CALBC: Releasing the Final Corpora

Şenay Kafkas[1*], Ian Lewin[2], David Milward[2], Erik van Mulligen[3], Jan Kors[3], Udo Hahn[4] and Dietrich Rebholz-Schuhmann[1]

[1]European Bioinformatics Institute, Wellcome Trust Genome Campus
Hinxton, CB10 1SD, U.K.

[2]Linguamatics LTD, St. John's Innovation Centre,
Cowley Road, Cambridge, UK

[3]Erasmus Medical Centre, Rotterdam, Nl

[4]Julie Lab, University of Jena, Jena, Germany

E-mail: kafkas@ebi.ac.uk, ian.lewin@linguamatics.com, david.milward@linguamatics.com, e.vanmulligen@erasmusmc.nl, j.kors@erasmusmc.nl, udo.hahn@uni-jena.de, rebholz@ebi.ac.uk

Abstract

A number of gold standard corpora for named entity recognition are available to the public. However, the existing gold standard corpora are limited in size and semantic entity types. These usually lead to implementation of trained solutions (1) for a limited number of semantic entity types and (2) lacking in generalization capability. In order to overcome these problems, the CALBC project has aimed to automatically generate large scale corpora annotated with multiple semantic entity types in a community-wide manner based on the consensus of different named entity solutions. The generated corpus is called the silver standard corpus since the corpus generation process does not involve any manual curation. In this publication, we announce the release of the final CALBC corpora which include the silver standard corpus in different versions and several gold standard corpora for the further usage of the biomedical text mining community. The gold standard corpora are utilised to benchmark the methods used in the silver standard corpora generation process and released in a shared format. All the corpora are released in a shared format and accessible at www.calbc.eu.

Keywords: Silver Standard, Named Entity Recognition, Corpora

1. Introduction

A number of challenges have been organized to deliver gold standard corpora (GSCs) and evaluate participating systems against them. BioCreative (Hirschman et al., 2005; Krallinger et al., 2008) and JNLPBA (Kim et al., 2004) are two such challenges. These challenges have played a pioneering role in terms of delivering GSCs and enabling the generation of efficient named entity recognition (NER) tools. However, because the GSCs are manually annotated they are often limited in size and in semantic entity types (e.g. proteins and genes, chemicals) i.e. up to only a few thousand documents annotated with a few semantic entity types. Training a named entity recognizer with a small sized corpus limits it generalization capability. Furthermore, currently, the biomedical text mining community can produce trained solution for a few number of semantic entity types only. The CALBC project (Collaborative Annotation of a Large Biomedical Corpus) has been aimed at overcoming these short-comings.

The purpose of the CALBC project is to automatically generate the first large-scale, community-wide annotated biomedical text corpus with multiple semantic entity types. The CALBC corpus includes about 1 million Medline abstracts from the domain of immunology. Two CALBC challenges have been organized to collect annotations. These challenges were similar to earlier ones in that the project partners[1] (PP) delivered a set of annotated corpora to the community for the reproduction of the corpora in a collaborative manner. However, they differ both in the size and generation methods of the corpora used: The annotations have been generated and integrated fully automatically based on our harmonization procedures by utilizing different NERs for the most important semantic entity types: Proteins and Genes (PRGE), Chemicals (CHED), Diseases and disorders (DISO), Living Beings (LIVB) (Rebholz-Schuhmann et al., 2010a; Rebholz-Schuhmann et al., 2010b). The resulting corpus is named the silver standard corpus (SSC).

The CALBC project brings the following novelties to the biomedical language processing (BioNLP) domain: (i) generation of a very large scale corpora (ii) usage of SSC rather than GSC (iii) analyzing the outcomes of usage of SSC generation approach.

This publication describes the final CALBC corpora which have been recently released and provides an overview on the generation of the corpora. The final corpora include SSC-III (including versions with annotations of a single semantic entity type and multiple type annotations), additional SSCs based on the CALBC

[1] European Bioinformatics Institute (EBI), Rebholz Group; Linguamatics (LM); Erasmus Medical Centre (EMC), Department of Medical Informatics; Friedrich Schiller University (FSU), Julie Lab.

project partners' annotations only and several GSCs used for benchmarking our harmonization approach. All the corpora described here are publicly available through www.calbc.eu.

