Multilingual semantic resources and parallel corpora in the biomedical domain: the CLEF-ER challenge

Dietrich Rebholz-Schuhmann[1,2], Simon Clematide[1], Fabio Rinaldi[1], Senay Kafkas[2], Erik M. van Mulligen[3], Chinh Bui[3], Johannes Hellrich[4], Ian Lewin[5], David Milward[5], Michael Poprat[6], Antonio Jimeno-Yepes[7], Udo Hahn[4], and Jan A. Kors[3].

(1) Department of Computational Linguistics, University of Zürich, Ch
(rebholz, clematide, rinaldi)@ifi.uzh.ch
(2) European Bioinformatics Institute, Wellcome Trust Genome Campus, Hinxton, Cambridge, CB10 1SD, U.K.
(3) Department of Medical Informatics, Erasmus University Medical Center, Rotterdam
(4) Universität Jena, Fürstengraben 30, D-07743 Jena
(5) Linguamatics Ltd, 324 Science Park, Milton Road, Cambridge CB4 0WG
(6) Averbis GmbH, Tennenbacher Strasse 11, D-79106 Freiburg
(7) National ICT Australia

Abstract. Multilingual terminological resources can be drawn from parallel corpora in the languages of interest, possibly exploiting machine translation solutions for term identification. This main objective of the CLEF-ER challenge involves parallel corpora in English and other languages.
The challenge organisers have gathered and normalized documents from the biomedical domain: titles from scientific articles, drug labels from the European Medicines Agency, and patent texts from the European Patent Office. The parallel units have been identified, marked-up and formatted for future use. The three different corpora show comparable sizes.
In preparation of the CLEF-ER challenge, the documents have been annotated with terminologies in English and non-English languages (de, fr, es, and nl) and the pre-existing terminological resource has been optimized for the entity recognition task in CLEF-ER. Finally a silver standard corpus for entity annotations and their identifiers has been produced on the English documents for the evaluation of challenge contributions.

Motivation

Biomedical IT solutions require terminological resources (TRs) to achieve interoperability of modules and data. Increasingly such IT solutions require multilingual TRs, since they are used in different countries to capture and encode patient related information in the home language. To this end, the biomedical terminologies have to be produced in different languages and entities have to

be identifiable across languages, i.e. the concepts should carry the same identifier, to enable their reuse and to improve the exchange of data across cultural and language borders. However, existing biomedical multilingual TRs are too limited in their size and their development has to be supported by automatic means that analyse available document resources for the acquisition of novel biomedical terms.

The production of multilingual TRs is time-consuming and requires novel approaches to produce them at a large scale. One way to improve the development is the exploitation of multilingual parallel documents ("corpora") and to use automatic annotation and alignment methods to identify relevant terms. Such multilingual documents are available from the European Patent Office[1], EMEA[2] and the Medline[3] distribution. Although they cover different topics, all of them contain potentially novel terms.

The documents from the available corpora can be annotated with automatic means for the identification of entities in the different languages and subsequently mined for novel terms. When different annotations have been provided, computational methods have to be applied to align the annotations in a single corpus (called "silver standard corpus", SSC). Under the provision of the SSC, we achieve two goals: (1) we can evaluate the results from other annotation solutions against the SSC, and (2) we can attribute the concept unique identifiers (CUIs) from the English annotated documents to the non-English documents. No gold standard corpus (GSC) is available for the assessment of the correct annotation of concept identifiers to the terms in the parallel corpora.

After the term candidates have been collected – including their CUIs – they have to be validate and integrated into the existing TR, i.e. the novel terms have to be aligned with the existing state of the art TRs.

