CT-EBM-SP - Corpus of Clinical Trials for Evidence-Based-Medicine in Spanish (version 3)

Project documentation

Companion file to article:

Transformer-based relation extraction and concept normalization using an annotated clinical trials corpus

Leonardo Campillos-Llanos[1,*], Ana Valverde-Mateos[2], Adrián Capllonch-Carrión[3], Sofía Zakhir-Puig[4], and Jónathan Heras-Vicente[5]

[1]ILLA, CSIC, Madrid, 28037, Spain.
[2]UTM RANME, Madrid, 28013, Spain
[3]CSM Retiro, Hospital Gregorio Marañón, Madrid, 28009, Spain
[4]IULMA, Universitat de València, Valencia, 46010, Spain
[5]Departamento de Matemáticas y Computación, Universidad de La Rioja, Logroño, 26004, Spain
[*]leonardo.campillos@csic.es **ORCID**: 0000-0003-3040-1756

Related Annotation Work

Corpus	Text type and size	Normalizations (count)
BioCreative[1]	6218 journal abstracts.	Normalization to yeast, fly and mouse genes identifiers.
CANTEMIST[2]	1301 anonymized oncology clinical cases.	16030 entities normalized to the ICD-O-3 tumour classification (850 unique codes).
DisTEMIST[3]	1000 anonymized clinical cases extracted from journal articles.	10 665 entities normalized to SNOMED-CT (7303 unique codes).
NCBI corpus[4]	793 abstracts.	Disease entities (6892) normalized to MeSH and OMIM.
CRAFT[5]	67 full-text journal articles.	Nearly 100 000 concepts from to the Gene Ontology.
ShARe (SemEval 2015 Task 14[6])	531 clinical notes.	Disorder entities (19 111) normalized to the UMLS.
MCN corpus[7]	100 discharge summaries.	10 919 concept mentions normalized to the UMLS.
QUAERO[8]	2500 Medline titles, 13 texts from EMEA, and 25 EPO patents in French	26 409 entity annotations were mapped to 26 281 UMLS CUIs (5797 unique codes).
MANTRA[9]	1450 texts from EMEA, Medline and biomedical patent claims in English, French, German, Spanish and Dutch.	5530 annotations (total for all languages) normalized to the UMLS.
E3C[10]	422 anonymized clinical cases extracted from journal articles in Basque, English, French, Italian and Spanish.	6475 entities (total for all languages) normalized to the UMLS.

Acronyms: *CT*: clinical trials; *CUI*: Concept Unique Identifier; *E3C*: European Clinical Case Corpus; *EMEA*: European Medicines Agency; *ICD-O-3*: International Classification of Diseases for Oncology vs 3; *EPO*: European Patent Office; *MCN*: Medical Concept Normalization; *MeSH*: Medical Subject Headings; *NCBI*: National Center for Biotechnology Information; *OMIM*: Online Mendelian Inheritance in Man; *SNOMED-CT*: Systematized Nomenclature of Medicine – Clinical Terms; *UMLS*: Unified Medical Language System.

Table 1. Biomedical and CT-related corpora for medical concept normalization.

Corpus	Text type and size	Annotations (count)
i2b2[11]	Patient reports (394 in train + 477 in test).	3 entity types (in total, 27 837 train, 5009 test), and 6 relation types (5264 train; 9069 test), as reported in[12].
ADE[13]	Medical reports (4272 sentences).	3 entity types: Drug (5063), Dosage (231) and Adverse-Effect (5776); 2 relation types: Adverse-Effect (6821) and Drug-dosage (279).
DDI[14]	1025 documents from DrugBank and Medline.	4 entity types (18 502) and 4 types of DDI relationships (5028).
CDR[15]	1500 abstracts from Pubmed.	Chemical entities (4409), Disease entities (5818) and Chemical-Disease relations (3116).
n2c2 2018 ADE[16]	505 discharge summaries.	9 entity types (83 869) and 8 relation types (59 810).
Chia[17]	1000 texts from ClinicalTrials.gov (12 409 elibility criteria).	15 entity types (41 487) and 12 different relationships (25 017).
Tseo et al. (2020)[18]	3314 trials from ClinicalTrials.gov	10 clinical entity types, 5 attributes and value limits (121 221); negation relation (5054).
Tseo et al. (2020)[19]	470 study protocols from a private database.	15 clinical entity types and 7 relation types.
Nye et al. (2021)[20]	1932 journal abstracts describing CTs.	Entities describing interventions and outcomes (13895), and relations between them (5054).
BioRED[21]	600 PubMed abstracts.	6 entity types (20 419), 8 relationships (6503) and relation pairs with novelty findings (4532).
SNPPhenA[22]	360 abstracts from PubMed.	SNPs (868), phenotypes (590), SNP-Phenotype association candidates (1300), neutral (308), negative (120) and positive candidates (872); negation and modality markers.
DGO[23]	1622 oncology exams and breast cancer notes in Portuguese.	10 entity types (146 769) and 3 relation types (111 716).
LEAF[24]	1006 CT descriptions from ClinicalTrials.gov (105 816 annotations).	50 entity types (41 487) and 51 relation types (25 017).
RCT-ART[25]	558 sentences from 6 disease areas.	3 entity types (3541) and 3 relationships (3182).

Table 2. Biomedical and CT-related corpora annotated with relationships.

