Cognitive Technologies

Georg Rehm *Editor*

European Language Grid

A Language Technology Platform for Multilingual Europe

OPEN ACCESS

Editor
Georg Rehm
Deutsches Forschungszentrum
für Künstliche Intelligenz GmbH (DFKI)
Berlin, Germany

The European Language Grid has received funding from the European Union's Horizon 2020 research and innovation programme under grant agreement no. 825627.

ISSN 1611-2482 ISSN 2197-6635 (electronic)
Cognitive Technologies
ISBN 978-3-031-17257-1 ISBN 978-3-031-17258-8 (eBook)
https://doi.org/10.1007/978-3-031-17258-8

This Springer imprint is published by the registered company Springer Nature Switzerland AG
The registered company address is: Gewerbestrasse 11, 6330 Cham, Switzerland

Chapter 17
European Clinical Case Corpus

Bernardo Magnini, Begoña Altuna, Alberto Lavelli, Anne-Lyse Minard, Manuela Speranza, and Roberto Zanoli

Abstract Interpreting information in medical documents has become one of the most relevant application areas for language technologies. However, despite the fact that huge amounts of medical documents (e. g., medical examination reports, hospital discharge letters, digital medical records) are produced, their availability for research purposes is still limited, due to strict data protection regulations. Aiming at fostering advanced information extraction technologies for medical applications, we present E3C, a corpus of clinical case narratives fully based on freely licensed documents. E3C (European Clinical Case Corpus) contains a vast selection of clinical cases (i. e., narratives presenting a patient's history) that cover different medical areas, are based on different styles and produced in different languages. A portion of the corpus has been manually annotated to be used for training and testing purposes, while a larger set of documents has been automatically tagged to serve as a baseline for future research in information extraction.

1 Overview and Objectives of the Pilot Project

The interest in information extraction from clinical narratives has increased in recent decades, including clinical entity extraction and classification (Schulz et al. 2020; Grabar et al. 2019; Dreisbach et al. 2019; Luo et al. 2017), clinical prediction systems, e. g., MIMIC III (Johnson et al. 2016), and the organisation of challenges at CLEF (Kelly et al. 2019), and Semeval. However, only a few shared datasets have been created, limiting the potential of developing applications in this area.

Bernardo Magnini · Alberto Lavelli · Manuela Speranza · Roberto Zanoli
Fondazione Bruno Kessler, Italy, magnini@fbk.eu, lavelli@fbk.eu, manspera@fbk.eu, zanoli@fbk.eu

Begoña Altuna
Fondazione Bruno Kessler, Italy, HiTZ Centre, University of the Basque Country, Spain, begona.altuna@ehu.eus

Anne-Lyse Minard
Université d'Orléans, France, anne-lyse.minard@univ-orleans.fr

G. Rehm (ed.), *European Language Grid*, Cognitive Technologies,
https://doi.org/10.1007/978-3-031-17258-8_17

We report upon the E3C (European Clinical Case Corpus) ELG pilot project, which resulted in a large collection of clinical cases in five European languages: English, Spanish, French, Italian and Basque. A clinical case is a statement of a clinical practice, presenting the reason for a clinical visit, the description of physical exams, and the assessment of the patient's situation. Clinical cases are typically reported and discussed in research papers, and are often used for education purposes in medicine. In addition, published clinical cases are de-identified, overcoming privacy issues, and are rich in clinical entities as well as temporal information.

> 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. Continuous superficial ulcers and spontaneous bleeding were observed under colonoscopy. Subsequent gastroscopy revealed mucosa with diffuse edema, ulcers, errhysis, and granular and friable changes in the stomach and duodenal bulb, which were similar to the appearance of the rectum. After ruling out other possibilities according to a series of examinations, a diagnosis of GDUC was considered. The patient hesitated about intravenous corticosteroids, so he received a standardized treatment with pentasa of 3.2 g/d. After 0.5 mo of treatment, the patient's symptoms achieved complete remission. Follow-up endoscopy and imaging findings showed no evidence of recurrence for 26 mo.

The sample clinical case reported in the box above is about a patient presenting gastric symptoms, who is finally diagnosed with gastroduodenitis associated with ulcerative colitis (GDUC). To reach the diagnosis, two medical tests (colonoscopy and gastroscopy) were performed. Treatment, outcome (complete remission) and follow-up (no evidence of recurrence) are also present in the text.

2 Corpus Collection and Annotation

The document collection was determined by the available resources for each language (e. g., PubMed, scientific journals, medicine leaflets). First, we identified possible document sources as well as their licenses and re-distribution policies. We selected sources that were either already available under Creative Commons licenses (i. e., CC-BY or CC-BY-SA), possibly asking for re-distribution permission to the right holders. In the case of the SPACCC[1] and NUBes[2] corpora, the texts were ready to be used by us in terms of licensing and formatting. We automated the text collection as much as possible, for example, in some cases we were able to identify and extract the section with the clinical case. All English and some French documents were automatically extracted from PubMed[3], through its API, while medicine leaflets were automatically crawled and stored in a single file for each language. Journal articles with clinical cases that could not be extracted automatically were filtered through the search query "clinical case" in the different languages. In addition to the

[1] https://github.com/PlanTL-GOB-ES/SPACCC

[2] https://github.com/Vicomtech/NUBes-negation-uncertainty-biomedical-corpus

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

extraction of the relevant documents, corresponding metadata was stored to allow accurate documentation.

