



## Second CALBC Workshop

---

*Hinxton Hall*

*Wellcome Trust Genome Campus*

*Hinxton, Cambridgeshire, UK*

*16<sup>th</sup>-18<sup>th</sup> March 2011*

---

*Dietrich Rebholz-Schuhmann*

*Şenay Kafkas*

*(editors)*



EMBL-EBI



Erasmus MC

*Erasmus*

 Linguamatics



# Foreword

---

Dietrich Rebholz-Schuhmann

Dear CALBC workshop participants,

Welcome to the second CALBC workshop. In this workshop the CALBC project partners and the CALBC challenge participants will present the work of the two previous years that leads to the development of the first large-scale annotated biomedical corpus. The un-annotated corpus was provided to the public for annotation as part of the CALBC challenge. The CALBC project is novel in the biomedical NLP research community in its scale, in its approach of using silver standard corpora rather than gold standard corpora, and in its examination of the consequences of taking a silver standard approach. The silver standard corpus is a corpus that has been automatically generated on the basis of alignment and harmonisation procedures operating on the input of a large variety of named entity taggers.

Biomedical text mining solutions face the condition that biomedical terminologies and document repositories are both increasing in size and complexity. The use of terminologies, the recognition of entities in the text and the normalisation of entities to concept identifiers form tasks that would profit from a large-scale annotated corpus. Unfortunately, it would require significant resources to build such a large-scale annotated corpus. The CALBC project tackles this task through the reconciliation of the annotations from different annotated corpora from the project partners as well as from participants to the CALBC challenge. In total, the challenges lie in the amount of data that has to be processed as part of the challenges, in the automatic alignment of corpora resulting from the challenges, the resolution of conflicts between the annotations from different systems, the normalisation of entities in the large-scale corpus and in the preparation of a consistent corpus.

During the workshop, the project partners will give presentations on the ongoing work and the results from the challenges. Challenge participants have the opportunity to present their annotation solutions. The first 1½ day of the workshop are reserved for the core of the CALBC challenge and the last ½ day is dedicated for a discussion of the prospect of the results from CALBC.

The project partners of the CALBC project are the European Bioinformatics Institute (D. Rebholz-Schuhmann), Erasmus Medical Center Rotterdam (J. Kors, E. van Mulligen), Friedrich-Schiller University in Jena (U. Hahn) and Linguamatics (D. Milward). Project participants are from Europe and Asia. The project started in January 2009 and terminates in June 2011. During this period the project partners organised the first and the second CALBC challenge in autumn 2009 and autumn 2010, respectively. The main outcome of the project will be an annotated corpus of about 1 million Medline abstracts providing annotations from the first and the second challenge.

Welcome again and enjoy the workshop!

The CALBC project partners

# Workshop program – Day 1

Time	Day 1 (Wednesday – March 16th, 2011)
12.00-13.00	Lunch
13.00-15.00	<b>Session 1: Introduction to the CALBC challenges, setup, participation</b>
13.00-13.15	Overview: Setup of the first and the second CALBC challenge, overview on the analyses (Dietrich Rebholz-Schuhmann, EBI, U.K.)
13.15-13.45	Methods for Matching of Annotations, Harmonisation and Evaluation (Jan Kors, Erasmus University Medical Center, Rotterdam, NI)
13.45-14.10	Concept identification by machine learning aided dictionary-based named entity recognition and rule-based entity normalisation (György Móra, University of Szeged, Hu)
14.10-14.35	Annotating the CALBC corpus with a machine learning harmonization approach (David Campos, Universidade de Aveiro, Pt)
14.35-15.00	Annotating large corpora with concept retrieval (Rafael Berlanga, Universitat Jaume I, Es)
15.00-15.30	Break
15.30-17.00	<b>Session 2, part 1: Annotation methods applied to the CALBC corpus</b>
15.30-15.55	Dictionary-based concept identification with UMLS (Max De Wilde, University of Antwerp, Be)
15.55-16.20	OntoGene at CALBC II and Some Thoughts on the Need of Document-Wide Harmonization (Simon Clematide, University of Zurich, Ch)
16.20-16.50	Challenge I vs. Challenge II: Set-up, Participation Data and Feedback (Udo Hahn, Friedrich-Schiller-University, Jena, De; D. Rebholz-Schuhmann)
16.50-17.00	Recap and initial open discussion on the setup and outcome of the CALBC project and challenge
17.00-18.00	<b>Keynote talk: Prof. Yves Moreau (Katholieke Universiteit, Leuven, Be)</b> <i>Candidate gene prioritization by genomic data fusion</i>
18:00	Bus ride to Cambridge
19:00	<b>Dinner, Queens College, Cambridge</b>