Acronyms: *ADE*: Adverse Drug Effects; *CDR*: Chemical-Disease Relations; *CTs*: Clinical Trials; *DDI*: Drug-Drug Interactions; *SNPs*: Single Nucleotide Polimorphisms.

Annotation Scheme

UMLS	ANAT: Anatomical structure, body substance or cell.	*músculo* (‘muscle’)
	CHEM: Chemical or pharmacological substance.	*antibiótico* (‘antibiotic’)
	DEVI: Medical device.	*sonda* (‘probe’)
	DISO: Pathology, disorder, neoplasms, injuries or signs and symptoms.	*cancer*, *fiebre* (‘fever’)
	GENE: Nucleotide sequence, gene or genome.	*HER2*, *BRAF*
	LIVB: Human; professional or population group; animal, bacterium, fungus; plant and virus.	*paciente* (‘patient’), *S. aureus*
	PHYS: Biologic, cell, genetic or physiologic function.	*digestión* (‘digestion’)
	PROC: Diagnostic, therapeutic or laboratory procedure; health care activity; research activity.	*hemograma* (‘hemogram’), *bypass*
Intervention	Contraindicated: Any procedure or drug that is a contraindication. [Attribute]	*Contraindicación a AAS* (‘ASA contraindication’)
	Dose: Drug dose or strength.	*2mg*, *2%*
	Form: Dosage form.	*pastilla* (‘pill’)
	Route: Route or mode of administration.	*per os* (‘oral’)
Temporal	Age. [entity or attribute]	*18 años* (‘18 y.o.’)
	Date: Specific or relative date.	*2020*, *hoy* (‘today’)
	Duration: Length of time that an event lasts.	*6 meses* (‘6 months’)
	Frequency: How often an event occurs.	*semanal* (‘weekly’)
	Time: Hour or part of the day.	*15:00*
Assertion	Neg_cue: Negation cue.	*no*
	Negated: Negated concept or event. [Attribute]	*sin fiebre* (’no fever’)
	Spec_cue: Speculation cue.	*posible* (‘possible’)
	Speculated: Speculated concept or event. [Attribute]	*probable diabetes*
Event temporality	Family_History_of: Record of medical conditions or procedures in a patient’s family. [Attribute]	*Antecedentes familiares: DB* (‘Family history: DB’)
	Future: Any expected event or a pending procedure. [Attribute]	*embarazo planificado* (’planned pregnancy’)
	History_of: Record of medical conditions or procedures. [Attribute]	*Antecedentes de ictus* (’history of stroke’)
	Hypothetical: Any event that is imagined, but it is not truly happening at the moment. [Attribute]	*si es mujer, debe...* (‘If female, subject must...’)
Experiencer	Family_member: A patient’s relative. [Attribute]	*padre* (’father’)
	Other: Any person that is not the patient/participant nor a relative. [Attribute]	*representante legal* (’legal representative’)
	Patient: ill person or participant in a trial. [Attribute]	*pacientes pediátricos* (’pediatric patients’)
Other	Concept: idea, evaluation scale, procedure approach or temporal concept not in TimeML.	*esperanza de vida* (‘life expectancy’)
	Food or Drink	*soja* (‘soy‘), *alcohol*
	Observation: Clinical findings or conditions that are not pathological states.	*normotenso* (’normal pressure’)
	Quantifier or Qualifier: modifiers/adjectives needed to understand the text.	*grave* (‘severe’)
	Result or Value: Outcome of a lab or a procedure.	> *140 kg*

Table 3. Annotated entities and attributes with examples.

	Relation type	Examples
Assertion	Negation	*no* → *fiebre* ('no fever')
	Speculation	*posible* → *diabetes* ('possible diabetes')
Intervention-related / temporal	Has_Dose_ or_Strength	*aspirina* → *500 mg*
	Has_Form	*aspirina* → *comprimido* ('tablet')
	Has_Route_or_Mode	*aspirina* → *oral*
	Has_Duration_or_ interval	*aspirina* → *un mes* ('one month')
	Has_Frequency	*aspirina* → *diaria* ('daily')
	Has_Age	*paciente* ('patient') → *50 años* ('50 y.o.')
	After	*sutura* ('suture') → *operación* ('surgery')
	Before	*anestesia* ('anesthesia') → *operación* ('surgery')
	Overlap	*operación* ('surgery') → *2009*
	Combined_with	*dolutegravir* → *lamivudina* ('lamivudine')
	Used_for	*aspirina* → *tratamiento* ('treatment')
Other	Causes	*VIH* ('HIV') → *SIDA* ('AIDS')
	Experiences	*paciente* ('patient') → *diabetes*
	Has_Quantifier_or _Qualifier	*asma* ('asthma') → *grave* ('severe')
	Has_Result_or_Value	*IMC* ('BMI') → *> 25*
	Location_of	*dolor* ('pain') → *oído* ('ear')

Table 4. Annotated relations with examples.

Annotation Results

We annotated 86391 entities, including 13982 nested entities (16.18%) and 534 (0.62%) discontinuous entities; 16574 attributes and 67593 relationships. 81.71% of entities were normalized. Tables 5 and 6 report the count of entities, attributes and relations. Table 7 shows the count of normalized entities, including the unclear normalization (i.e., entities marked with the '?' character) and the complex normalization (i.e., entities mapped to two or more UMLS CUIs). Figure 1 also shows a chord diagram with the 20 most frequent relations between entities (we used the normalized CUIs to count the relations, and showed the English preferred term).

