Creating Multilingual Gold Standard Corpora for Biomedical Concept Recognition

Jan A. Kors[1], Simon Clematide[2], Saber A. Akhondi[1], Erik M. van Mulligen[1], and Dietrich Rebholz-Schuhmann[2]

[1] Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands
{j.kors,s.ahmadakhondi,e.vanmulligen}@erasmusmc.nl
[2] Institute of Computational Linguistics, University of Zurich, Zurich, Switzerland
{clematide,rebholz}@ifi.uzh.ch

Abstract. We describe our approach to create gold standard corpora for biomedical concept recognition in multiple languages, including English, French, German, Spanish, and Dutch. The annotations are based on a subset of the Unified Medical Language System and cover a wide variety of semantic groups.

Keywords: gold standard corpus, multilingual, concept annotation

1 Introduction

Gold standard corpora (GSCs) are essential for the development and evaluation of systems that perform natural language processing tasks. Currently available GSCs are only in English, often contain annotations for a limited set of semantic types, and generally do not link the annotations to ontological information.

In the MANTRA project (http://www.mantra-project.eu), community efforts are solicited to provide two types of resources: enriched multilingual biomedical terminologies and semantically annotated multilingual documents for a wide range of semantic types. To achieve these goals, the MANTRA project capitalizes on a variety of existing parallel corpora and terminologies. The corpora include multilingual titles of scientific abstracts, drug labels, and biomedical patents. The terminologies are drawn from the Unified Medical Language System (UMLS). The quality of the newly generated resources has to be evaluated on multilingual GSCs in which entity mentions of different semantic types are mapped to unique concept identifiers. In this paper, we describe our approach to construct such GSCs and report initial results.

2 Methods

2.1 Corpora

The GCSs are based on three multilingual corpora that have been collected in the MANTRA project: abstract titles from Medline, drug labels from the European Medicines Agency (EMEA) (freely available through the OPUS collection, http://opus.lingfil.uu.se/EMEA.php), and patents in the biomedical domain from IFI Claims (http://ificlaims.com). The languages of interest in the MANTRA project include English, German, French, Spanish, and Dutch. The Medline titles are bilingual, always in English and one of the other languages. The EMEA labels are available in all languages, the patents only in English, German, and French. Each document in the MANTRA corpora consists of one or more units of text, where a unit may contain a title (Medline abstracts), sentence (EMEA labels), or a paragraph of text (patents). From each MANTRA corpus, units were randomly selected for constructing a GSC: 100 units from the EMEA labels, 100 units from each set of bilingual Medline titles (400 units in total), and 50 units from the patents. Another 20 English units (11 titles, 5 labels, 4 patents) were selected for the development of annotation guidelines.

2.2 Terminology

The annotators had to make their annotations based on the terminology that is used in the MANTRA project. The MANTRA terminology contains a subset of the UMLS, including MeSH, MedDRA, and SNOMED-CT. For each concept in these terminologies, all terms were culled together with their semantic type and concept unique identifier (CUI). Concepts were included if their semantic type belonged to one of the following semantic groups [1]: anatomy, chemicals and drugs, devices, disorders, geographic areas, living beings, objects, phenomena, and physiology.

2.3 Annotation Process

The annotations are made independently by at least three annotators, using the brat rapid annotation tool [2]. The annotation process consists of the following steps:

1. For each unit, pre-annotations are provided based on the annotations made by the concept recognition systems participating in the MANTRA project. A pre-annotation consists of the span of text corresponding with the concept, and its preferred name, semantic type, semantic group, and CUI (all based on the MANTRA terminology).
2. All English units are annotated. Annotators have to correct the pre-annotations if they are wrong, and add annotations that were missed by the systems. To find further information on a marked span of text in brat (pre-annotated or marked by the annotator), annotators can easily link out to the UMLS Terminology Services (https://uts.nlm.nih.gov/home.html) or to the Mantra terminology.
3. The English GSCs are established by harmonizing the individual annotations. For harmonization we use the e-centroid method, an extension of the centroid method that was developed in the CALBC project [3].
4. The non-English units are annotated. For each unit, the annotators are provided with the pre-annotations and with the gold-standard annotations of the corresponding English unit. This should make the concept recognition and annotation in the non-English units less demanding for the annotators.
5. The non-English GSCs are established using the same approach as for the English GSCs.