Background

The main terminological resources in the biomedical domain are maintained and partially produced by the National Library of Medicine (NLM). In addition, the NLM provides sophisticated resources for the processing of natural language text such as MetaMap, a text categorizer that attributes CUIs to text passages. The following terminological resources - amongst others - play an important role in the biomedical domain: The Medical Subject Headings (MeSH) have been produced to categorize the abstracts in PubMed in such a way that information retrieval is performed at high efficiency. The headings are assigned manually to text documents, but computer programmes support the assignment by attributing also headings by automatic means (see MetaMap above). The Systematized Nomenclature of Human and Veterinary Medicine (SNOMED-CT) resource is a medical terminology for the encoding of diseases. It forms a standard that is well established throughout health organisations in different countries and has

[1] `http://www.epo.org/`
[2] `http://opus.lingfil.uu.se/EMEA.php`
[3] `http://www.ncbi.nlm.nih.gov/pubmed/`

been included into the distribution of Unified Medical Language System (UMLS) under specific licensing requirements. The Medical Dictionary for Regulatory Activities (MedDRA) is a terminological resource that enables encoding of adverse drug events for regulatory affairs.

The analysis of clinical and biomedical documents requires tools and resources for efficient processing. Advanced technologies have been developed in recent years that enable semantic annotations, e.g. entity recognitions of a large number of biomedical entities, as well as the efficient and precise parsing of the literature. The research has focused on the development of solutions for the bioinformatics research community, but increasingly solutions from the research domain are used in the clinical environment as well.

Clinical data is best described by the standardized reporting of clinical parameters, for example the measurements of physiological and patho-physiological parameters from a sample of blood, and of phenotypic parameters given in the natural language of the country. It remains a challenge to read the notes of the clinical doctor, but efficient solutions have been developed to transform the patient record into a standardized representation for further exploitation.

A number of challenges have been introduced for the assessment of tools and solutions that process biomedical text and normalize identified facts and entities. In the domain of molecular biology, the main focus resided on the correct identification of entities such as genes, proteins, drugs and diseases, and after that, the extraction of molecular interactions, gene regulatory events and gene-disease associations. The sequel of BioCreAtIve challenges engaged the research community into several challenges with different characteristics. As an alternative, the BioNLP Shared Task tackled similar problems, and from early on supported the exploitation of ontological resources for the challenges. All the challenges provide manually annotated corpora as a gold standard corpus to measure performances of the annotation solutions. As an alternative to this general paradigm, the CALBC challenge made use of a corpus at a very large scale that has been generated with automatic means from existing annotation solutions, and it thus represented a "silver-standard" corpus.

The "Conference and Labs of the Evaluation Forum" (CLEF) has been formed to tackle challenges and their evaluation as a joint effort. It is organized on a yearly basis and has over and over again suggested new challenges for the research community, including the analysis of patents for improved information retrieval (CLEF-IP), the analysis of medical records (CLEFeHealth) and even the analysis of multilingual and multimodal data.

The CLEF-ER challenge tackles a combination of different tasks: (1) entity mention annotation, (2) entity normalisation, and (3) also multilingual analysis in the sense that participants could use a resource in English to annotate the non-English parallel document. Furthermore, the challenge is not only tuned to a single corpus, but includes patent texts and scientific medical texts alike, provides a reference terminological resource, and demands to process a large-scale corpus, larger than typically available in challenges that make use of a gold standard corpus.

Material and Method

The CLEF-ER challenge is focused to the languages English (en), Spanish (es), French (fr), German (de), and Dutch (nl). This selection has been motivated by the availability of resources, i.e. terminologies and documents alike, in the different languages. It was an important requirement that documents in a non-English language must have a parallel document in English, and – at the same time – it was not relevant that a pair of documents in English and a non-English language should be accompanied by yet another document in a third language.

Terminologies

A number of resources have been prepared for the CLEF-ER challenge: terminological resources and documents. The terminological resource is based on the UMLS and makes use of the contained terms, the standardized file formats and the licensing server of the National Center of Bioinformatics (NCBI) for access to the terminological resource.