# Workshop program – Day 2

Time	Day 2 (Thursday - March 17th, 2011)
<b>9.00-11.00</b>	<b>Session 2, part 2: CALBC II challenge and gold standard data</b>
9.00-9.30	Building the SSC and evaluation of the annotation systems against the SSC (Ian Lewin, EBI, U.K.; Jan Kors; Dietrich Rebholz-Schuhmann)
9.30-10.00	Evaluation of the SSC against the Gold Standard Corpora (Senay Kafkas, EBI, U.K.; Ian Lewin, EBI, U.K.; Dietrich Rebholz-Schuhmann)
10.00-10.30	A CRF-based approach to harmonize heterogeneous gene/protein annotations (David Campos, Universidade de Aveiro, Pt)
10.30-11.00	Use cases for the Silver Standard Corpora (David Milward, Linguamatics, Cambridge, U.K.; Peter Corbett, Linguamatics, Cambridge, U.K.)
<b>11.00-13.00</b>	<b>Keynote Talk: Lynette Hirshman (Mitre, U.S.A.) <i>Coupling Evaluation to End Users: Case Studies in Text Mining for Biomedicine</i></b>
<b>12.00-13.00</b>	Lunch time
<b>13.00-15.00</b>	<b>Session 3: Normalisation of data and Semantic Web</b>
13.00-13.30	Normalisation of lexical entities: Jochem, LexEBI, cross-comparisons (Dietrich Rebholz-Schuhmann et. al.)
13.30-14.00	Normalisation of the silver standard corpus using normalised lexical resources: evaluation of annotation solutions (Ernesto J. Ruiz, Oxford University, U.K.; Ian Lewin)
14.00-14.40	A Semantic Model for Federated Queries Over a Normalized Corpus (Samuel Croset, EBI, U.K.; C. Grabmüller, EBI, U.K.)
14.40-15.00	Integration of literature with biomedical data resources- the SESL project (Christoph Grabmüller, EBI, U.K.; Samuel Croset, EBI, U.K.; Dietrich Rebholz-Schuhmann)
<b>15.00-15.30</b>	Break
<b>15.30-16.30</b>	<b>Keynote Talk: Therese Vachon (Novartis Institutes for Biomedical Research, Basel, Ch)</b>
<b>16.30-18.00</b>	<b>Poster Session</b>

# Workshop program – Day 3

---

Time	Day 3 (Friday- March 18th, 2011)
9.00-10.00	Keynote talk: Timo Hannay (Digital Science / Nature Publishing Group, London) <i>When content meets technology</i>
10.00-12.00	Session 5: Large-scale annotation – next generation?
10.00-10.20	Is the interoperability and normalisation of semantic resources solved?
10.20-10.40	Is the integration of different semantic types in the same corpus solved?
10.40-11.00	Which use cases from involving Semantic Web technology would profit from the integration of the scientific literature?
11.00-11.20	Does the CALBC approach scale in such a way that all scientific publications in the biomedical domain would be annotated with automatic means in the future?
11.20-11.40	Can we tackle the problem of multi-linguality with the CALBC approach?
11.40-12.00	What other data resources apart from the scientific biomedical literature should be annotated with the CALBC approach?
12.00-12.30	Wind-up and closing
<b>12.00-13.00</b>	Lunch time

