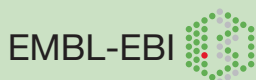




CALBC Project - the collaborative annotation of a large-scale biomedical corpus

www.calbc.eu

Partners

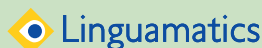


European Bioinformatics Institute
Dr Dietrich Rebholz-Schuhmann,
Project Coordinator



Erasmus University Medical Center,
Rotterdam, The Netherlands
Dr Erik van Mulligen

Friedrich-Schiller
University, Jena,
Germany
Prof. Udo Hahn



Linguamatics Ltd, Cambridge, UK
Dr David Milward

The sheer volume of biomedical scientific literature has stimulated research into automated methods to make this body of knowledge more accessible for researchers and medical practitioners. This involves automatically analysing and extracting knowledge from the texts using natural language processing techniques and, in particular, exploiting their potential to recognise named entities of relevance to biomedical research, for example proteins or genes. CALBC aims to integrate the annotations from different named entity recognition systems to create a large, annotated corpus for the automatic analysis of scientific literature. Information from these systems is gathered via the CALBC Challenges. Challenge II will be open for submissions from 13 September – 15 December 2010.

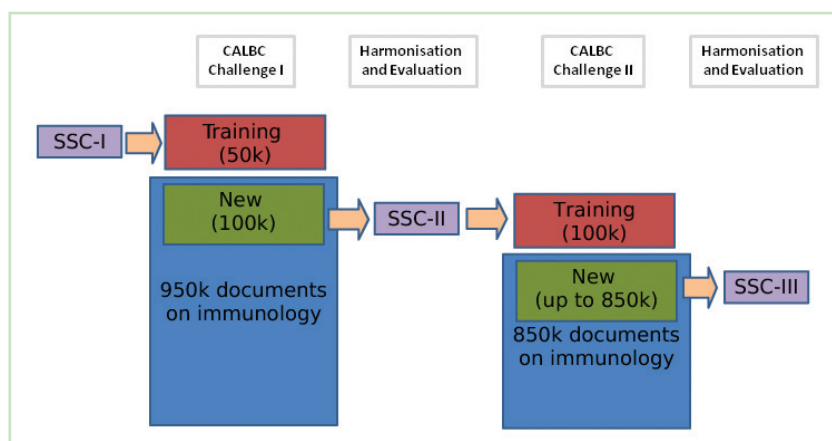
What is CALBC?

CALBC (Collaborative Annotation of a Large-scale Biomedical Corpus) is a European support action addressing the automatic generation of a very large, community-wide, shared text corpus annotated with biomedical entities [1].

Through the CALBC challenges, we gather community submissions to build the large, annotated corpus holding about one million Medline immunology-related abstracts. Participation is open to any team that is willing to submit annotations obtained with their own named entity recognition system. The annotations of all participating systems will be automatically integrated with additional metadata to develop a 'Silver Standard Corpus' (SSC). The SSC will therefore have a broader scope than any single system. The scale of the corpus means that manual curation is not possible [2] and all harmonisation steps have to be performed automatically [3].

In phase 1 (pilot), the four project partners measured the performance of their systems to develop the pilot corpus (SSC-I) [4]. In phase 2, participants submitted their annotation as part of Challenge I and received an evaluation of their results against the pilot corpus [5]. The harmonisation of these annotations has led to the development of the second Silver Standard Corpus (SSC-II).

Now in phase 3, the SSC-II will be used in Challenge II, the result of which will be the generation of the third and final Silver Standard Corpus (SSC-III).



What can I do with CALBC?

CALBC, participating in the challenge:

- If you train your named entity recognition solutions against the SSC, you will be able to extract a large number of semantic groups from any other corpus.
- Contribute to the development of the next SSC and test different types of annotation (more specific or more general) against the current SSC.

CALBC, the corpus:

The resulting corpus can be exploited for different goals:

- The text mining community can train existing text mining solutions to reproduce the CALBC annotations.
- Novel text mining solutions can be developed using the corpus, such as new methods for the disambiguation of entities.
- CALBC will provide a larger body of biomedical information than is currently available to the text mining community.

CALBC, the data resource – In addition to the IeXML annotation format, the corpus will be delivered in a Resource Description Framework (RDF) representation so that it can be integrated in the Semantic Web. The corpus will serve as a data resource for data mining solutions that contribute to the understanding of immunological questions.

Participating in the CALBC Challenge II

Sign up: Participants must sign up to the CALBC mailing list by sending a request to public@calbc.eu. The sign-up confirmation also contain the participant log-in details for the submission site where the challenge data can be downloaded.

Corpus: The Challenge II corpus consists of 875,000 Medline abstracts on immunology. These are separated into two parts: a smaller group of 175,000 abstracts for all participants to process, and a larger group of 700,000 abstracts for groups able to handle large sets of text. Participants should download the corpus and adapt their annotation system to the formats proposed by CALBC.

Submission: Annotated corpora must comply with the annotation guidelines of the challenge and be uploaded through the submission site.

Three types of annotations are considered in the evaluation: term boundaries, semantic type assignments, and concepts assignments (optional). Participants can indicate which annotation types are provided when they submit their results.

Task A (named entity recognition): Annotate the corpus with entity boundaries for one or more semantic groups, e.g. genes/proteins, diseases, species, chemicals, others.

Task B (concept identification): Annotate the corpus with boundaries and concept identifiers for the entities.

Evaluation and feedback: After submission, a fully-automated analysis system will instantly start the analysis and alignment process. The results of the alignment of the submitted corpus against the Silver Standard will be reported as soon as the alignment is finished (estimated to take approximately one day). The analysis process will deliver statistical parameters that help to interpret the contained annotations and their performance. The annotation results of the participating systems will be made available to each participant for a subset of 50,000 abstracts of the corpus.

Further reading

[1] Rebholz-Schuhmann, D., *et al.* The CALBC Silver Standard Corpus. *J. Bioinform.Comput. Biol.*, 8, 163-179 (2010)

[2] Krallinger, M. *et al.* Evaluation of text-mining systems for biology: overview of the second BioCreative community challenge. *Genome Biology*, 9, S2: 1-9 (2008)

[3] Rebholz-Schuhmann, D. *et al.* Text Processing through Web Services: Calling Whatizit. *Bioinformatics*, 24, 296-298 (2008)

[4] Rebholz-Schuhmann, D., *et al.* The CALBC Silver Standard Corpus for Biomedical Named Entities. A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers. Proc. LREC'10, Valletta, Malta. (2010)

[5] Rebholz-Schuhmann, D., & Hahn, U. Proceedings of the First CALBC workshop. (2010) <http://workshop.calbc.eu/FirstProceedings.pdf>

Support

CALBC is supported by the 7th Framework Programme of the European Commission, as part of the 'Intelligent Content and Semantics' theme (ICT-2007.4.2), grant agreement number 231727.

Need help?

URL: www.calbc.eu
e-mail: public@calbc.eu
Tel: +44 (0) 1223 492594
Fax: +44 (0) 1223 494468

Post:
Dietrich Rebholz-Schuhmann
EMBL-European Bioinformatics
Institute
Wellcome Trust Genome Campus
Cambridge
CB10 1SD
UK