In principle, the full set of UMLS terms could be relevant for the annotation of the multilingual documents, but a number of constraints have to be considered. First, not all terms are relevant for the documents that have been prepared for the CLEF-ER challenge. Second, the overhead for the processing of the full set of terms may be high and may distract from the challenge tasks; in other terms, it is advantageous for challenge participants to handle only a reduced set of the terms relevant for the challenge. Third, the evaluation of the results can be improved by reducing the term set, since the excluded terms and categories will reduce ambiguities, i.e. less semantic categories (called "semantic types" and "semantic groups") could be distinguished if a term is polysemous with regards to the semantic categories.

Table 1. (Terminological resource): The English part of the TR contains most terms. Only Spanish is covered in SNOMED-CT. MedDRA terms have been translated in all languages.

Terms	MeSH	SNOMED-CT	MedDRA
en	764,000	1,184,005	56,061
de	77,249	-	50,128
fr	105,758	-	49,586
es	59,678	1,089,723	49,499
nl	40,808	-	50,932

The TR is required for the term normalization: the contained English and non-English concepts are provided with their CUIs. The CUIs form the key result as part of the annotation task in the CLEF-ER challenge, since the challenge

participants have to annotate the text with the CUIs, and the challenge organizers evaluate the anntotations against a silver standard corpus (SSC). This subselection of the UMLS terminological resource is called the MANTRA terminological resource (MTR). The terminology is delivered in the OBO file format, which has been proposed by the Open Biomedical Ontology (OBO) Foundry and is maintained by the National Center of Biomedical Ontologies. Its aim is to create a common format for controlled vocabularies.

The UMLS licence agreement requires that users validate their licenses when accessing the TR. This task is performed by querying the license server from the NLM with the right credentials, for example using restful services through a server[4] validating the username and the password of the licensee. The Terminological Resource is accessible through the download site at the Erasmus University Medical Center Rotterdam.

Selection of parallel corpora

The corpora have been selected from different resources and for the purpose to serve two objectives: (1) enabling extraction of novel terms for the multilingual TR as part of the MANTRA project work and the outcomes from the CLEF-ER challenge, and (2) offering parallel corpora to the CLEF-ER challenge participants as an input to solve term recognition tasks with and without machine-translation solutions.

A number of requirements have to be met to fulfill the given objectives. The corpora have to cover the domain knowledge that is under investigation. This is the case for the scientific literature, but also for documents that deal with drug labels. In addition, the MANTRA project partners selected patent documents from the European Patent Office, since this distribution of documents covers a significant amount of documents, deal with biomedical domain knowledge, but also differ in their language from the above mentioned types of documents.

The diversity between the document repositories is high with regards to the amount of available content, the type of the documents and the languages that are supported in the repository. From the patent corpus, mainly the claim section is available as parallel document, and from Medline only the titles of the scientific articles can be aligned across languages. The EMEA repository allows to identify full documents in parallel. For the patent claims we could produce parallel documents for English, German and French, whereas EMEA and Medline covers all selected languages. The EMEA corpus has already been exploited to train and test statistical machine translation solutions.

Note that Medline abstracts have always been translated from the non-English language into English and as a consequence the parallel units, i.e. the titles, are restricted to language pairs between the English and the non-English language. More in detail, all non-English units, i.e. patent claim sections, Medline titles, or EMEA document, have a parallel unit in English. Also, every English

[4] `https://uts-ws.nlm.nih.gov/restful/isValidUMLSUser?licenseCode=...`

unit has a parallel non-English unit, but only a smaller portion of English units has a parallel non-English unit in two, three or four languages.