# Table of Content

---

Invited Speakers.....	7
Yves Moreau .....	7
Lynette Hirschman .....	7
Therese Vachon .....	7
Timo Hannay .....	8
Abstracts – CALBC project in general.....	9
Overview: Setup of the first and the second CALBC challenge, overview on the analyses.....	9
Methods for Matching of Annotations, Harmonisation and Evaluation .....	12
Challenge I vs. Challenge II: Set-up, Participation Data and Feedback.....	13
Building the SSC and evaluation of the annotation systems against the SSC.....	14
Evaluation of the SSC against the Gold Standard Corpora .....	15
A CRF-based approach to harmonize heterogeneous gene/protein annotations .....	17
Use Cases for the Silver Standard Corpora .....	19
Abstracts – CALBC Challenge II: system descriptions – project partners .....	28
Annotating the second CALBC corpus with Peregrine .....	31
Robust Annotation of BioMedical Entities for CALBC using I2E .....	33
Annotating the CALBC Challenge II Corpora with the JULIE Lab Tools .....	34
Abstracts – CALBC Challenge II: system descriptions – participants .....	37
Concept identification by machine learning aided dictionary-based named entity recognition and rule-based entity normalisation.....	37
Annotating large corpora with concept retrieval .....	40
Annotating the CALBC corpus with a machine learning harmonization approach .....	43
Dictionary-based concept identification with UMLS .....	46
OntoGene at CALBC II and Some Thoughts on the Need of Document-Wide Harmonization.....	48
Biomedical Named Entity Recognition in CALBC Challenge II using Dictionaries and SVMs .....	52
Brief Description of ITNLP System for CALBC II.....	55
MetaMap in the CALBC Workshop II .....	57
Scalable Interlinking of Bio-Medical Entities and Scientific Literature in Linked Life Data .....	59
List of attendees.....	62

## Invited Speakers

### Yves Moreau

Katholieke Universiteit, Leuven, Be

**Title: Candidate gene prioritization by genomic data fusion.**

### Lynette Hirschman

The MITRE Corporation, U.S.A.

**Title: Coupling Evaluation to End Users: Case Studies in Text Mining for Biomedicine**

This talk will examine evaluations of text processing for the biomedical literature as well as the increasing activity in evaluation of clinical records. As these areas begin to intersect, for example, in pharmacogenomics applications or in genotype and phenotype-wide association studies, the potential applications begin to drive the development of resources and the creation of associated challenge evaluations. We will use the BioCreative evaluations of text mining for the biomedical literature and the i2b2 evaluations for clinical records as case studies to explore how such evaluations bring together end user and developer communities to develop new challenges and create useful tools for end users.

### Therese Vachon

Novartis Institutes for Biomedical Research, Basel, Ch

**Title: The Knowledge Integration Framework, Backbone for knowledge federation, assisted annotation, data curation and answering of complex queries in Drug Discovery.**

Complex drug discovery questions require federation of a vast amount of internal and external data sources (e.g. Genomics, Proteomics, Chemogenomics, Chemistry, Portfolio and Competitive Intelligence). Cross-linking of information to the original sources is a key need (e.g. genes, proteins, targets, metabolic pathways, modes of action, diseases, biomarkers, anatomy, cell lines, tissues, assays, products, projects, compounds, ligands). The use of data standards, Terminologies and Ontologies is mandatory to integrate data sources and to elucidate, model and share knowledge about chemical, biological and disease mechanism information. Data are poorly annotated and need to be curated and mapped to other sources.

The Knowledge Integration Framework provides a federation layer based on well controlled terminologies and referential knowledge. The data are produced with text mining and knowledge mapping techniques on a large set of scientific repositories covering the scientific concepts represented in the framework. The terminologies that we implemented aim at a uniform wording across scientific data repositories. The framework supports consistent annotation of data, data curation, terminology curation, collaborative management of terminology, data integration, semantic mapping and normalization, search and contextual navigation.