2. Silver Standard Corpora Generation

The final SSC has been generated as a result of two CALBC challenges. Figure 1 shows the set-up of the CALBC challenges and SSC generation process. Briefly, SSC-I is the seed corpus. It contains annotations derived from the PP only and formed the basis of the first challenge. The annotations from the PP systems have been harmonized into SSC-I by applying a voting-scheme. The SSC-I data consists of 50k training and 100k test datasets. The first challenge received 60 submissions from 22 different teams (including PP). All the submissions were evaluated against the SSC-I and annotations from the best performing ones were harmonized in to SSC-II. Outcomes of the first challenge and detailed information on the generation of SSC-I and SSC-II have been provided in (Rebholz-Schuhmann et al., 2011).

In the second challenge, participants were provided with the 100k SSC-II training set and two test datasets namely SSC-II-Small (175k) and SSC-II-Big (714k). Annotation of the small set was mandatory while it was optional to annotate the big set. The second challenge received 56 submissions from 16 different teams (including PP). The SSC-III corpus has been generated through harmonizing the annotations for all four semantic entity types separately using the centroid harmonization method developed in the scope of the CALBC project (see section 3.1).

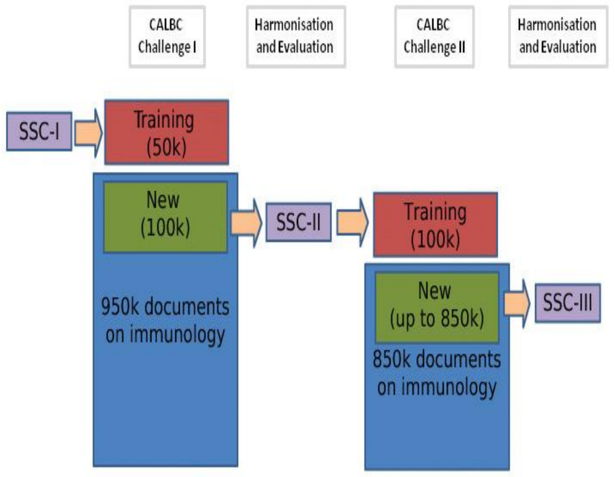
CALBC Challenge I
Harmonisation and Evaluation
CALBC Challenge II
Harmonisation and Evaluation
SSC-I
Training (50k)
New (100k)
950k documents on immunology
SSC-II
Training (100k)
New (up to 850k)
850k documents on immunology
SSC-III

Figure 1: Set-up of two CALBC challenges and generations of the SSCs

3. Silver Standard Corpora

3.1 Centroid Harmonization Method

The centroid method has been developed and used in the scope of the CALBC project to harmonize and represent annotations within a single semantic entity type. The centroid represents the "centre of all provided annotations (of semantic entity type X)" along with the distribution of alternative boundaries. Briefly, in this method, the input texts provided from different annotation systems are tokenized at the character level and votes of adjacent name-internal characters are taken to determine both the centroid itself and the possible alternative boundaries around it. Details of the method are provided in our parallel submission to LREC (Lewin, Kafkas & Rebholz-Schuhmann, 2012).

3.2 Data Format

The data format used to represent the centroid along with the annotation boundary distribution is illustrated in Figure2.

***Context*:**
Identification of a 68 - kilodalton nuclear
ATP - binding phosphoprotein
encoded by bovine papillomavirus type 1
Annotations:
(1) 68-kilodalton nuclear *ATP-binding phosphoprotein*
(2) nuclear *ATP-binding phosphoprotein*
(3) *ATP-binding phosphoprotein*

(a)

Identification of a 68 - kilodalton nuclear
<e b="l:0:1|l:19:1|l:7:1|r:0:3">ATP - binding
phosphoprotein</e>
encoded by bovine papillomavirus type 1

(b)

Figure 2: Sample annotation representation (a) Annotations (b) Centroid with alternative boundaries

"ATP - binding phosphoprotein" is the centroid of all the annotations listed in Figure 2-(a). The centroid is marked up with the <e> tag and the alternative boundaries are defined with the b attribute. Boundaries are represented as tuples <direction:position:value> and separated with the bar character. "Direction" defines the type of boundary (left or right). "Position" shows the position of the boundary (with respect to the centroid) and the "value" shows the number of alternative annotations that support this boundary. For example l:0:1 indicates that there is a left boundary at position 0 supported by 1 annotation. Similarly r:0:3 indicates that there is a right boundary at position 0 supported by 3 annotations.

3.3 Entities Spanning Several Semantic Entity Types

Participants of the CALBC challenges were encouraged to submit annotations in up to 16 different semantic entity types. They were also encouraged to attach external database identifiers (e.g. a UMLS concept id) to their annotations.

