The E3C Project: European Clinical Case Corpus

El proyecto E3C: European Clinical Case Corpus

Bernardo Magnini[1], **Begoña Altuna**[1,2], **Alberto Lavelli**[1],
Manuela Speranza[1], **Roberto Zanoli**[1]
[1]Fondazione Bruno Kessler
[2]Universidad del País Vasco/Euskal Herriko Unibertsitatea
{magnini;altuna;lavelli;manspera;zanoli}@fbk.eu

Abstract: The European Clinical Case Corpus (E3C) project aims at collecting and annotating a large corpus of clinical documents in five European languages (Spanish, Basque, English, French and Italian), which will be freely distributed. Annotations include temporal information, to allow temporal reasoning on chronologies, and information about clinical entities based on medical taxonomies, to be used for semantic reasoning.

Keywords: Clinical data, corpus, multilingual, temporal information, clinical entities.

Resumen: El proyecto European Clinical Case Corpus (E3C) pretende reunir y anotar un gran crorpus de documentos clínicos en cinco lenguas europeas (español, euskera, inglés, francés e italiano) que será distribuido libremente. Las anotaciones incluyen información temporal para permitir el razonamiento temporal en cronologías e información sobre entidades clínicas basada en taxonomías médicas para su uso en razonamiento semántico.

Palabras clave: información clínica, multilingüe, información temporal, entidades clínicas.

1 *Introduction*

E3C, *European Clinical Case Corpus*, is a one-year project (started in July 2020) aiming at creating a corpus of clinical documents in five European languages: Spanish, Basque, English, French and Italian. The project is partially supported by the European Language Grid project through its open call for pilot projects[1]. On its hand, the European Language Grid project has received funding from the European Union's Horizon 2020 Research and Innovation programme under Grant Agreement no. 825627 (ELG). The project is led by FBK (Fondazione Bruno Kessler) and part of the activities has been subcontracted to the Université d'Orléans.

The core of the corpus is a manually annotated dataset of clinical cases. A clinical case reports 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 (Box 1).

> A 25-year-old man with a history of Klippel-Trenaunay syndrome presented to the hospital with mucopurulent bloody stool and epigastric persistent colic pain for 2 wk. Colonoscopy showed continuous superficial ulcers and bleeding. Subsequent gastroscopy revealed mucosa with diffuse edema, ulcers, errhysis, and granular and friable changes in the stomach and duodenal bulb. A diagnosis of GDUC was considered. The patient hesitated about iv corticosteroids, so he was treated with pentasa 3.2 g/d. After 0.5 mo of treatment, the symptoms achieved complete remission. Follow-up examinations showed no evidence of recurrence for 26 mo.

Box 1: Sample clinical case.

2 *Motivation and Related Work*

The main motivation of the project is creating a clinical document corpus that can be freely redistributable and that contains temporal information and clinical entity annotations. The annotation of temporal informa-

[1] https://www.european-language-grid.eu/open-calls/

tion is the core effort of the project, so we expect the E3C corpus to be useful in tasks such as event ordering and chronology generation. These decisions are justified by three common issues on Natural Language Processing (NLP) for the clinical domain, listed below.

First, the E3C corpus is formed of already published documents, mostly clinical cases in journals so as to overcome the patient privacy issues that Electronic Health Records (EHR) often convey. We have opted for selecting documents that already allow redistribution to ensure the E3C corpus will be easily usable for the research community.

Secondly, as E3C is a multilingual corpus, it helps reducing the gap between English and other languages in terms of available data for research. A large dataset of clinical cases in Spanish is already available, SPACCC (Intxaurrondo et al., 2018), and we have expanded and enriched it. For a low-resourced language as Basque, instead, E3C is the first clinical narrative corpus. In addition, the project will also provide the NLP community with the first Creative Commons clinical case corpora for French and Italian.

Thirdly, E3C is centered on temporal information in clinical narratives, which has not been often targeted by scholars. Most of the attention has been focused on clinical entity extraction and classification (Schulz et al., 2020; Grabar et al., 2019; Dreisbach et al., 2019; Luo et al., 2017) and only a few key initiatives have addressed temporal information processing, e.g. the THYME annotation scheme (Styler et al., 2014b). THYME and other off-spins have been used to annotate clinical case corpora such as the i2b2 temporal relation corpus (Sun, Rumshisky, and Uzuner, 2013), and have been used in clinical narratives processing challenges, e.g., CLEF eHealth (Kelly et al., 2019). This information could be then merged with the information on structured data collections, e.g. MIMIC III (Johnson et al., 2016), enriching it.

3 Data Collection and Distribution

The E3C corpus is a collection of both existing corpora (e.g., the SPACCC corpus) and published texts extracted from different sources, such as PubMed[2] (journal abstracts), SciELO[3] and the PanAfrican Medical Journal[4] (clinical cases).