2.4 Annotation Guidelines

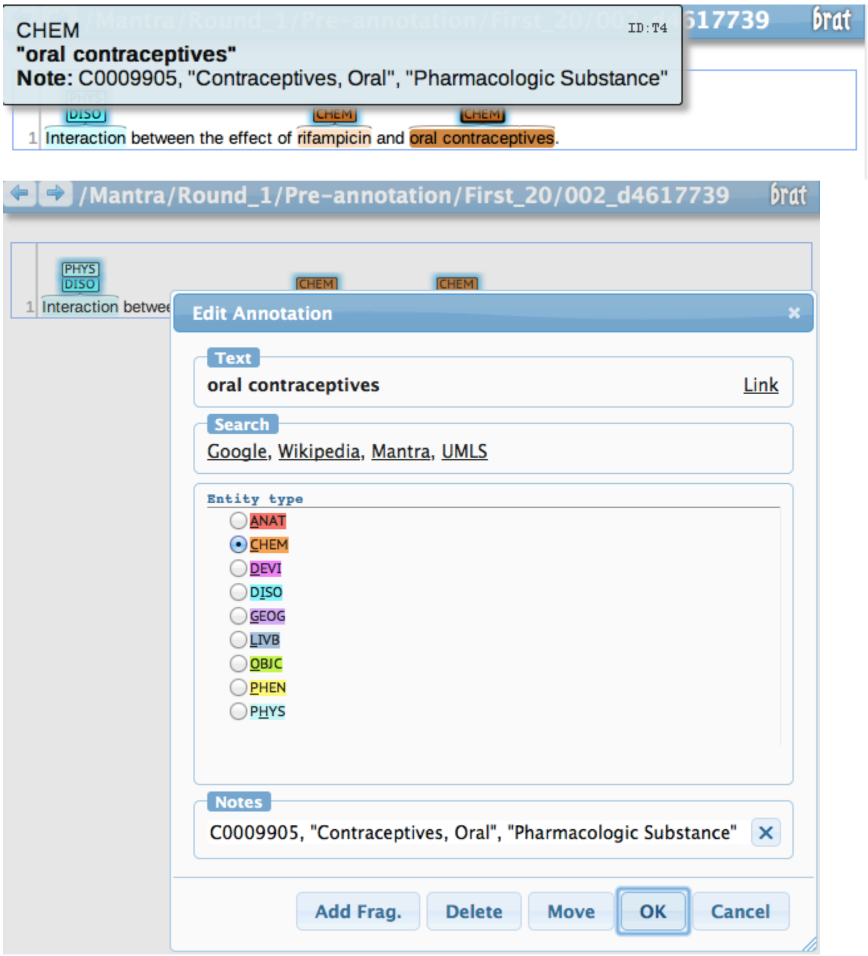
Annotation guidelines were established based on the 20 units that were selected for development purposes. In case of multiple pre-annotations of the same stretch of text, the annotators should try to disambiguate. If the difference in meaning between the concepts is not clear or the context provides insufficient information to disambiguate, all annotations are kept. When an entity is nested within another entity, only the most detailed description of the entity is annotated. The general principle is to annotate the entity that is more specific and informative. Only concepts that are part of the MANTRA terminology should be annotated.

2.5 Inter-Annotator Agreement

Inter-annotator agreement was measured by the F-score (harmonic mean of recall and precision) between two annotators or between one annotator and the gold standard. Note that the F-score is invariant to who of the two annotators is taken as the reference when computing precision and recall.

3 Results

Fig. 1 shows two screen shots of the brat annotation tool for one of the English units. Information on the (pre-)annotated concepts is shown when the cursor is hovered over the annotations. Double-clicking a word or phrase shows a window that allows to make modifications or to link out to further information.



CHEM ID:T4
"oral contraceptives"
Note: C0009905, "Contraceptives, Oral", "Pharmacologic Substance"

1 Interaction between the effect of rifampicin and oral contraceptives.

/Mantra/Round_1/Pre-annotation/First_20/002_d4617739

1 Interaction betwee

Edit Annotation

Text: oral contraceptives

Search: Google, Wikipedia, Mantra, UMLS

Entity type:
○ ANAT
● CHEM
○ DEVI
○ DISO
○ GEOG
○ LIVB
○ OBJC
○ PHEN
○ PHYS

Notes: C0009905, "Contraceptives, Oral", "Pharmacologic Substance"

Fig. 1. Example of a Medline title with pre-annotated concepts, color-coded by semantic group. When the cursor is hovered over an annotation, the corresponding CUI, preferred term, semantic type and semantic group are shown (upper screen). When a text is double-clicked, a pop-up window appears to edit the annotation or to link out to other resources, such as the UMLS Technology Services (lower screen).

The annotation guidelines were developed based on the annotations of the 20 English training units by three independent annotators and subsequent discussions. A harmonized annotation for this set was automatically constructed, and inter-annotator agreement scores were computed (Table 1). The F-scores indicate good to excellent agreement between annotators 1 and 3 and the harmonized set.

The annotation of the 550 units in the full GSC has been started and is still work in progress.

Table 1. Agreement (F-score) between three annotators and the harmonized set (H) on annotations in 20 English training units.

Annotator	2	3	H
1	.65	.83	.93
2		.58	.70
3			.88

4 Discussion

We described our approach to create multilingual GSCs for biomedical concept recognition. First steps have been taken, including the development of annotation guidelines and a flexible annotation environment, and the selection of multilingual text units from different document types. Inter-annotator agreement scores on a small development set suggest that the annotations of different annotators are in good agreement.

To our knowledge, this is the first attempt to create GSCs for biomedical concept recognition in languages other than English. Other distinguishing features are the wide variety of semantic groups that are being covered, and the diverse text genres from which units have to be annotated.

The creation of the GSCs is currently under development. Once available, the GSCs will be made publicly available.

References

1. McCray, A.T., Burgun, A., Bodenreider, O.: Aggregating UMLS Semantic Types for Reducing Conceptual Complexity. Stud. Health Technol. Inform. 10, 216-220 (2001)
2. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: Brat: a Web-based Tool for NLP-assisted Text Annotation. In: Proceedings of the Demonstrations Session at EACL 2012, pp. 103-107. Association for Computational Linguistics (2012)
3. Lewin, I., Kafkas, S., Rebholz-Schuhmann, D.: Centroids: Gold Standards with Distributional Variation. In: Proceedings of the Eighth International Conference on Language Resources and Evaluation (LREC-2012), pp. 3894-3900. European Language Resources Association (2012)