	Abstracts	**EudraCT**	**Total**
ANAT	2742	4159	6901
CHEM	4345	4974	9319
DEVI	437	190	627
DISO	4310	8888	13198
GENE	16	75	91
LIVB	3461	4955	8416
PHYS	874	1475	2349
PROC	9235	10103	19338
Contraindicated	4	263	267
Dose	869	223	1092
Form	166	68	234
Route	546	422	968
Age	527	1051	1578
Date	605	1302	1907
Duration	1310	1312	2622
Frequency	300	126	426
Time	343	70	413
Concept	2257	1041	3298
Observation	1718	2221	3939
Food_or_Drink	146	44	190
Quantifier_or_Qualifier	1853	3100	4953
Result_or_Value	221	1407	1628
Neg_cue	1195	1564	2759
Negated	1583	2102	3685
Spec_cue	351	464	815
Speculated	509	611	1120
Family_history_of	0	16	16
Future	10	391	401
History_of	60	2716	2776
Hypothetical	9	113	122
Family_member	30	63	93
Other	77	386	463
Patient	3005	3956	6961
Total entities	37512	48879	86391
Nested entities	5945	8037	13982 (16.18%)
Discontinuous entities	334	200	534 (0.62%)
Total attributes	5602	10972	16574

Table 5. Distribution of annotated entities and attributes per type.

	Abstracts	EudraCT	Total
After	754	455	1209
Before	3113	2960	6073
Causes	1634	2163	3797
Combined_with	753	677	1430
Experiences	7739	8989	16728
Has_Age	253	573	826
Has_Dose_or_Strength	920	294	1214
Has_Drug_Form	185	78	263
Has_Duration_or_Interval	1029	1111	2140
Has_Frequency	370	167	537
Has_Quantifier_or_Qualifier	1927	3754	5681
Has_Result_or_Value	486	1635	2121
Has_Route_or_Mode	661	501	1162
Location_of	2761	4349	7110
Negation	1554	2050	3604
Overlap	3366	4493	7859
Speculation	488	600	1088
Used_for	2246	2505	4751
Total	30239	34849	67593

Table 6. Distribution of annotated relations per type.

	Abstracts	EudraCT	Total
Normalized entities	30202 (80.51%)	40388 (82.63%)	70590 (81.71%)
Not-normalized entities	7310 (19.49%)	8491 (17.37%)	15801 (18.29%)
Unclear normalization	1208 (3.22%)	1136 (2.32%)	2344 (2.71%)
Complex normalization	5323 (14.19%)	6113 (12.51%)	10304 (11.93%)
Total entities	37512	48879	86391

Table 7. Count of normalized entities.

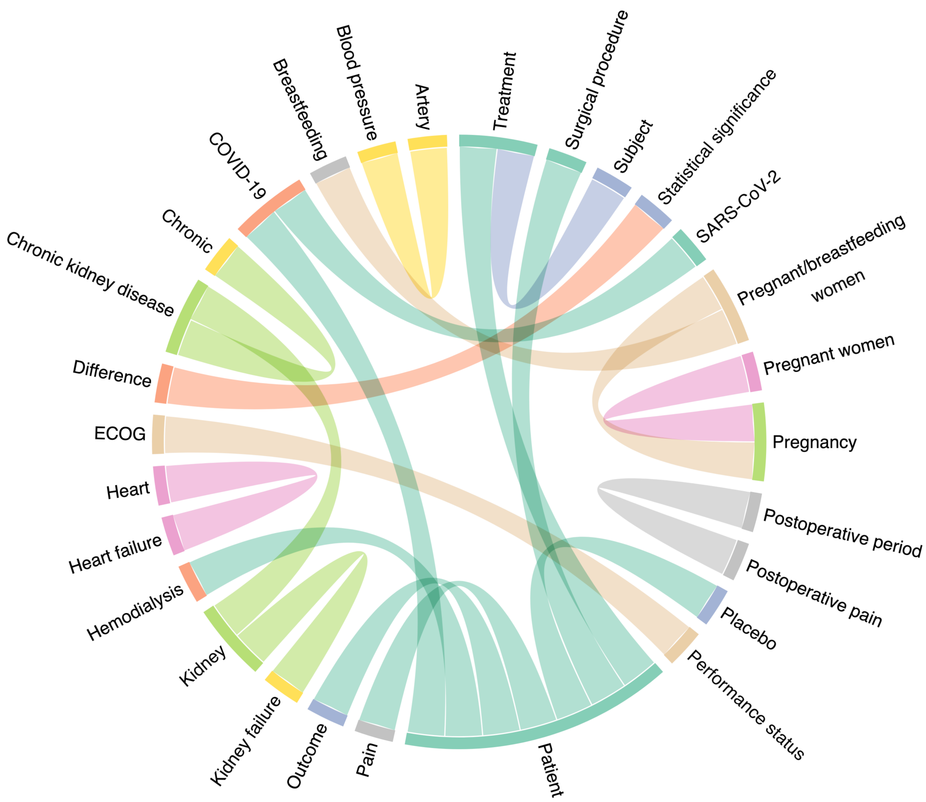

Figure 1. Chord diagram of the 20 most frequent relations between entities.