Language	L1	L2	L3
Spanish	81	162	1,772
Basque	88	113	1,232
English	84	171	9,986
French	81	168	9,111
Italian	84	174	10,217

Table 1: Documents in the different layers for each language.

3.1 Corpus Organisation

For each language, the E3C corpus is organized into three layers, with different purposes:

Layer 1: about 25K tokens per language (around 80 documents) of clinical narratives with full manual annotation of clinical entities, temporal information and factuality, for benchmarking and linguistic analysis.

Layer 1 is the core of the E3C corpus and special attention has been awarded to creating a balanced document set in terms of size. Short (<200 tokens), medium (200–400 tokens) and long (400–600 tokens) texts have been selected, as we presumed that text length would directly affect the temporal information in text; the longer the text, the more complex the temporal graph.

Layer 2: 50–100K tokens per language of clinical narratives with automatic annotation of clinical entities and manual check of the annotation of a small sample (about 10%).

Layer 3: about 1M tokens per language of non-annotated medical documents (not necessarily clinical narratives) to be exploited by semi-supervised approaches.

All the layers are covered for Spanish, English, Italian and French. For Basque, Layer 2 (14K token) and 3 (600K token) are covered only partially. In Table 1 we summarize the distribution of documents per layer and language. In the case of L1, the amount of texts provides the information on how many distinct temporal graphs or chronologies we will be able to build from the dataset.

3.2 Corpus Distribution

The final E3C corpus will be available for download from the ELG platform repository[5]. All documents will be released under

[2] https://pubmed.ncbi.nlm.nih.gov/

[3] https://scielo.org/

[4] https://www.panafrican-med-journal.com/

[5] https://live.european-language-grid.eu

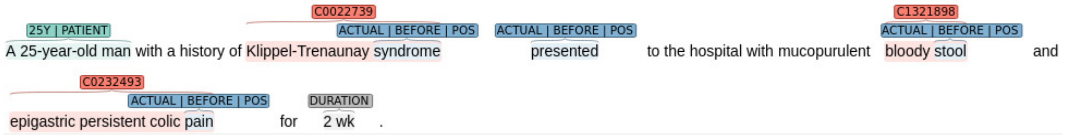
25Y | PATIENT
C0022739
ACTUAL | BEFORE | POS
ACTUAL | BEFORE | POS
C1321898
ACTUAL | BEFORE | POS
A 25-year-old man with a history of Klippel-Trenaunay syndrome presented to the hospital with mucopurulent bloody stool and
C0232493
ACTUAL | BEFORE | POS
DURATION
epigastric persistent colic pain for 2 wk .

Figure 1: A sentence in a clinical case annotated with both temporal information and clinical entities (i.e., disorders) with their UMLS codes (marked in red).

Creative Commons licenses. This is possible as a large part of the texts in the corpus were already released under Creative Commons licenses and as permission for free distribution has been requested and obtained from the original owners of some of the documents.

4 *Corpus Annotation*

The E3C corpus will contain two types of annotations: (i) temporal information and factuality, and (ii) annotation of clinical entities, specifically disorders.

Manual annotation has been performed by a team of NLP students and researchers. More precisely, temporal information and clinical entity annotation guidelines have been defined by two teams of three and four experts respectively. For the manual annotation effort, eight people have been trained and are completing the annotation of temporal information, while clinical entity annotation is being conducted by five people.

4.1 Temporal Information Annotation

Temporal information annotation is performed following the THYME annotation guidelines (Styler et al., 2014a) with some minor adaptations (Magnini et al., 2020). This scheme provides tags for events, time expressions, temporal relations between events and/or time expressions, and aspectual relations between events. For each tag, a set of attribute-value pairs allow to make the relevant features explicit. In order to mark information that further contributes to the clinical history of a patient, we have added three categories: measurements and test results, actors (for the patient itself, health professionals and other participants), body parts.

In Figure 1 a simplified temporal information annotation is displayed. Events are in dark blue and their contextual modality (ACTUAL), document time relation (BEFORE) and polarity (POS) are highlighted. The *2 wk* time expression (in gray) is classified as a duration and the information about the actor (the patient) is represented in light blue.

4.2 Clinical Entity Annotation

Clinical entity annotation focuses on disorders. Following UMLS[6] a disorder is defined as “a definite pathologic process with a characteristic set of signs and symptoms”.

In E3C, we mark disorders mentioned in the text and assign them an UMLS concept unique identifier (CUI). Disorder identification and coding is performed following an adaptation of the ShARe annotation guidelines (Elhadad et al., 2012). In concept selection, we restrict to the UMLS semantic group Disorder, which includes the Finding semantic type in addition to those proposed by ShARe. Figure 1 shows disorders, marked in red, and their UMLS codes.

