine-tuning, although requiring annotated data, produces significantly better results.

Despite using different pre-trained models trained on diverse domain-specific data (generic domain vs medical domain), the top-performing team (Simple Ideas), along with the second-placed team (ExtremITA) and the baseline model based on multilingual BERT (mBERT), achieved remarkably similar results.

CRF (Polimi), as the exclusive traditional machine learning algorithm involved in the task, obtained a preci-

sion (70.34) in line with that of the top models (71.10). Nevertheless, its relatively lower recall (27.12), in comparison to the recall of the best-performing models (60.62), results in moderately satisfactory outcomes in terms of F_1 score (39.15).

One team (Simple Ideas) conducted an evaluation of their pipeline-based approach in two distinct tasks: the Italian CLinkaRT task (F_1 62.99) and the parallel TESTLINK task at IberLEF 2023, focusing on Basque (F_1 72.65) and Spanish (F_1 68.38). Interestingly, this approach demonstrated superior performance results across all three languages.

Based on the outcome of our analysis of systems' performance in relation to two distinct diagonal dimensions, i.e. n-ary and one-to-one relations on one hand, and numerical and non-numerical RMLs on the other (see Tables 3 and 4, we can observe that extracting n-ary relations is more challenging than extracting one-to-one relations, which is not surprising. Moreover, the task of extracting relations involving numerical RMLs seems easier than extracting relations involving non-numerical entities, which may be correlated to the lower agreement obtained on the latter in the IAA test.

An analysis of the entities involved in the relations extracted by the participants' systems shows that recognising EVENTs seems to be generally harder than recognising RMLs (Table 5). One possible explanation for this is that EVENTs are commonly identified by their syntactic head (leaving out the other elements in the phrase) which can sometimes be quite challenging.

Participants report two key reasons for the incorrect tagging produced by their models. On one hand, BERT tokenizers struggle splitting correctly medical terms (e.g. antitrombina -> anti trombina), which leads to wrongly setting the boundaries of the annotations. In addition, the difficulty of capturing the most peripheral elements in the entity mentions has also been a cause for failing to detect correctly the entity spans. This is the case of "punte di [circa 1200 pg/ml]" or "pari a 0 o [inferiori a 1.5 mg/dl]" in which only the tokens between the brackets have been annotated by the systems.

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 is expected to provide a preliminary indication of the task's difficulty, achieves better results on the Italian CLinkaRT task (F_1 30.88) compared to the parallel TESTLINK task for Basque (F_1 23.96) and Spanish (F_1 22.10). However, the participating systems, such as the Simple Idea's system, showed contrasting results.

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, especially as far as recall is concerned. As this was the first time that we were proposing this task, 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.

Acknowledgments

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

References

[1] K. Jain, V. Prajapati, NLP/Deep Learning Techniques in Healthcare for Decision Making, Primary Health Care 11 (2021). URL: https://www.iomcworld.org/open-access/nlpdeep-learning-techniques-in-healthcare-for-decision-making-66608.html.

[2] O. Sankoh, P. Byass, Cause-specific mortality at INDEPTH Health and Demographic Surveillance System Sites in Africa and Asia: concluding synthesis, Global health action 7 (2014). doi:10.3402/gha.v7.25590.

[3] A. E. Johnson, T. J. Pollard, L. Shen, L.-w. H. Lehman, M. Feng, M. Ghassemi, B. Moody, P. Szolovits, L. Anthony Celi, R. G. Mark, MIMIC-III, a freely accessible critical care database, Scientific Data 3 (2016). URL: http://www.timeml.org/publications/timeMLpubs/IWCS-v4.pdf.

[4] O. Trigueros, A. Blanco, N. Lebeña, A. Casillas, A. Pérez, Explainable ICD multi-label classification of EHRs in Spanish with convolutional attention, International Journal of Medical Informatics 157 (2022) 104615. URL: https://www.sciencedirect.com/science/article/pii/S1386505621002410. doi:https://doi.org/10.1016/j.ijmedinf.2021.104615.

[5] S. Santiso, A. Pérez, A. Casillas, Adverse Drug Reaction extraction: Tolerance to entity recognition errors and sub-domain variants, Computer Methods and Programs in Biomedicine 199 (2021) 105891. URL: https://www.sciencedirect.com/science/article/pii/S0169260720317247. doi:https://doi.org/10.1016/j.cmpb.2020.105891.

[6] D. Newman-Griffis, G. Divita, B. Desmet, A. Zirikly, C. P. Rosé, E. Fosler-Lussier, Ambiguity in medical concept normalization: An analysis of types and coverage in electronic health record datasets, Journal of the American Medical Informatics Association 28 (2020) 516–532. URL: https://doi.org/10.1093/jamia/ocaa269. doi:10.1093/jamia/ocaa269.

[7] G. Alfattni, N. Peek, G. Nenadic, Extraction of temporal relations from clinical free text: A systematic review of current approaches, Journal of Biomedical Informatics 108 (2020) 103488. URL: https://www.sciencedirect.com/science/article/pii/S1532046420301167. doi:https://doi.org/10.1016/j.jbi.2020.103488.

[8] T. Hao, H. Liu, C. Weng, Valx: A System for Extracting and Structuring Numeric Lab Test Comparison Statements from Text, Methods of information in medicine 55 (2016) 266–75. doi:10.3414/ME15-01-0112.

[9] B. Percha, Modern Clinical Text Mining: A Guide and Review, Annual Review of Biomedical Data Science 4 (2021) 165–187. URL: https://doi.org/10.1146/annurev-biodatasci-030421-030931. doi:10.1146/annurev-biodatasci-030421-030931, pMID: 34465177.