In 2010, the usage of the Framework has grown significantly (number of users, number of applications integrating the framework) and became part of the NIBR IT and Informatics community strategy.

The presentation will include some use cases to demonstrate the usage in key areas.

## **Timo Hannay**

Digital Science / Nature Publishing Group, London, U.K.

### **Title: When content meets technology**

Science publishers have long been concerned with content, but only recently, and somewhat reluctantly, with technology. Yet the real opportunity to transform scientific discovery for a digital age lies in combining them both. In this talk, Timo Hannay, formerly Publishing Director of Nature.com and now Managing Director of the Nature spin-off business, Digital Science, assesses these opportunities, and describes what he and his colleagues are doing to help realize them.

## Abstracts – CALBC project in general

### Overview: Setup of the first and the second CALBC challenge, overview on the analyses

Dietrich Rebholz-Schuhmann

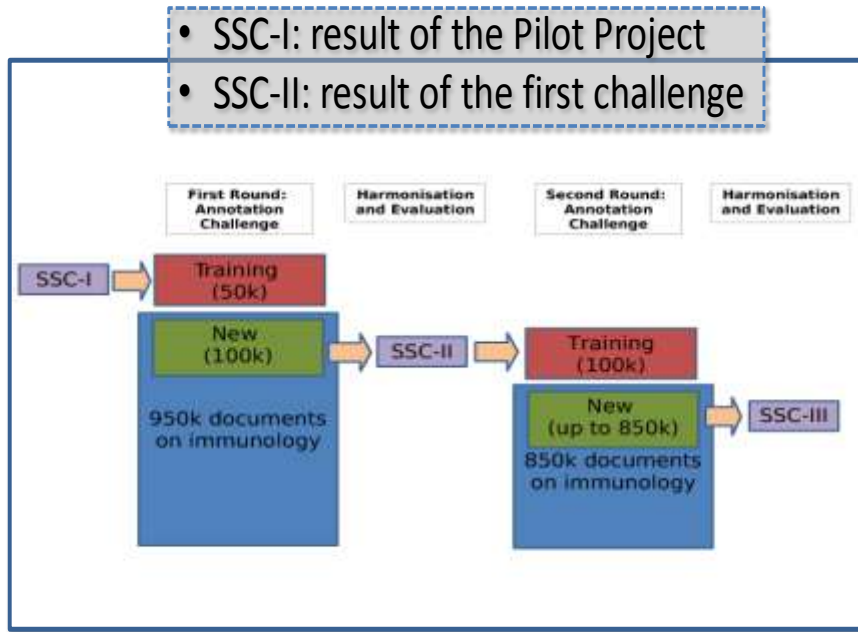
European Bioinformatics Institute  
Wellcome Trust Genome Campus  
Hinxton, Cambridge, CB10 1SD, U.K.  
rebholz@ebi.ac.uk

This presentation will give an overview on previous work in the CALBC project, on the current status of the CALBC project and on organisation of the first and the second CALBC challenge.

In the CALBC project the project partners have approached a number of problems that in principle lead to better support of the automatic annotation of the scientific literature. First, the project partners have harmonised annotations from different annotation solutions. Second, they have generated several corpora that contain a large number of entities of different semantic types. Third, they have performed a number of performance measurements to assess the quality of the Silver Standard Corpus against the Gold Standard corpora [1,2]. Last, they have generated a semantic resource that enables exploitation of the annotated corpus as part of the public Semantic Web infrastructure.

Part of the project work was concerned with the generation of the harmonised corpora and the evaluation of annotated corpora against the harmonised corpus. This required continued work on the methods for harmonisation and evaluation. Over the past year, the methods for the harmonisation of the corpus and the alignment of the annotated corpora have been improved and tested on a large scale. New solutions for the normalisation of terminological resources have been developed and the normalisation of the harmonised set has been progressed, tested and evaluated. Finally, the project partners have performed a significant number of tests of their own solutions against gold standard corpora, the generation of Silver Standard Corpora from the alignment of their results based on the gold standard corpora, and the measurements of the Silver Standard Corpora against the Gold Standard Corpora.