Most only submitted data for some types. Furthermore, most submitted data for their different types using different copies of the original corpus. The bringing together of annotations of the same corpus for different semantic entity types represents a quite novel challenge for named entity research.

The CALBC "multiple-type" corpus that we release represents the results of our initial investigations into this task. Our approach is first to harmonize the corpus several times, once for each semantic entity type, using the centroid method described in section 3.1. This gives us the "hearts" of the challenge participants' contributions. Secondly, we join the harmonized corpora by choosing "longest boundary extent" (the most informative name) where there is conflict but we assign to it all the database identifiers assigned to any part of it, and note when entities contain nested terms of other types.

<e ct=”prge” nt=”diso|ched” id=”...”>Cystic fibrosis transmembrane conductance regulator </e> This entity is identified as a protein/gene which contains within it disease and chemical terms. The id attribute (shown opposite) details the database identifiers assigned by various participants	umls:C0056889:T116:prge uniprot:Q6PQZ2:T116:prge umls:C0056889:T123:ched chemlist:4250345::ched disease:C0010674::diso meddra2:10011762::diso nci:Cystic_Fibrosis::diso uniprot:Q07DZ6:T116:prge UMLS:C0010674:T047:diso

Figure 3: Sample annotation of single corpus representation of multiple types

3.4 Released SSCs

We have generated different harmonised corpora based on all available annotations from PP and challenge participants ("all-comers") or on PP alone ("partners"). There is one corpus (centroid representation) segmented for 4 different semantic entity types (PRGE, CHED, DISO, LIVB) and one corpus with the multiple type representation. Finally, we also release our corpora in 2 different sizes. Table 1 shows the multiple type SSC-III data while Table 2 shows the single type data. SSCs based on the partners annotations are provided as additional data (see Table 3).

Corpus	#Abs.	SET	#Annot.	#V
SSC-III Small ("all comers")	174,999	ALL	2,548,900	5
SSC-III Big ("all comers")	714,283	ALL	10,304,172	4

Abs.: Abstracts, ST: Semantic Entity Type, Annot.: Annotations, V: Votes

Table 1: Multiple Type SSCs

Corpus	#Abs.	SET	#Annot.	#V	#S
SSC-III Small ("all comers")	174,999	PRGE	782,598	5	13
SSC-III Small ("all comers")	174,999	CHED	556,716	5	11
SSC-III Small ("all comers")	174,999	DISO	597,085	5	12
SSC-III Small ("all comers")	174,999	LIVB	775,371	5	12
SSC-III Big ("all comers")	714,283	PRGE	3,122,527	4	9
SSC-III Big ("all comers")	714,283	CHED	2,549,941	4	9
SSC-III Big ("all comers")	714,283	DISO	2,559,251	4	10
SSC-III Big ("all comers")	714,283	LIVB	3,284,426	4	10

Abs.: Abstracts, SET: Semantic Entity Type, Annot.: Annotations, V: Votes, S:cSubmissions

Table 2: Single Type SSCs

Corpus	#Abs.	SET	#Annot.	#V	#S
SSC-III Small ("partners")	174,999	PRGE	501,637	2	4
SSC-III Small ("partners")	174,999	CHED	497,561	2	4
SSC-III Small ("partners")	174,999	DISO	496,573	2	4
SSC-III Small ("partners")	174,999	LIVB	735,059	2	4
SSC-III Big ("partners")	714,283	PRGE	2,675,768	2	4
SSC-III Big ("partners")	714,283	CHED	2,995,690	2	4
SSC-III Big ("partners")	714,283	DISO	2,239,680	2	4
SSC-III Big ("partners")	714,283	LIVB	3,284,426	2	4

Abs.: Abstracts, SET: Semantic Entity Type, Annot.: Annotations, V: Votes, S:Submissions

Table 3: Partner SSCs

4. Gold Standard Corpora

In addition to the SSC data, we also release several GSCs in which annotations are represented in the official format of CALBC, IeXML (Rebholz-Schuhmann, Kirsch & Nenadic, 2006). These GSCs are used to benchmark our SSC generation approach. For this purpose, an SSC is generated by harmonizing the solutions from individual named entity taggers. Analysis results show that the generated SSC outperforms any single annotation solution when they are evaluated against several GSCs. Details on our evaluation methods and results will be the focus of another publication which we plan to publish it in the near future. The released GSCs are listed in Table 4.