Relation Extraction Results

	Precision	Recall	F1	Accuracy
`bert-base-multilingual-cased`	0.884 (±0.006)	0.874 (±0.003)	**0.879 (±0.005)**	**0.917 (±0.001)**
`bert-base-multilingual-cased vs 2`	0.880 (±0.006)	**0.876 (±0.005)**	0.878 (±0.003)	0.916 (±0.001)
`dccuchile-bert-base-spanish-wwm-uncased`	0.868 (±0.009)	0.857 (±0.006)	0.862 (±0.006)	0.907 (±0.003)
`dccuchile-bert-base-spanish-wwm-uncased vs 2`	0.877 (±0.008)	0.854 (±0.005)	0.864 (±0.006)	0.908 (±0.003)
`microsoft-mdeberta-v3-base`	0.873 (±0.010)	0.859 (±0.005)	0.864 (±0.005)	0.909 (±0.005)
`microsoft-mdeberta-v3-base vs 2`	**0.886 (±0.003)**	0.857 (±0.007)	0.869 (±0.005)	0.911 (±0.003)
`RoBERTa-bsc-bio-ehr-es`	0.867 (±0.007)	0.852 (±0.006)	0.858 (±0.006)	0.907 (±0.005)
`RoBERTa-bsc-bio-ehr-es vs 2`	0.875 (±0.009)	0.858 (±0.004)	0.865 (±0.006)	0.910 (±0.004)

Table 2. Results of the relation extraction task (average ± standard deviation of 5 experimental rounds).

One-way analysis of variance					
P value	< 0.0001				
P value summary	***				
Are means signif. different? (P < 0.05)	Yes				
Number of groups	8				
F	8.594				
R square	0.6528				
Bartlett's test for equal variances					
Bartlett's statistic (corrected)	1.512				
P value	0.9819				
P value summary	ns				
Do the variances differ signif. (P < 0.05)	No				
ANOVA Table	**SS**	**df**	**MS**		
Treatment (between columns)	0.001849	7	0.0002641		
Residual (within columns)	0.0009833	32	3.073e-005		
Total	0.002832	39			
Bonferroni's Multiple Comparison Test	**Mean Diff.**	**t**	**Significant?** P < 0.05**?**	**Summary**	**95% CI of diff**
mBERT vs mBERT v2	0.001833	0.5229	No	ns	-0.01011 to 0.01378
mBERT vs BSC-BIO	0.01717	4.897	Yes	***	0.005220 to 0.02911
mBERT vs BSC-BIO v2	0.01333	3.803	Yes	*	0.001386 to 0.02528
mBERT vs BETO	0.02083	5.942	Yes	***	0.008886 to 0.03278
mBERT vs BETO v2	0.01417	4.041	Yes	**	0.002220 to 0.02611
mBERT vs mDeBERTa	0.01450	4.136	Yes	**	0.002553 to 0.02645
mBERT vs mDeBERTa v2	0.009167	2.615	No	ns	-0.002780 to 0.02111
mBERT v2 vs BSC-BIO	0.01533	4.374	Yes	**	0.003386 to 0.02728
mBERT v2 vs BSC-BIO v2	0.0115	3.280	No	ns	-0.0004470 to 0.02345
mBERT v2 vs BETO	0.01900	5.419	Yes	***	0.007053 to 0.03095
mBERT v2 vs BETO v2	0.01233	3.518	Yes	*	0.0003863 to 0.02428
mBERT v2 vs mDeBERTa	0.01267	3.613	Yes	*	0.0007195 to 0.02461
mBERT v2 vs mDeBERTa v2	0.007333	2.092	No	ns	-0.004614 to 0.01928
BSC-BIO vs BSC-BIO v2	-0.003833	1.093	No	ns	-0.01578 to 0.008114
BSC-BIO vs BETO	0.003667	1.046	No	ns	-0.008280 to 0.01561
BSC-BIO vs BETO v2	-0.003000	0.8557	No	ns	-0.01495 to 0.008947
BSC-BIO vs mDeBERTa	-0.002667	0.7606	No	ns	-0.01461 to 0.009280
BSC-BIO vs mDeBERTa v2	-0.008000	2.282	No	ns	-0.01995 to 0.003947
BSC-BIO v2 vs BETO	0.007500	2.139	No	ns	-0.004447 to 0.01945
BSC-BIO v2 vs BETO v2	0.0008334	0.2377	No	ns	-0.01111 to 0.01278
BSC-BIO v2 vs mDeBERTa	0.001167	0.3327	No	ns	-0.01078 to 0.01311
BSC-BIO v2 vs mDeBERTa v2	-0.004167	1.188	No	ns	-0.01611 to 0.007780
BETO vs BETO v2	-0.006667	1.902	No	ns	-0.01861 to 0.005280
BETO vs mDeBERTa	-0.006333	1.806	No	ns	-0.01828 to 0.005614
BETO vs mDeBERTa v2	-0.01167	3.328	No	ns	-0.02361 to 0.0002804
BETO v2 vs mDeBERTa	0.0003332	0.09504	No	ns	-0.01161 to 0.01228
BETO v2 vs mDeBERTa v2	-0.005000	1.426	No	ns	-0.01695 to 0.006947
mDeBERTa vs mDeBERTa v2	-0.005333	1.521	No	ns	-0.01728 to 0.006614

Table 7. One-way analysis of variance and Bonferroni's multiple comparison test between models for the F1 measure of the relation extraction task.