5 *Conclusions and Future Work*

The E3C project aims at reducing the lack of available resources for clinical NLP, gathering a large number of clinical narratives and focusing on languages other than English. After completing the project, the E3C corpus and the associated resources (baselines, scorers, etc.) will be available for research under a Creative Commons license, which will facilitate its acquisition and reusability. More specifically, since the corpus contains information for temporal reasoning as well as clinical entity mentions, it will be useful for works on semantic interpretation of clinical texts. The fact that the corpus contains texts in five languages will allow linguistic comparisons as well as experimentation on transfer learning. We also consider that the E3C corpus is a resource that could be employed in a series of evaluation challenges due to its atypical contents and types of annotations.

Acknowledgements

This work was partially supported by the European Language Grid project through its

[6] `https://uts.nlm.nih.gov/uts/umls/home`

open call for pilot projects (EU grant no. 825627), and by the Basque Government post-doctoral grant POS_2020_2_0026.

References

[Dreisbach et al.2019] Dreisbach, C., T. A. Koleck, P. E. Bourne, and S. Bakken. 2019. A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data. *International Journal of Medical Informatics*, 125:37–46.

[Elhadad et al.2012] Elhadad, N., G. Savova, W. Chapman, G. Zaramba, D. Harris, and A. Vogel. 2012. ShARe Guidelines for the Annotation of Modifiers for Disorders in Clinical Notes. Technical report, Columbia University.

[Grabar et al.2019] Grabar, N., C. Grouin, T. Hamon, and V. Claveau. 2019. Recherche et extraction d'information dans des cas cliniques. Présentation de la campagne d'évaluation DEFT 2019. In *Actes du Défi Fouille de Textes 2019*, pages 7–16, Toulouse, France. Actes DEFT 2019.

[Intxaurrondo et al.2018] Intxaurrondo, A., M. Marimón, A. González-Agirre, J. A. López-Martín, H. Rodríguez, J. Santamaría, M. Villegas, and M. Krallinger. 2018. Finding Mentions of Abbreviations and Their Definitions in Spanish Clinical Cases: The BARR2 Shared Task Evaluation Results. In *Proceedings of the Third Workshop on Evaluation of Human Language Technologies for Iberian Languages (IberEval 2018)*, pages 280–289, Seville, Spain. Spanish Society for Natural Language Processing.

[Johnson et al.2016] 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.

[Kelly et al.2019] Kelly, L., H. Suominen, L. Goeuriot, M. Neves, E. Kanoulas, D. Li, L. Azzopardi, R. Spijker, G. Zuccon, H. Scells, and J. Palotti. 2019. Overview of the CLEF eHealth Evaluation Lab 2019. In F. Crestani, M. Braschler, J. Savoy, A. Rauber, H. Müller, D. E. Losada, G. Heinatz Bürki, L. Cappellato, and N. Ferro, editors, *Experimental IR Meets Multilinguality, Multimodality, and Interaction*, pages 322–339, Cham. Springer International Publishing.

[Luo et al.2017] Luo, Y., W. K. Thompson, T. M. Herr, Z. Zeng, M. A. Berendsen, S. R. Jonnalagadda, M. B. Carson, and J. Starren. 2017. Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review. *Drug Safety*, 40:1075–1089.

[Magnini et al.2020] Magnini, B., B. Altuna, A. Lavelli, M. Speranza, and R. Zanoli. 2020. The E3C Project: Collection and Annotation of a Multilingual Corpus of Clinical Cases. In *Proceedings of the Seventh Italian Conference on Computational Linguistics*, Bologna, Italy, December. Associazione Italiana di Linguistica Computazionale.

[Schulz et al.2020] Schulz, S., J. Ševa, S. Rodríguez, M. Ostendorff, and G. Rehm. 2020. Named Entities in Medical Case Reports: Corpus and Experiments. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4495–4500, Marseille, France. European Language Resources Association.

[Styler et al.2014a] Styler, W., G. Savova, M. Palmer, J. Pustejovsky, T. O'Gorman, and P. C. deGroen. 2014a. THYME Annotation Guidelines. Technical report, University of Colorado. `http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf`.

[Styler et al.2014b] 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. 2014b. Temporal Annotation in the Clinical Domain. *Transactions of the Association for Computational Linguistics*, 2:143–154.

[Sun, Rumshisky, and Uzuner2013] Sun, W., A. Rumshisky, and O. Uzuner. 2013. Annotating temporal information in clinical narratives. *Journal of Biomedical Informatics*, 46(Supplement):S5–S12. 2012 i2b2 NLP Challenge on Temporal Relations in Clinical Data.