Optimizing the TR for the challenge

The MANTRA project partners have annotated the documents with their in-house annotation solutions, e.g. OntoGene, Peregrine, UIMA-based annotation solutions from the JulieLab and from Averbis. All annotation solutions apply the provided UMLS terminological resource. After this phase, the English annotated corpora have been harmonized and evaluated. This included the assessment on the number of annotations that resulted from the different semantic categories in the UMLS. From the distribution of the annotations, the MANTRA project partners took the decision to reduce the number of categories in the UMLS terminological resource and to extract those categories from UMLS that have higher relevance to the multilingual terminological resource and show - in addition - good coverage in the annotated corpora. The final solution is called the MTR and covers the semantic groups: anatomy, chemicals and drugs, devices, disorders, geographic areas, living beings, objects, phenomena and physiology (ANAT, CHEM, DEVI, DISO, GEOG, LIVB, OBJC, PHEN, PHYS). All terms that are categorized with other semantic groups have then been removed to produce the final terminological resources for the CLEF-ER challenge.

Preparation of the silver standard corpus

The annotated corpora in English have been processed to produce the silver standard corpus as describe in.

Resource and evaluation

The following resources have been made available to the challenge participants: terminologies and corpora.

Table 2. (Units and word counts, all corpora): The number of units (and words) is highest in English for Medline. German and French are evenly well covered in all three corpora, and Spanish shows similar coverage, except that Spanish (and Dutch) are not represented for patent texts.

Units	EMEA	Medline	Patent	Words	EMEA	Medline	Patent
en	364,005	1,593,546	120,638	en	5,120,067	15,775,814	6,034,104
de	364,005	719,232	120,637	de	4,571,203	5,996,504	5,194,032
fr	373,152	572,176	120,636	fr	5,515,157	6,023,945	6,689,812
es	366,769	247,655		es	5,897,467	2,573,056	
nl	360,418	54,483		nl	5,130,890	435,390	

Statistics across the corpora

The corpora differ in their quality and also in the type of units that have been identified from the documents, i.e. Medline titles in contrast to paragraphs in EMEA documents and claim sections in the patents. On the other side, all corpora are sufficiently large to support the needs of the CLEF-ER: the corpora pose a challenge to the participants since the annotation of the corpora requires automatic means, and also a challenge to the MANTRA project partners, since the harmonization of the annotations should deliver a valuable resource to the public for future exploitation (as in the CALBC challenge).

The corpora also differ in their sizes (see tbl. 2). The Medline titles form the biggest corpus regarding the number of units and the number of words (as well as the number of characters). The EMEA corpus appears to be larger than the patent corpus when considering the number of units, but is evenly large when measuring the size based on words and characters. Since the EMEA corpus is available in all languages, it is well suited for the CLEF-ER challenge.

The units from the patent corpus are available in English, French and German as stated before, whereas the other two corpora are provided in all languages. In addition, the annotation of the patent corpus has shown that the language in the corpus has a high diversity and the terms from the MTR are often used in a non-specific way with regards to the biomedical purpose of the MTR, which makes the patent corpus less suitable for the entity recognition task.

By contrast, the pairwise units from the Medline titles show high heterogeneity and at the same time, the use of the terminology in Medline is most specific with regards to the purpose of the MeSH, MedDRA, and SNOMED-CT terms in the MTR.

Evaluation

The partners have contributed sets of annotated documents where each corpus is either in English or a non-English language (de, fr, es, nl). All identified mentions of an entity have been annotated with a CUI. The following analyses have been performed to determine which corpora will be included into the challenges, and what languages should be covered in the challenges.

One assumption is that those corpora are most suitable for the challenges that comply best with the terminological resources and – after all – with the domain knowledge over all. The following parameters can be used to measure "compliance" between the terminological resources and/or the domain knowledge and the corpora.

1. The number of annotations that can be identified using the prepared terminological resource in the English corpus (L1 for language 1, see fig. 1).
2. The number of annotations that can be identified using the prepared terminological resource in the non-English language in the parallel corpus (L2 for language two, i.e. the non-English language, see fig. 1).

