November 2009

INFSO.E2
Technologies for Information Management

Annual Public Report 2009

CALBC Annual Report

www.calbc.eu

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 abstracts annotated with approximately a dozen semantic types) by automatically integrating the annotations from different named entity recognition systems.

The CALBC challenge 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 to develop a ‘silver standard’ corpus which will have a broader scope than any single annotation system.

Summary of Activities

The project activities in the first year were dedicated to the generation of the first harmonised corpus that integrates the annotations for genes/proteins, diseases, species and chemical entities delivered from the initial project partners.

- The first harmonised corpus is available for the CALBC challenge. It consists of 150,000 Medline abstracts on immunology and contains about 1,500,000 annotations.
- The project partners have established an IT infrastructure that enables the automatic evaluation of annotated corpora at a large scale (see www.calbc.eu). This IT infrastructure processes all submitted annotated corpora from the participants on the EBI’s local computer network and delivers the results from the assessment to the participants.
- Several harmonisation schemes and evaluation measures for the alignment of semantic annotations in the corpora have been tested and the best solution has been used to generate the annotated corpus.

Corpus formats for automatic alignment and evaluation of submissions

The formatting of the annotated corpus is crucial to the success of the CALBC project. The definition of the format enables integration of the annotations into the documents. This is important to enable alignment of the annotated corpora in an efficient IT infrastructure to identify the agreement and disagreement between the annotations from the participant's corpus and the harmonised corpus. Furthermore, we can expect that the document format can be exploited for the semantic enrichment of scientific documents.

The following characteristics have been implemented for the annotation format:

- All the annotations will be based on XML, since XML supports readability for both computers and humans.
- Annotations within the document are preferred over annotations at the end of the document (inline annotations over standoff annotations).
- A namespace is used to identify the concept in the original knowledge source.
- The exact boundaries of the entity are specified including the annotation of a semantic group.

Example: **<e id="Uniprot:P01308:T028:PRGE|UMLS:C1337112:T028:PRGE">INS gene</e>**

Submission system

The submission site allows challenge participants to submit several annotated sets of documents. The participants require an account that is created when they subscribe to the challenge.

Documents are uploaded to the CALBC Web site using the web interface. Uploaded documents can be compared to corpora available at the site; each comparison is performed as an independent compute process on EBI's computer farm.

Every participant is notified by email about the progress of the compute job and once the job has terminated. A single comparison is completed in a few minutes or only after several hours depending on the size of the submitted corpus. The results from the comparison are listed on the submission Web site.

Different types of evaluation measures

The project partners have tested different evaluation measures. The so-called "majority voting" calculation measures pair-wise agreement between two individual partners. If the evaluation system identifies the agreement amongst a larger number of participants, then it produces a lower number of annotations that fulfill this criterion. In the latter case, when comparing the participants' systems against the harmonized corpus, the precision of the participants' systems increases, the recall decreases and overall the F-measure becomes lower.

An alternative measure is the cosine similarity of the tokens found within the boundaries delivered from both annotation systems. This measure is less strict than the previous measure in the sense that it enables the comparison of annotations even if the boundaries do not match strictly ("fuzzy matching"). Tokens that cause mismatches at the left or right boundaries of the annotations are weighted with their IDF score that has been determined from the whole corpus. If the boundary disagreement between two annotations is induced by a token with a low IDF weight then the cosine score of the aligned annotations differs only to a small extend and will be accepted.

Rollout of the challenge

50,000 annotated documents will be made available to the public for training (middle of November). The remaining 100,000 unannotated documents will be delivered for testing at a later stage. The submission of the final sets of annotated documents (100,000 documents) is expected to take place at the end of January 2010.

Challenge / Task 1 (Named Entity Recognition): The participant's system provides annotations of the boundaries and semantic groups of the found entities.

Challenge / Task 2 (Concept Identification): The participant's system provides annotations of the boundaries and concept identifiers of the found entities.

User Involvement, Promotion and Awareness

Several activities improved the public awareness to the project.

A public Web site has been set up (www.calbc.eu) in combination with a document repository and mailing services (www.google.com/a/calbc.eu). Mailing lists are available for corpus related work (corpus@calbc.eu), the challenge (challenge@calbc.eu) and the general public (public@calbc.eu). Participants into the challenge and the project have to send a request for registration to be included into the challenge.

The project has been presented as part of public presentations at several conferences: (1) the industry programme meeting for small enterprises (Vienna, Austria), (2) the BioCreative II.5 workshop (Madrid, Span), (3) the IEEE EMBC conference (Minneapolis, Minnesota, U.S.A), (4) the Workshop on discourse analysis at the ISWC 2009 conference (Washington, USA), and (5) the LBM 2009 conference (Jeju Islands, Korea).

First joint publications between the project partners have been submitted to LBM2009 and to LREC2010.

Future Work

The current annotated and harmonised corpus provides a consensus of annotations. The consensus is defined by the automatic methods used to produce an alignment of the annotations from the different project partners.

Future work is concerned with the integration of more semantic types, the empirical evaluation of similarity measures for the harmonisation of the annotations and the exploration of new approaches for the harmonisation of the corpus.

In 2009 and 2010, the project partners will organise and finalize the first CALBC challenge and will organise a scientific workshop for the challenge participants to exchange their experiences and scientific results.

Further Information

The project flyer is available from the following link:
http://www.ebi.ac.uk/Rebholz-srv/CALBC/flyer.pdf

Annotation guidelines for the CALBC challenge:
http://www.ebi.ac.uk/Rebholz-srv/CALBC/challenge_guideline.pdf

Full corpus containing 150,000 Medline abstracts on immunology:
ftp://ftp.ebi.ac.uk/pub/software/textmining/corpora/immunology/immu.subset.sent.xml.gz

Access to 50,000 annotated documents from the full corpus (see above) to be used as training or testing corpus through the CALBC submission Web site.