The annotation of temporal information was performed following an adaptation of the THYME annotation guidelines (Styler et al. 2014).[4] Temporal information refers to the events in a text as well as to chronological references and relations. To encode temporal information, we defined the following tags and relation types. Events, time expressions, temporal relations and aspectual relations are widely used in temporal information tasks, while actor, body part and RML annotations were added as they convey relevant information of the clinical domain.

- *Events* are the events or states relevant to the patient's clinical timeline.
- *Time expressions* refer to points and intervals in time.
- *Temporal relations* (TLINK) implement relations that chronologically order events and time expressions.
- *Aspectual relations* (ALINK) are created between an aspectual event and its subordinated non-aspectual event.
- *Actors* are the people (or animals) mentioned in the text.
- *Body parts* are the parts of the body that are bigger than cells.
- *Results, measurements and lab and test results (RML)* are lab test and analytics' results, formulaic measurements and measurement values.

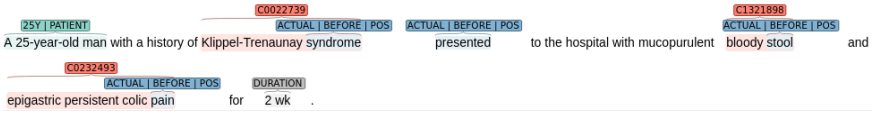

Fig. 1 A sentence in a clinical case annotated with both temporal information and clinical entities (i. e., disorders) with their UMLS codes (marked in red)

The annotation of clinical entities is mainly based on the guidelines of SEMEVAL 2015 Task 14 "Analysis of Clinical Text"[5] and on the ASSESS CT guidelines (Miñarro-Giménez et al. 2018). The annotation of Layer 1 was done fully manually, while for Layer 2 the automatic annotation was produced with a distant supervision method that matches clinical entities with disorder concepts in UMLS.

3 Implementation

The E3C corpus is organised in three different layers:

Layer 1: about 25k tokens per language of clinical narratives with full manual or manually checked annotation of clinical entities, temporal information and factuality, for benchmarking and linguistic analysis.

[4] http://clear.colorado.edu/compsem/documents/THYME_guidelines.pdf

[5] http://alt.qcri.org/semeval2015/task14/data/uploads/share_annotation_guidelines.pdf

Layer 2: 50-100k tokens per language of clinical narratives with automatic annotation of clinical entities. Distant supervision was used to annotate 8,972 clinical entities with their corresponding concepts in UMLS.

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

Table 1 shows the sizes of the layers (document and token numbers). Table 2 shows the numbers of Layer 1 tags to indicate information density in clinical cases.

	English	French	Italian	Spanish	Basque
Layer 1	84 / 25142	81 /25196	86 / 24319	81 / 24681	90 / 22505
Layer 2	171 / 50371	168 / 50490	174 / 49900	162 / 49351	111 / 12541
Layer 3	9779 / 1075709	25740 / 66281501	10213 / 13601915	1876 / 1030907	1232 / 518244

Table 1 Documents/tokens in each language and layer in the E3C corpus.

Entity	English	French	Italian	Spanish	Basque
CLINENTITY	1024	1327	869	1345	1910
EVENT	4885	4312	3385	4767	7910
ACTOR	682	427	338	319	505
BODYPART	968	659	328	814	1410
TIMEX3	380	333	298	383	638
RML	480	508	383	391	1101
ALINK	114	71	109	92	113
TLINK	4852	4084	1150	4700	7981

Table 2 Annotations in each language in Layer 1 in the E3C corpus.

4 Evaluation

For temporal information and clinical entity annotation tasks, we performed inter-annotator agreement (IAA) tests. We measured whether the guidelines had been defined and were understood correctly, and we ensured that the quality of annotations in the corpus was similar. The IAA phase had been done on the English part of the corpus. IAA for temporal entities (EVENT, TIMEX3, ACTOR, BODYPART) was measured using three annotators and six documents. To compute the agreement, we used the F1-measure metric, which produced the same results as using the Dice coefficient. The agreement is high for EVENT and ACTOR entities (with an average of 0.81 and 0.87), but a bit lower for TIMEX3 and BODYPART (with an average of 0.50 and 0.57). The IAA for temporal relations (TLINK) was split in two phases: three documents were annotated, the results discussed by the annotators and

then three new documents were annotated. To measure the agreement, we used the Tempeval-3 scorer (UzZaman and Allen 2011), implemented for the evaluation of systems based on the comparison of temporal graphs built from annotations. The average F1-measure for the first phase was 0.43 and 0.53 for the second.