	CUI-L2 (all)				CUI-L1 (all)				Quota of L2 to L1			
Row Label	EMEA	Mdl	EPO	Average	EMEA	Mdl	EPO	Average	EMEA	Mdl	EPO	Average
DISO	103,330	262,783	19,348	153,291	220,079	463,479	122,570	301,923	47.3%	54.1%	15.8%	44.4%
CHEM	84,297	86,094	61,837	80,949	213,813	169,783	227,607	198,309	40.1%	52.0%	27.3%	42.6%
LIVB	52,152	83,747	9,433	57,310	117,664	200,580	107,945	149,817	44.5%	41.0%	8.9%	36.6%
ANAT	25,418	83,940	20,445	48,455	55,699	201,492	95,793	122,632	46.3%	42.8%	21.3%	40.3%
OBJC	9,240	12,304	20,515	12,543	37,783	66,553	183,593	76,063	25.3%	21.8%	11.2%	21.3%
PHYS	22,303	38,612	11,254	26,966	54,148	102,003	62,128	75,176	42.8%	39.1%	18.7%	36.9%
DEVI	4,669	5,616	12,919	6,646	26,023	22,414	112,048	40,862	17.6%	26.3%	11.5%	20.2%
PHEN	5,334	9,606	16,958	9,195	17,006	34,124	72,760	34,146	33.5%	33.2%	23.2%	31.5%
GEOG	5,257	15,941	218	8,711	12,247	22,933	12,129	16,597	43.0%	62.7%	1.8%	43.6%
CONC	50,765	72,418	27,662	55,644	572,009	674,307	1,137,178	723,502	10.0%	13.7%	2.4%	10.2%
OCCU	1,245	37,850	1,164	16,205	10,322	56,770	4,686	28,299	12.9%	64.9%	24.9%	36.3%
ORGA	2,070	7,489	150	3,938	3,538	19,316	863	9,506	67.5%	37.9%	17.5%	46.3%
GENE	177	776	1,795	742	1,267	2,557	11,041	3,682	24.6%	31.3%	16.6%	26.0%

Fig. 1. (Annotations with all terms): The table shows the number of annotations, i.e. the number of provided CUIs with the annotations, for the different corpora: EMEA, Medline (Mdl), patents (EPO) and an average figure over all corpora. All terms from the provided terminological resource have been used. L1 refers to the English documents, and L2 refers to the non-English documents. The average has been calculated the following way: for a single language all annotations from all partners have been added and then averaged across all three corpora. Then these average values have been averaged across all languages (for L2), which was not required for English alone (L1). The right section shows the percentage of annotations from L1 that have been re-identified in the non-English languages (L2). Note: the quota is not the direct fraction of L2 over L1, but has been calculated as the average fraction across the different languages and corpora.

3. The previous parameter, but now only all those English annotations are counted where non-English translation is available for the same CUI in the non-English language (L2, see fig. 2).

Counting all English annotations as reference (see tbl. 1, parameter 2) gives an analysis that is less generous to the annotation solutions ("pessimistic" or "real world" evaluation) than counting only those English annotations that comply with the third parameter ("optimistic" or "idealistic" evaluation, parameter 3). The latter evaluate the performances under the condition that for each English term there exists a non-English transcript in the TR, i.e. it ignores a number of English annotations where no translation of the term can be found in L2 anyway.

A number of open questions have been resolved through the analysis of the results that were given by the annotation of the corpora in English and the non-English languages from the project partners. The open questions were concerned with the selection of the languages for the challenge, the selection of corpora and of the semantic groups in the TR.

First, the languages have been limited to German, Spanish, French and Dutch apart from English. Second, the corpora have been limited to: Medline, patents, and EMEA. Actually, the sizes of these three corpora is quite similar and the diversity between the corpora should contribute to the diversity in the challenge.

Fig. 2. (Annotations with selected terms): The numbers in the table have been produced in the same way as the numbers in the previous figure (see fig. 1), only the number of CUIs has been determined in a different way. In this table, only those CUIs have been counted for the English corpora, where the non-English transcription of the English term exists in the terminological resource. This way, a number of English annotations have been ignored, since the non-English term would not exist anyways. This table shows values that are more generous to the annotation solutions, since they would not be able to identify terms that do not exist in the non-English form in the TR.