Corpus	#Sentences /Abstracts	SET	#Annot.	YR
BioCreative II Gene Mention Test Dataset*	4,171 sentences	PRGE	5,144	2005
PennBioIE Oncology	1,414 abstracts	PRGE	18,148	2008
JNLPBA Test Dataset	401 abstracts	PRGE	6,142	2004
FSU-PRGE	3,236 abstracts	PRGE	59,483	2009
Arizona* Disease	2,775 sentences	DISO	3,206	2008
SCAI-Test dataset	100 abstracts	CHED	1,206	2008

SET: Semantic Entity Type, Annot.: Annotations, YR: Year of Release

*Sentences in the released versions of BioCreative-II and Arizona Disease datasets do not include all the sentences from the original version of the corpora. Those original sentences were mapped, in so far as we were able to, into their original PubMed abstracts and it is the abstracts (not just the sentences) that were tagged by CALBC participants. Therefore, some content may be lost in this mapping process.

Table 4: GSCs

5. Discussion and Conclusion

In this paper, we present the recently released final CALBC data. The data includes SSCs in different versions and several GSCs. The SSCs are generated in a fully automatic manner by harmonizing solutions from different NER systems for multiple semantic entity types. NER systems differ in their application scopes and annotation strategies. Therefore, it could be expected that they complement each other. Consequently, the generated SSCs reflect these different scopes and strategies.

The generated SSCs are the first large-scale corpora in the BioNLP domain. A large-scale corpus is expected to allow more reliable statistics and less likelihood of overfitting to small datasets. A large-scale SSC could potentially lack in quality compared a given GSC. However, the balance between scale and quality depends on the task for which the corpus is used.

NER developers generally evaluate their systems against different corpora. However, generated corpora are available to the public in different formats. This requires NER developers to adapt their systems according to a given corpus. Lack of a shared format across all corpora hinders progress in the development as well as evaluation of NER systems. In order to overcome this problem, the CALBC corpora (the SSCs and the GSCs) are released in a shared format called IeXML which is a standardized format for representing the annotations. We expect that a shared format across all corpora would not only reduce the workload of NER developers but also the error rate when the system is evaluated against the corpora.

We believe the released CALBC data will serve as useful resources for developing and improving text mining systems. For example, the released SSCs can be mined for multi semantic entity types as well as context of the annotations.

6. Acknowledgements

This work was funded by the EU Support Action grant 231727 under the 7th EU Framework Programme within Theme “Intelligent Content and Semantics” (ICT 2007.4.2).

This work was partially funded by the EU FP7 Support Action grant 231727 (ICT 2007.4.2) Collaborative Annotation of a Large Biomedical Corpus and 296410 (ICT 2011.4.1) Multilingual Annotation of Named Entities and Terminological Resource Acquisition.

We also very gratefully acknowledge the support of the CALBC challenge participants.

7. References

Hirschman, L., Yeh, A., Blaschke, C., Valencia, A. (2005). Overview of BioCreAtIvE: Critical assessment of information extraction for biology, BMC Bioinformatics, 6(Suppl 1):S1.

Kim JD, Ohta T, Tsuruoka Y, Tateishi Y, Collier N: Introduction to the bio-entity recognition task at JNLPBA. Proceedings of the JNLPBA-04 Geneva, Switzerland; 2004, 70-75.

Krallinger, M., Morgan, A., Smith, L., Leitner, F., Ta-nabe, L., Wilbur, J., Hirschman, L., Valencia, A. (2008). Evaluation of textmining systems for biology: Overview of the Second BioCreAtIvE Community Challenge, Genome Biology, 9(Suppl 2):S1.

Lewin, I., Kafkas, S., Rebholz-Schuhmann, D. (2012): Centroids: Gold Standards with Distributional Variations, Proceedings of the International Conference on Language Resources and Evaluation 2012.

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

Rebholz-Schuhmann, D., Yepes, J.A.J., van Mulligen, E.M., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Tomanek, K., Beisswanger, E., Hahn, U. (2010a). The CALBC Silver Standard Corpus for Biomedical Named Entities: A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers. Proceedings of the LREC 2010 ELRA, Valletta, Malta.

Rebholz-Schuhmann, D., Yepes, J.A.J., Van Mulligen, E., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U. (2010b). CALBC Silver Standard Corpus, Journal of Bioinformatics and Computational Biology, 8(1):163-79.

Rebholz-Schuhmann, D. et al. (2011). Assessment of NER solutions against the first and second CALBC Silver Standard Corpus, Journal of Biomedical Semantics, 2(Suppl 5):S11.