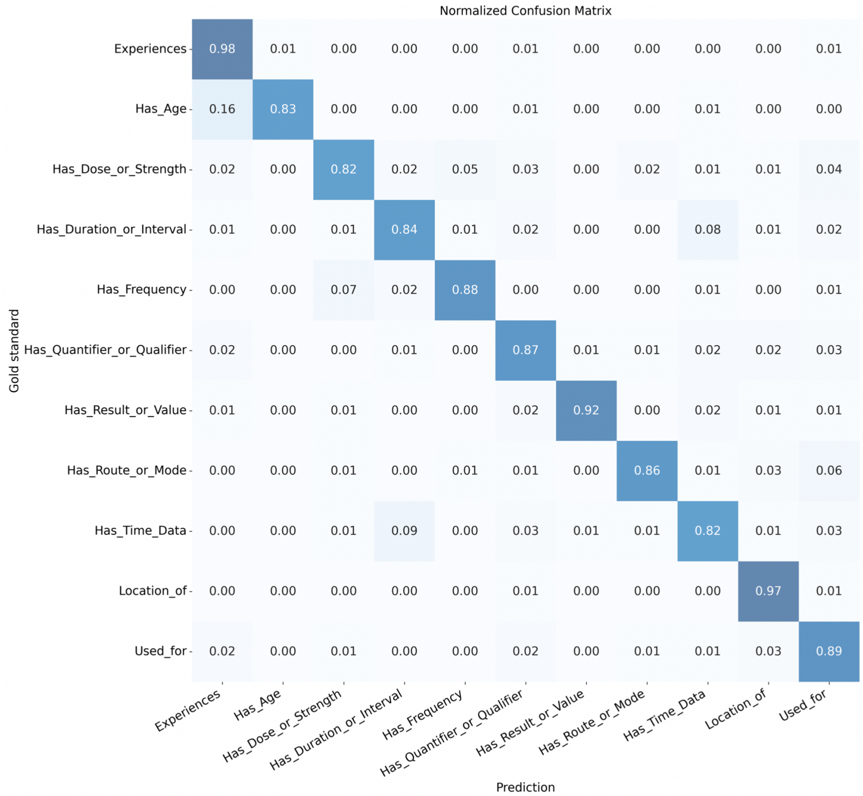

Figure 2. Confusion matrix of the bert-base-multilingual-cased vs 2 model.

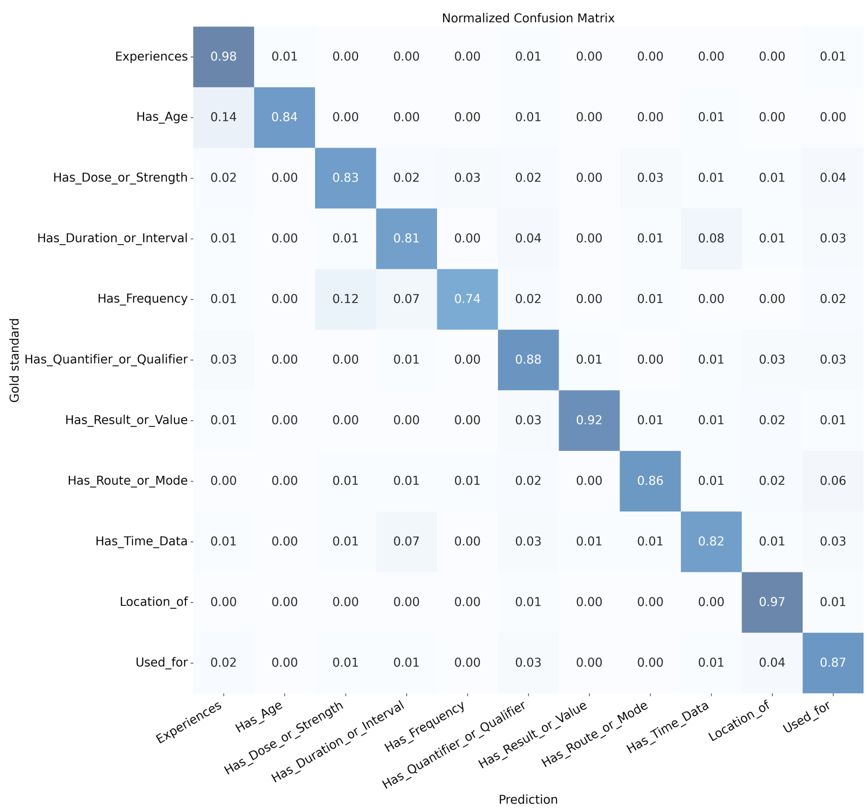

Figure 3. Confusion matrix of the dccuchile-bert-base-spanish-wwm-uncased vs 2 model.

