A Clinical Trials Corpus with UMLS Entities was produced in Order to Improve Medicine

Nortehz Garipert*

Department of life science, Stellenbosch University, South Africa

Abstract

Healthcare workers find it difficult to keep up with the latest studies that support Evidence-Based Medicine due to the huge volume of medical literature. Natural language processing makes it easier to find relevant information, and gold standard corpora are needed to make systems better. We gathered the Clinical Trials for Evidence-Based Medicine in Spanish (CT-EBM-SP) corpus to contribute a fresh dataset for this domain. Anatomy (ANAT), pharmacological and chemical compounds (CHEM), diseases (DISO), and lab tests, diagnostic or therapeutic procedures were used to annotate clinical trials (PROC). We used F-measure to measure inter-annotator agreement (IAA) after we double-annotated the corpus. As an example, we use neural network models to perform medical entity recognition investigations. Our findings suggest that this resource is sufficient for trials using cutting-edge biomedical named entity recognition techniques.

Keywords

ANAT, PROC, IAA, CHEM, Clinical trials, Medicine

Correspondence to:

Nortehz Garipert
Department of life science,
Stellenbosch University, South Africa,
Email: ngaripert.su@su.za

Citation: Garipert N (2022). A Clinical Trials Corpus with UMLS Entities was produced in Order to Improve Medicine. EJBI. 18(1):3-4.

DOI: 10.24105/ejbi.2022.18.1.3-4

Received: 05-Jan-2022, Manuscript No. ejbi-22-53329;
Editor assigned: 06-Jan-2022, PreQC No. ejbi-22-53329(PQ);
Reviewed: 20-Jan-2022, QC No. ejbi-22-53329;
Revised: 24-Jan-2022, Manuscript No. ejbi-22-53329(R);
Published: 31-Jan-2022

1. Introduction

The goal of this project is to provide the first annotated collection of materials in the Spanish language about clinical research and trial announcements. The purpose of this resource is to undertake medical NER experiments and build systems that address the challenges listed. Clinical trial and retrospective study abstracts from PubMed and the SciELO repository, as well as clinical trial notifications from EudraCT, have been annotated. Pathologies (DISO), anatomic entities (ANAT), biochemical or pharmacological substances (CHEM), and diagnostic or therapeutic procedures and lab tests are the four semantic areas covered by the Unified Medical Language System (UMLS) (PROC) [1].

As a proof-of-concept, we concentrated on those four entity categories to see if the annotation and named entity recognition tasks on these data produced satisfactory results. The results of the studies presented here suggest that the annotation scheme and methodology were adequate. The present resource is open to the research community for no charge. Furthermore, the methodology can be applied to other languages that have similar data sources. This article begins with a review of the literature before going over the methodologies, which include text selection and sources, annotation process and scheme, content analysis, inter-annotator agreement assessment, and use case experiments. The results are then presented, including the number of texts and annotations, therapeutic areas covered inter-annotator agreement, and experimental findings. Before coming to a conclusion, we discuss our findings. The topics of this paper are summarized in a supplementary graphical abstract. The PICO model is a frequently used framework for formalizing clinical trial data: a population or group of patients (P) with a medical problem undergoes an experimental intervention involving a standard therapy or comparator (C), in the hopes of improving outcomes (O). Corpora for named entity recognition, on the other hand, include entities annotated with both PICO and other domain labels (e.g. diseases or drugs) [2].

NICTA-PIBOSO, a collection of many biomedical abstracts, is one of the earliest annotated corpora of evidence-based texts. Sentences were manually labelled with PIBOSO elements, similar to the work reported in (Population, Intervention, Background, Outcome, Study Design, and Other). The dataset was utilised to conduct experiments aimed at identifying essential sentences and putting machine learning NER models to the test [3]. More than matched with PubMed articles about RCTs are collected in the Evidence Inference corpus. The prompts and the evidence-supporting texts were matched by medical doctors. They also documented the relationship between the Intervention, the Comparator, and the Outcomes: the results might show a considerable rise or reduction in comparison to the comparator, or no significant difference at all. The dataset was utilised in evidence inference machine learning research.

Finally, the Chia corpus compiles annotations on patient eligibility requirements from clinical trials involving a variety of diseases.

Two medical experts created executable queries by annotating entities and relationships, which can be represented as annotation graphs. To the best of our knowledge, other teams have annotated eligibility requirements as well, but this is the most comprehensive freely available resource. The corpus was built for electronic phenotyping and information extraction experiments. The Drug Semantics corpus is a set of product-specific summaries (SPCs) [4]. In fields where publically available data is rare, a text selection approach is essential for constructing a corpus of sufficient size and generalizability. Large datasets may sufficient if enough sources are accessible; nevertheless, experiments in the medical arena have already proven that larger datasets do not always give superior findings. This is why we used the KL distance on the semantic annotations and lexical similarity to choose texts based on their similar length or semantic content. These strategies are complementary and more appropriate for our purpose than other options such as picking texts based on the demographics of the writers or the publication channel [5].

2. Conclusion

The CT-EBM-SP corpus, which contains approximately clinical trials studies and announcements in Spanish, was created using the methods outlined here. This is the first resource in this language dedicated to medical natural language processing of clinical trials. We demonstrated that using the current version of the CT-EBM-SP corpus, we were able to evaluate state-of-the-art neural biological named entity recognizers with competitive results. The methods described here can be applied to other languages with similar sources, such as English, French, or German.

3. References

1. Uzuner O, Solti I, Cadag E. Extracting medication information from clinical text. J Am Med Inform Assoc. 2010; 17(5):514-518.
2. Sun W, Rumshisky A, Uzuner O. Evaluating temporal relations in clinical text: 2012 i2b2 challenge. J Am Med Inform Assoc. 2013; 20(5):806-813.
3. Kim JD, Ohta T, Tsujii J. Corpus annotation for mining biomedical events from literature. BMC Bioinform. 2008; 9(1):10.
4. Bada M, Eckert M, Evans D, Garcia K, Shipley K, Sitnikov D, et al. Concept annotation in the CRAFT corpus. BMC Bioinform. 2012; 13(1):161.
5. Ury F, Butler A, Yuan C, Fu Lh, Sun Y, Liu H, et al. Chia, a large annotated corpus of clinical trial eligibility criteria. Sci Data. 2020; 7(1):1-11.