CALBC Project - towards the collaborative annotation of the immunology research literature

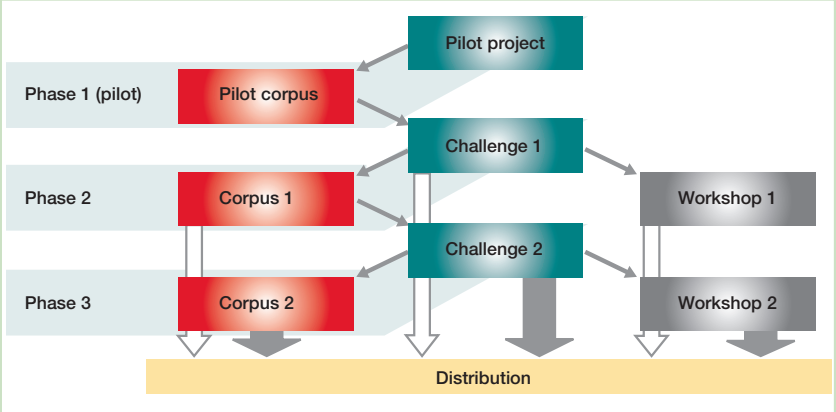
www.calbc.eu

The sheer volume of biomedical scientific literature has stimulated research in making this body of knowledge more accessible for researchers and medical practitioners. This involves automatically analysing the texts and extracting knowledge from them 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 [1]. CALBC aims to integrate the annotations from several named entity recognition systems to create a large annotated corpus for the automatic analysis of scientific literature.

What is CALBC?

CALBC (Collaborative Annotation of a Large Biomedical Corpus) is a European support action addressing the automatic generation of a very large, community-wide shared text corpus annotated with biomedical entities. We propose to create a broadly scoped and diversely annotated corpus (150,000 Medline immunology-related abstracts annotated with approximately a dozen semantic types) by automatically integrating the annotations from different named entity recognition systems.

The CALBC challenge consists of building such a large annotated corpus. 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. The integrated corpus will therefore have a broader scope than any single system.



The challenge will be performed in three phases. In phase 1 (pilot project) the four project partners deliver annotations and measure the performance of their solutions to develop the pilot corpus. In the second phase, participants submit their annotation as part of Challenge 1 and receive an evaluation of their results against the pilot corpus. The integration of these annotations leads to the development of the first corpus (Corpus 1). In phase 3, Corpus 1 is used for another challenge that results in the generation of the final silver standard corpus, Corpus 2.

Partners

EMBL-EBI

European Bioinformatics Institute
Dr Dietrich Rebholz-Schuhmann,
Project Coordinator

Erasmus MC

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

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

seit 1558

Linguamatics

Linguamatics SA, Cambridge, UK
Dr David Milward

Published by the EMBL-EBI – a part of the European Molecular Biology Laboratory

What can I do with CALBC?

CALBC, the challenge:

- The annotations from different solutions will be harmonised in a single corpus.
- The scale of the corpus means that manual curation will not be possible (see BioCreative I and II [2, 3]). All harmonisation steps have to be performed automatically.
- The evaluation of the annotations against the silver standard will be by purely automatic means [4, 5] on a farm of compute servers to enable acceptable response cycles.

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 – 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

Corpus: The corpus consists of 150,000 Medline abstracts on immunology. Participants to the CALBC challenge should download the corpus and adapt their annotation system to the formats proposed by CALBC.

Submission: Participants must be register at the submission site (accessed from www.calbc.eu). Annotated corpora must comply with the annotation guidelines of the challenge. The annotated corpora must be uploaded through the submission site to take part in the challenge.

Three types of annotations are considered in the evaluation: term boundaries, semantic type assignments, and concepts assignments (optional). Typically, each annotated term will have a corresponding semantic type (and concept identifier, if applicable). Participants can provide semantic type ids (or concept ids) for every sentence or abstract, without reference to the corresponding term. Participants can indicate which annotation types are provided when they submit their results.

No restrictions are put on the semantic types and concept identifiers provided by the systems. However, participants are encouraged to make use of a set of preferred semantic types (i.e. the UMLS type system) and vocabularies (BioThesaurus, BioLexicon, UMLS). If no preferred resources are used, participants must provide a description of their type system and vocabulary.

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 the performance of the annotations in comparison to the silver standard. 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] Krauthammer, M. & Nenadic, G. Term identification in the biomedical literature. J. Biomed. Inform. 37, 512-26 (2004)

[2] Hirschman, L. *et al*. Overview of BioCreAtIvE: critical assessment of information extraction for biology. BMC Bioinformatics. 6, S1 (2005)

[3] 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)

[4] Rebholz-Schuhmann, D. *et al*. Text Processing through Web Services: Calling Whatizit. Bioinformatics. 24, 296-98 (2008)

[5] Rebholz-Schuhmann, D., Kirsch, H. & Nenadic, G. IeXML: towards a framework for interoperability of text processing modules to improve annotation of semantic types in biomedical text. Proc. of BioLINK, ISMB 2006, Fortaleza, Brazil. (2006)

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: challenge@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

SEVENTH FRAMEWORK PROGRAMME