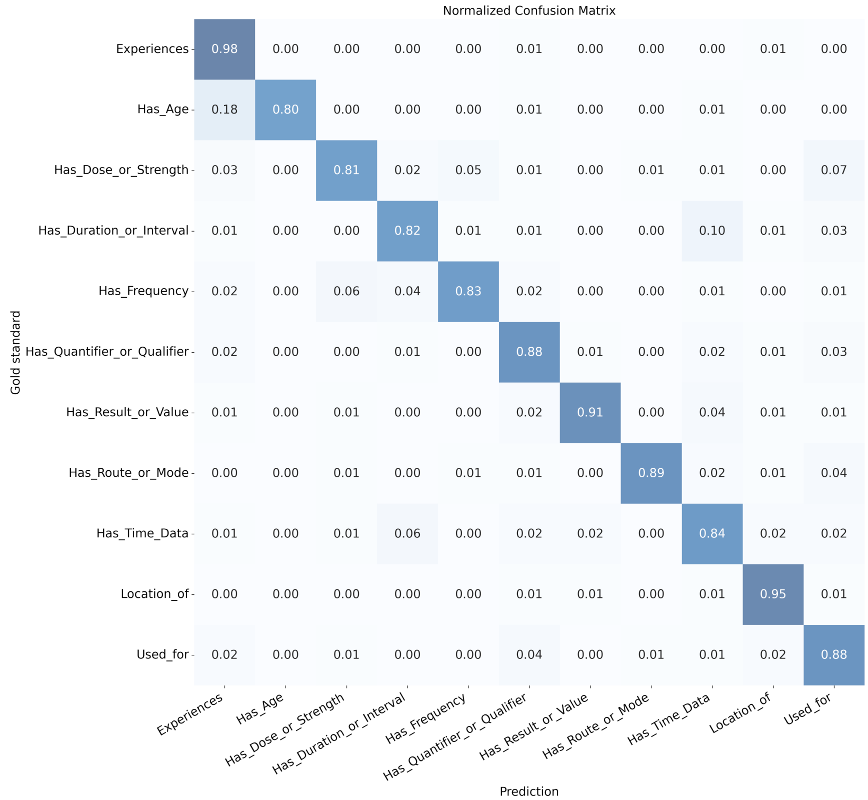

Figure 4. Confusion matrix of the microsoft-mdeberta-v3-base vs 2 model.

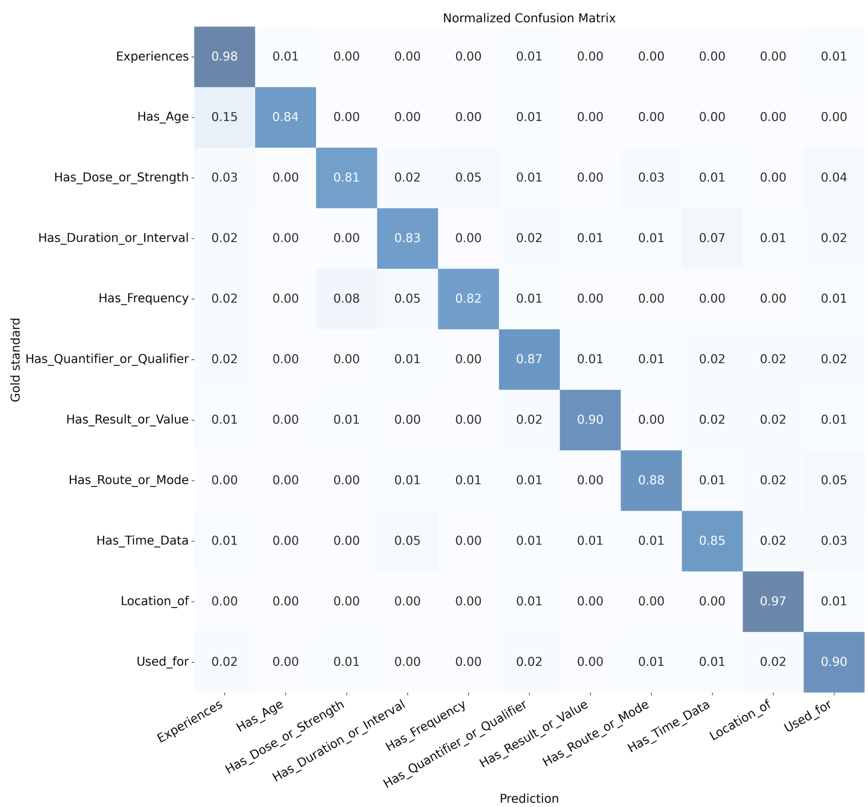

Figure 5. Confusion matrix of the RoBERTa-bsc-bio-ehr-es vs 2 model.

Explainability

We applied an explainable artificial intelligence (XAI) method to visualize the `bert-base-multilingual-cased` model's attention weights for specific input tokens (Figures S4-S7, Supplementary material). This analysis tends to show that Transformer models focus on key words in the medical entities from the sentence context to predict a relationship.

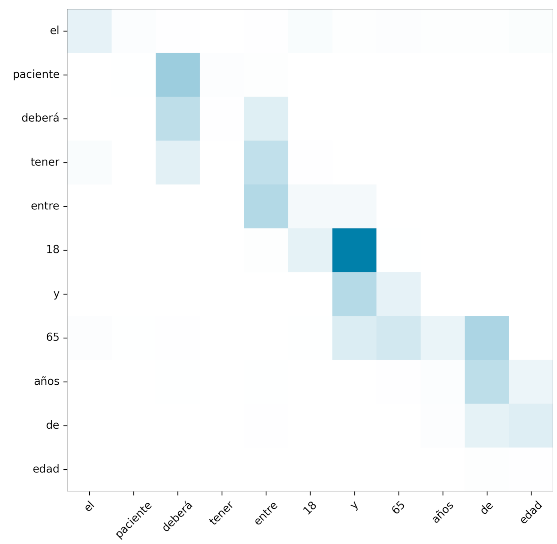

Figure 6. Attention weights predicting relation Has_Age in sentence 'The patient must be between the ages of 18 and 65'.