	CUI-L2 (preselected)				CUI-L1 (preselected)				Quota of L2 to L1			
Row Label	EMEA	Mdl	EPO	Average	EMEA	Mdl	EPO	Average	EMEA	Mdl	EPO	Average
DISO	103,330	262,783	19,348	153,291	183,481	400,212	62,974	250,234	25.3%	21.8%	11.2%	21.3%
CHEM	84,297	86,094	61,837	80,949	134,539	144,534	171,659	145,377	40.1%	52.0%	27.3%	42.6%
LIVB	52,152	83,747	9,433	57,310	104,471	150,968	55,924	114,666	33.5%	33.2%	23.2%	31.5%
ANAT	25,418	83,940	20,445	48,455	42,792	154,856	50,750	90,084	47.3%	54.1%	15.8%	44.4%
PHYS	22,303	38,612	11,254	26,966	41,682	80,768	47,785	58,782	24.6%	31.3%	16.6%	26.0%
OBJC	9,240	12,304	20,515	12,543	16,961	30,983	86,685	35,374	43.0%	62.7%	1.8%	43.6%
DEVI	4,669	5,616	12,919	6,646	18,468	15,735	83,534	29,690	46.3%	42.8%	21.3%	40.3%
PHEN	5,334	9,606	16,958	9,195	10,354	24,530	46,652	22,753	67.5%	37.9%	17.5%	46.3%
GEOG	5,257	15,941	218	8,711	9,131	19,203	6,669	12,804	17.6%	26.3%	11.5%	20.2%
CONC	48,514	72,418	27,662	54,501	97,689	126,404	77,498	105,765	44.5%	41.0%	8.9%	36.6%
OCCU	1,245	37,850	1,164	16,205	10,195	53,041	4,675	26,719	10.0%	13.7%	2.4%	10.2%
ORGA	2,070	7,489	150	3,938	2,558	16,861	770	8,084	12.9%	64.9%	24.9%	36.3%
GENE	177	776	1,795	742	910	2,168	10,676	3,309	42.8%	39.1%	18.7%	36.9%

Finally, the terminological resource is limited to the semantic groups: ANAT, CHEM, DEVI, DISO, GEOG, LIVB, OBJC, PHEN, PHYS. The other groups have been excluded, since they did not show enough coverage (GENE, ORGA, OCCU), or it was too unspecific, i.e. not specific enough for the biomedical domain (CONC) resulting to very heterogeneous annotation results across the different corpora and languages. It is obvious that the TR has been reduced to the sets of terms that are required for the challenge.

Conclusions

The CLEF-ER challenge organized from the MANTRA project partners is the first of its kind tackling a number of challenges in a large-scale annotation task allowing multilingual approaches. The main objective is the identification of biomedical entities (and concepts) in multilingual documents, where the documents are available as part of parallel corpora involving the English language and at least one non-English language.

The corpora stem from the scientific literature (Medline abstracts), drug labels (EMEA documents) and from the patent text (European patent office). A reference terminology has been provided from UMLS and has been optimized for the challenge participants. The English corpora have been annotated with the terms from the MTR using the annotation solutions of the project partners leading to a silver standard corpus, which has been distributed to the challenge participants. These annotated corpora give the challenge participants different opportunities to contribute to the challenge.

Altogether, the CLEF-ER challenge will work towards solutions that identify biomedical terms in multilingual documents of different kinds, where the proposed solutions have to cope with large amounts of terms and large data resources. The overall outcome will help to establish the semantic web in healthcare and to allow interoperability of IT solutions across country borders and languages.

Acknowledgement

This work was funded by the European Commission STREP grant number 296410 ("Mantra", FP7-ICT-2011-4.1).