The CALBC project partners have performed two challenges and have used different corpora to compare the annotations from the different annotation solutions. The second challenge was opened on 13<sup>th</sup> September 2010 and was announced on the BioCreative III meeting in Washington DC. 100,000 documents have been made available to the public for training and two corpora of Medline abstracts (175k + 714k) have been made available for annotation (see fig. 1). On 20<sup>th</sup> January 2011, the second challenge closed. Thereafter, the contributions from the project partners and challenge participants were processed to present the results on this workshop.



**Fig. 1.** The figure gives an overview how the documents from the first CALBC challenge have been used for the second CALBC challenge. In addition, the corpus for the second CALBC challenge has been filled with documents from the publicly available gold standard corpora.

During the workshop we will follow presentations of different kinds. First, the challenge participants will present their annotation solutions to better understand the outcome of the CALBC II challenge. Second, the CALBC project partners will present the approaches taken to measure performances of the annotation solutions. This includes the approaches to assess the value of different SSCs: a number of gold standard corpora were collected and converted to leXML format to compare the performances of the annotation solutions against the GSCs and the SSCs at the same time [3]. Also, the gold standard corpora have been integrated into the corpus for the second CALBC challenge and will be used to gauge the performance of the participants' annotation solutions and harmonised set. Third, additional talks are concerned with the usage of the SSC and the exploitation of the SSC as part of the integration of the corpus into the bioinformatics data infrastructure and into the Semantic Web.

Finally, for the last day of the workshop the project partners have prepared input to discuss the prospects of the CALBC approach. The essential questions are concerned with the achievements regarding the interoperability and normalisation of semantic resources and the integration of different semantic types in the same corpus overall. Other questions demand, what use cases from involving Semantic Web technology would profit from the integration of the scientific literature and whether the CALBC approach scales in such a way that all scientific publications in the biomedical domain would be annotated with automatic means in the future. Also, the CALBC approach could be used to process multi-lingual corpora and to produce large-scale multi-lingual terminologies.

## Acknowledgements

This work is funded by the EU FP7 Support Action grant 231727 (ICT 2007.4.2).

## References

1. Rebold-Schuhmann,D., A. Jimeno Yepes, E. Van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, E. Beisswanger, and U. Hahn. (2010) "CALBC Silver Standard Corpus." J Bioinform Comput Biol. 2010 Feb;8(1):163-79.
2. Rebold-Schuhmann,D., A.J. Jimeno Yepes, E.M. van Mulligen, N. Kang, J. Kors, D. Milward, P. Corbett, E. Buyko, K. Tomanek, E. Beisswanger, and U. Hahn. (2010) The CALBC Silver Standard Corpus for Biomedical Named Entities: A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers. Proc. LREC 2010.
3. Rebold-Schuhmann,D., Kirsch,H., and Nenadic,G. (2006) leXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules. SIG BioLink, ISMB 2006, Fortaleza, Brasil.

## Methods for Matching of Annotations, Harmonisation and Evaluation

Jan Kors

Erasmus Medical Center, the Netherlands

In this talk, several methodological issues related to the CALBC project will be examined. First, we will review several approaches for matching the annotations of two different systems. Annotation matching is a necessary step in the computation of performance measures such as precision and recall. Second, we will present a number of harmonisation approaches that were studied in the CALBC project. We explored entity-, token-, and character-based voting schemes to harmonise the annotations of different named entity taggers and arrive at a silver standard corpus (SSC). These harmonisation methods focus on the named entity recognition task in CALBC (i.e., boundary recognition of entities and semantic group assignment); we will also describe the concept identification task (i.e., the assignment of concept identifiers to the recognised entities), and how concept identification in the harmonised set is accomplished. Finally, gold standard corpora (GSCs) provide a means to assess the performance of the silver standard for named entity recognition. We will briefly describe the different GSCs considered in the CALBC project.