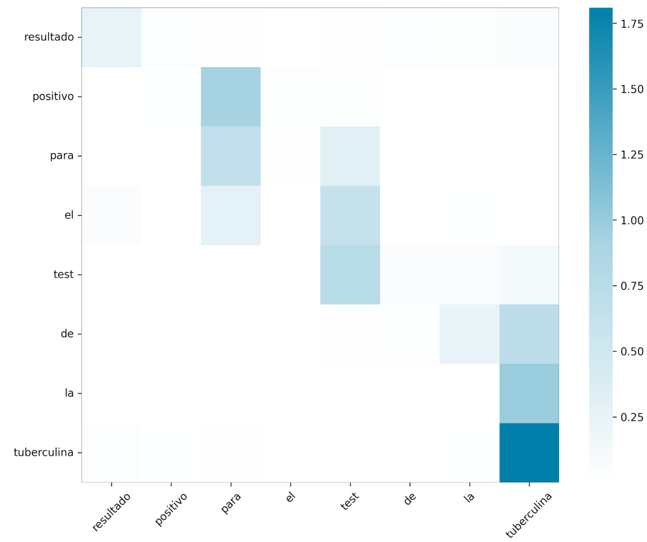

Figure 7. Attention weights predicting relation Has_Result_or_Value in sentence 'Positive result in the tuberculin test'.

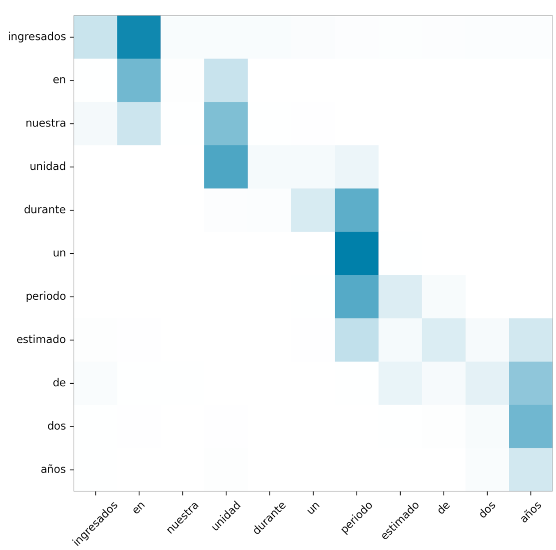

Figure 8. Attention weights predicting relation Has_Duration in sentence 'Admitted to our unit for an estimated period of two years'.

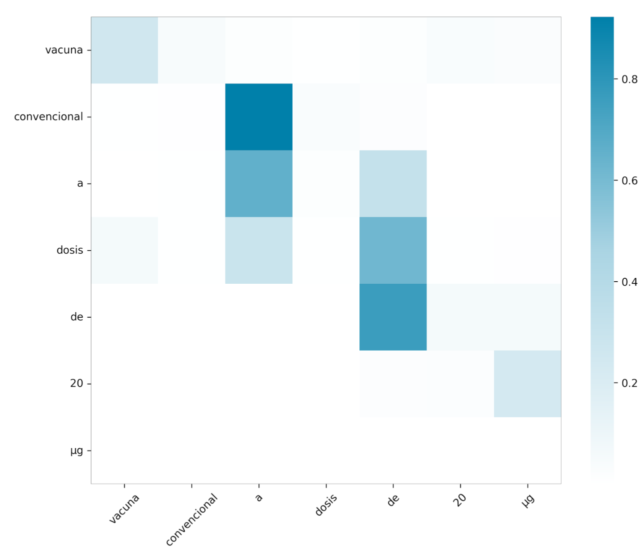

Figure 9. Attention weights predicting relation Has_Dose_or_Strength in sentence 'Conventional vaccine at a dose of 20 μg'.