The annotation of the clinical entities in Layer 1 was performed by four annotators. Again, the agreement is calculated using F1, whereas for the CUI attribute we computed the accuracy taking into consideration only the entities identified by two annotators. The agreement for clinical entity recognition is 0.70 on average (from 0.64 to 0.78). In the entity linking task, the accuracy on entities identified by both annotators starts at 0.86 (on average 0.89).

The clinical entities in Layer 2 were annotated automatically using distant supervision and UMLS as a controlled vocabulary. A manual assessment of the quality of these annotated entities would be too demanding in terms of human resources. For this reason, the quality of Layer 2 has been estimated through an indirect evaluation that uses the results obtained by distant supervision on Layer 1 (Table 3) as an estimation of the quality of the Layer 2 annotations. This approximation is possible because the documents in Layer 1 and Layer 2 are clinical cases and because they were extracted from the same kind of publications or from the same existing corpora.

	English	French	Italian	Spanish	Basque
Accuracy	48.33	54.92	58.09	63.64	55.35

Table 3 Estimated accuracy (F_1-measure) of the clinical entities in Layer 2.

5 Conclusions and Results of the Pilot Project

The E3C pilot project aims at fostering advanced information extraction technologies for medical applications. Results include a large corpus of annotated clinical cases in five languages. The corpus is available on the ELG platform.

Acknowledgements The work described in this article has received funding from the EU project European Language Grid as one of its pilot projects and from the Basque Government post-doctoral grant POS_2020_2_0026.

References

Dreisbach, Caitlin, Theresa A. Koleck, Philip E. Bourne, and Suzanne Bakken (2019). "A systematic review of natural language processing and text mining of symptoms from electronic patient-authored text data". In: *Int. Jour. of Medical Informatics* 125, pp. 37–46. DOI: 10.1016/j.ijmedinf.2019.02.008.

Grabar, Natalia, Cyril Grouin, Thierry Hamon, and Vincent 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*. Toulouse, France: Actes DEFT 2019, pp. 7–16. URL: https://www.irit.fr/pfia2019/wp-content/uploads/2019/07/actes_DEFT_CH_PFIA2019.pdf.

Johnson, Alistair E.W., Tom J. Pollard, Lu Shen, Li-wei H. Lehman, Mengling Feng, Mohammad Ghassemi, Benjamin Moody, Peter Szolovits, Leo Anthony Celi, and Roger G. Mark (2016). "MIMIC-III, a freely accessible critical care database". In: *Scientific Data* 3. DOI: 10.1038/sdata.2016.35.

Kelly, Liadh, Hanna Suominen, Lorraine Goeuriot, Mariana Neves, Evangelos Kanoulas, Dan Li, Leif Azzopardi, Rene Spijker, Guido Zuccon, Harrisen Scells, and João Palotti (2019). "Overview of the CLEF eHealth Evaluation Lab 2019". In: *Experimental IR Meets Multilinguality, Multimodality, and Interaction*. Ed. by Fabio Crestani, Martin Braschler, Jacques Savoy, Andreas Rauber, Henning Müller, David E. Losada, Gundula Heinatz Bürki, Linda Cappellato, and Nicola Ferro. Cham: Springer, pp. 322–339.

Luo, Yuan, William K. Thompson, Timothy M. Herr, Zexian Zeng, Mark A. Berendsen, Siddhartha R. Jonnalagadda, Matthew B. Carson, and Justin Starren (2017). "Natural Language Processing for EHR-Based Pharmacovigilance: A Structured Review". In: *Drug Safety* 40 (11), pp. 1075–1089. DOI: 10.1007/s40264-017-0558-6.

Miñarro-Giménez, José Antonio, Catalina Martínez-Costa, Daniel Karlsson, Stefan Schulz, and Kirstine Rosenbeck Gøeg (2018). "Qualitative analysis of manual annotations of clinical text with SNOMED CT". In: *PLoS ONE* 13.12. URL: https://www.ncbi.nlm.nih.gov/pmc/articles/PMC6307753/pdf/pone.0209547.pdf.

Schulz, Sarah, Jurica Ševa, Samuel Rodríguez, Malte Ostendorff, and Georg Rehm (2020). "Named Entities in Medical Case Reports: Corpus and Experiments". In: *Proceedings of the 12th Language Resources and Evaluation Conference*. Marseille, France: ELRA, pp. 4495–4500. URL: https://www.aclweb.org/anthology/2020.lrec-1.553.

Styler, William F., Steven Bethard, Sean Finan, Martha Palmer, Sameer Pradhan, Piet C. de Groen, Brad Erickson, Timothy Miller, Chen Lin, Guergana Savova, et al. (2014). "Temporal Annotation in the Clinical Domain". In: *Transactions of the Association for Computational Linguistics* 2. Ed. by Ellen Riloff, pp. 143–154. URL: http://aclweb.org/anthology/Q14-1012.

UzZaman, Naushad and James Allen (2011). "Temporal Evaluation". In: *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies*. Portland, Oregon, USA: ACL, pp. 351–356. URL: https://aclanthology.org/P11-2061.