[10] M. Lai, S. Menini, M. Polignano, V. Russo, R. Sprugnoli, G. Venturi, Evalita 2023: Overview of the 8th evaluation campaign of natural language processing and speech tools for italian, in: Proceedings of the Eighth Evaluation Campaign of Natural Language Processing and Speech Tools for Italian. Final Workshop (EVALITA 2023), CEUR.org, Parma, Italy, 2023.

[11] B. Magnini, B. Altuna, A. Lavelli, M. Speranza, R. Zanoli, The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases, in: Proceedings of the Seventh Italian Conference on Computational Linguistics, Associazione Italiana di Linguistica Computazionale, Bologna, Italy, 2020. URL: http://ceur-ws.org/Vol-2769/paper_55.pdf.

[12] W. F. Styler, S. Bethard, S. Finan, M. Palmer, S. Pradhan, P. C. de Groen, B. Erickson, T. Miller, C. Lin, G. Savova, et al., Temporal Annotation in the Clinical Domain, Transactions of the Association for Computational Linguistics 2 (2014) 143–154. URL: http://aclweb.org/anthology/Q14-1012.

[13] J. Pustejovsky, J. M. Castaño, R. Ingria, R. Saurí, R. J. Gaizauskas, A. Setzer, G. Katz, D. R. Radev, TimeML: Robust Specification of Event and Temporal Expressions in Text, New directions in question answering 3 (2003) 28–34. URL: http://www.timeml.org/publications/timeMLpubs/IWCS-v4.pdf.

[14] R. Zanoli, A. Lavelli, D. Verdi do Amarante, D. Toti, Assessment of the E3C corpus for the recognition of disorders in clinical texts, Natural Language Engineering (2023) 1–19. doi:10.1017/S1351324923000335.

[15] L. R. Dice, Measures of the amount of ecologic association between species, Ecology 26 (1945) 297–302. URL: http://www.jstor.org/pss/1932409.

[16] 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.

[17] H. Touvron, T. Lavril, G. Izacard, X. Martinet, M. Lachaux, T. Lacroix, B. Rozière, N. Goyal, E. Hambro, F. Azhar, A. Rodriguez, A. Joulin, E. Grave, G. Lample, Llama: Open and efficient foundation language models, CoRR abs/2302.13971 (2023). URL: https://doi.org/10.48550/arXiv.2302.13971. doi:10.48550/arXiv.2302.13971. arXiv:2302.13971.

[18] S. Schweter, Italian BERT and ELECTRA models. Version 1.0.1, 2020. URL: https://doi.org/10.5281/zenodo.4263142. doi:10.5281/zenodo.4263142.

[19] V. Sanh, L. Debut, J. Chaumond, T. Wolf, DistilBERT, a distilled version of BERT: smaller, faster, cheaper and lighter, in: 5th Workshop on Energy Efficient Machine Learning and Cognitive Computing @ NeurIPS 2019, 2019. URL: http://arxiv.org/abs/1910.01108.

[20] T. M. Buonocore, C. Crema, A. Redolfi, R. Bellazzi, E. Parimbelli, Localising in-domain adaptation of transformer-based biomedical language models, ArXiv abs/2212.10422 (2022).

[21] 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).

[22] G. Sarti, M. Nissim, IT5: Large-scale Text-to-text Pretraining for Italian Language Understanding and Generation, ArXiv preprint 2203.03759 (2022). URL: https://arxiv.org/abs/2203.03759.

[23] R. Taori, I. Gulrajani, T. Zhang, Y. Dubois, X. Li, C. Guestrin, P. Liang, T. B. Hashimoto, Stanford Alpaca: An Instruction-following LLaMA model, https://github.com/tatsu-lab/stanford_alpaca, 2023.

[24] A. McCallum, W. Li, Early results for named entity recognition with conditional random fields, feature induction and web-enhanced lexicons, in: Proceed-

ings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003, 2003, pp. 188–191. URL: https://aclanthology.org/W03-0430.

[25] F. Tamburini, How "BERTology" Changed the State-of-the-Art also for Italian NLP, in: J. Monti, F. dell'Orletta, F. Tamburini (Eds.), Proceedings of the Seventh Italian Conference on Computational Linguistics, CLiC-it 2020, Bologna, Italy, March 1-3, 2021, volume 2769 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2769/paper_79.pdf.

[26] O. Bodenreider, The Unified Medical Language System (UMLS): integrating biomedical terminology., Nucleic Acids Res. 32 (2004) 267–270. URL: http://dblp.uni-trier.de/db/journals/nar/nar32.html#Bodenreider04.

[27] T. Brown, B. Mann, N. Ryder, M. Subbiah, J. D. Kaplan, P. Dhariwal, A. Neelakantan, P. Shyam, G. Sastry, A. Askell, S. Agarwal, A. Herbert-Voss, G. Krueger, T. Henighan, R. Child, A. Ramesh, D. Ziegler, J. Wu, C. Winter, C. Hesse, M. Chen, E. Sigler, M. Litwin, S. Gray, B. Chess, J. Clark, C. Berner, S. McCandlish, A. Radford, I. Sutskever, D. Amodei, Language models are few-shot learners, in: H. Larochelle, M. Ranzato, R. Hadsell, M. Balcan, H. Lin (Eds.), Advances in Neural Information Processing Systems, volume 33, Curran Associates, Inc., 2020, pp. 1877–1901. URL: https://proceedings.neurips.cc/paper_files/paper/2020/file/1457c0d6bfcb4967418bfb8ac142f64a-Paper.pdf.