References

1. Hirschman, L., Colosimo, M., Morgan, A. & Yeh, A. Overview of BioCreAtIvE task 1B: normalized gene lists. *BMC bioinform* **6**, 1–10, https://doi.org/10.1186/1471-2105-6-S1-S11 (2005).
2. Miranda, A., Farré, E. & Krallinger, M. Named Entity Recognition, Concept Normalization and Clinical Coding: Overview of the Cantemist Track for Cancer Text Mining in Spanish: Corpus, Guidelines, Methods and Results. *Proc. IberLEF* 303–23 (2020).
3. Miranda-Escalada, A. *et al.* Overview of DisTEMIST at BioASQ: Automatic detection and normalization of diseases from clinical texts: results, methods, evaluation and multilingual resources. In *CLEF (Working Notes)*, 179–203 (2022).
4. Doğan, R. I., Leaman, R. & Lu, Z. NCBI disease corpus: a resource for disease name recognition and concept normalization. *J Biomed Inf.* **47**, 1–10, https://doi.org/10.1016/j.jbi.2013.12.006 (2014).
5. Bada, M. *et al.* Concept annotation in the CRAFT corpus. *BMC Bioinform* **13**, 161, https://doi.org/10.1186/1471-2105-13-161 (2012).
6. Elhadad, N. *et al.* SemEval-2015 task 14: Analysis of clinical text. In *Proc. of SemEval 2015*, 303–310 (Association for Computational Linguistics, 2015).
7. Luo, Y.-F., Sun, W. & Rumshisky, A. MCN: a comprehensive corpus for medical concept normalization. *J Biomed Inf.* **92**, 103132, https://doi.org/10.1016/j.jbi.2019.103132 (2019).
8. Névéol, A., Grouin, C., Leixa, J., Rosset, S. & Zweigenbaum, P. The Quaero French Medical Corpus: A Ressource for Medical Entity Recognition and Normalization. In *Proc. of BioTxtM 2014*, 24–30 (2014).
9. Kors, J. A., Clematide, S., Akhondi, S. A., van Mulligen, E. M. & Rebholz-Schuhmann, D. A multilingual gold-standard corpus for biomedical concept recognition: the Mantra GSC. *JAMIA* **22**, 948–56, https://doi.org/10.1093/jamia/ocv037 (2015).
10. Magnini, B. *et al.* European Clinical Case Corpus. In *European Language Grid: A Language Technology Platform for Multilingual Europe*, 283–288, https://doi.org/10.1007/978-3-031-17258-8_17 (Springer Internat. Publ. Cham, 2022).
11. Uzuner, Ö., South, B. R., Shen, S. & DuVall, S. L. 2010 i2b2/VA challenge on concepts, assertions, and relations in clinical text. *JAMIA* **18**, 552–556, https://doi.org/10.1136/amiajnl-2011-000203 (2011).
12. Chikka, V. R. & Karlapalem, K. A hybrid deep learning approach for medical relation extraction. *arXiv preprint arXiv:1806.11189* (2018).
13. Gurulingappa, H. *et al.* Development of a benchmark corpus to support the automatic extraction of drug-related adverse effects from medical case reports. *J Biomed Inf.* **45**, 885–892, https://doi.org/10.1016/j.jbi.2012.04.008 (2012).
14. Herrero-Zazo, M., Segura-Bedmar, I., Martínez, P. & Declerck, T. The DDI corpus: An annotated corpus with pharmacological substances and drug–drug interactions. *J Biomed Inform* **46**, 914–20, https://doi.org/10.1016/j.jbi.2013.07.011 (2013).
15. Li, J. *et al.* BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database* **2016**, https://doi.org/10.1093/database/baw068 (2016).
16. Henry, S., Buchan, K., Filannino, M., Stubbs, A. & Uzuner, O. 2018 n2c2 shared task on adverse drug events and medication extraction in electronic health records. *JAMIA* **27**, 3–12, https://doi.org/10.1093/jamia/ocz166 (2020).
17. Kury, F. *et al.* Chia, a large annotated corpus of clinical trial eligibility criteria. *Sci Data* **7**, 1–11, https://doi.org/10.1038/s41597-020-00620-0 (2020).
18. Tseo, Y., Salkola, M., Mohamed, A., Kumar, A. & Abnousi, F. Information extraction of clinical trial eligibility criteria. In *Proc. of the KDD Workshop on Applied Data Science for Healthcare. 2020*, 1–4 (2020).
19. Chen, M., Du, F., Lan, G. & Lobanov, V. S. Using Pre-trained Transformer Deep Learning Models to Identify Named Entities and Syntactic Relations for Clinical Protocol Analysis. In *Proc. AAAI Spring Symposium*, 1–8 (2020).
20. Nye, B. E. *et al.* Understanding clinical trial reports: Extracting medical entities and their relations. In *AMIA Jt Summits Transl Sci Proc*, vol. 2021, 485 (American Medical Informatics Association, 2021).
21. Luo, L., Lai, P.-T., Wei, C.-H., Arighi, C. N. & Lu, Z. BioRED: a rich biomedical relation extraction dataset. *Brief Bioinform* **23**, bbac282, https://doi.org/10.1093/bib/bbac282 (2022).
22. Bokharaeian, B., Dehghani, M. & Diaz, A. Automatic extraction of ranked SNP-phenotype associations from text using a BERT-LSTM-based method. *BMC Bioinform* **24**, 144, https://doi.org/10.1186/s12859-023-05236-w (2023).
23. Silva, D., Rosa, W., Mello, B., Vieira, R. & Rigo, S. Exploring named entity recognition and relation extraction for ontology and medical records integration. *Inf. Med Unlocked* **43**, 101381, https://doi.org/10.1016/j.imu.2023.101381 (2023).
24. Dobbins, N. J., Mullen, T., Uzuner, Ö. & Yetisgen, M. The Leaf Clinical Trials Corpus: a new resource for query generation from clinical trial eligibility criteria. *Sci Data* **9**, 490, https://doi.org/10.1038/s41597-022-01521-0 (2022).
25. Whitton, J. & Hunter, A. Automated tabulation of clinical trial results: A joint entity and relation extraction approach with transformer-based language representations. *Artif Intell Med* **144**, 102661, https://doi.org/10.1016/j.artmed.2023.102661 (2023).