## Challenge I vs. Challenge II: Set-up, Participation Data and Feedback

Udo Hahn<sup>1</sup>, Kerstin Hornbostel<sup>1</sup> and Dietrich Rebholz-Schuhmann<sup>2</sup>

<sup>1</sup> Jena University Language & Information Engineering (JULIE) Lab  
Friedrich-Schiller-Universität Jena,  
Jena, Germany  
<http://www.julielab.de>

<sup>2</sup> EMBL Outstation Hinxton  
European Bioinformatics Institute  
Hinxton, Cambridge, U.K.  
<http://www.ebi.ac.uk/Rebholz/>

We will compare the first and the second CALBC Challenge and report on the set-up of both challenges as well as fundamental participation data. To generate such data we have run an opinion survey which helped us collect data on the kind of institutions which participated, on the reasons for and benefits from participation in the challenges. We also solicited more opinion-style data such as the problems participants encountered and recommendations participants suggest for future annotation campaigns.

## Building the SSC and evaluation of the annotation systems against the SSC

Ian Lewin<sup>1</sup> Jan Kors<sup>2</sup> and Dietrich Rebholz-Schuhmann<sup>1</sup>

<sup>1</sup>EMBL Outstation Hinxton  
European Bioinformatics Institute  
Hinxton, Cambridge, U.K.  
<http://www.ebi.ac.uk/Rebholz/>

<sup>2</sup>Erasmus Medical Center, the Netherlands

We discuss the generation of consensus "Silver Standards" from the set of annotations supplied by a) the CALBC project partners and b) the CALBC challenge participants. The main operation used is a simple voting scheme over the pairs of consecutive characters which participants annotate as being within one entity. Then, a threshold is applied to the number of votes cast. The resulting silver standards can be used to measure performances and gain a picture of the shape of the data.

## Evaluation of the SSC against the Gold Standard Corpora

Şenay Kafkas, Ian lewin and Dietrich Rebholz-Schuhmann

European Bioinformatics institute, Wellcome Trust Genome Campus

Hinxton, CB10 1SD, U.K.

{kafkas, lewin, rebholz}@ebi.ac.uk

<http://ebi.ac.uk/rebholz>

The Silver Standard Corpus (SSC) is generated with automatic alignment methods through the reconciliation of annotations from different annotated corpora. The quality of the SSC has to be assessed against available Gold Standard Corpora (GSC) to judge the outcome of the reconciliation. In principle, the SSC should deliver the same annotations as required for the GSCs, since the GSC implement the standards for the respective entity types.

In the current phase of the CALBC project, we have generated three types of SSCs and evaluated them against three publicly available GSC for protein/gene (PRGE) annotations. The GSC which have been used in our experiments are the following corpora: PennBioIE-Oncology [1], BioCreative-II Gene Mention [2] and FSU-PRGE. We report on the performances in terms of F-measure using 98% cosine similarity scoring.

The first SSC has been produced by harmonizing the annotations that have been generated from the annotation systems of the project partners (EBI, LM, FSU-Julie, EMC) only. Consensus of at least two out of four systems was required for the inclusion of the annotations in the SSC. Two votes maximized the average F-score of the four systems against the SSC. This SSC is called the Partner-SSC. Our evaluation of the Partner-SSC against the GSCs shows that on average its performance is better than any annotation solution from any of the partners' solutions.

It was our hypothesis that we should be able to improve the quality of the SSC by using an increased number of systems delivering annotations, since these different systems would have different and complementary characteristics regarding the annotation of the corpus. Therefore, we have generated a second type of SSC based on the consensus of all submissions (4 from challenge participants and 14 from project partners) with different voting schemes. This SSC is called the All-SSC and requires an agreement of 3 out of 18 systems in the harmonisation procedure. The performance assessment of the All-SSC against the GSCs shows that that on average the All-SSC can outperform the majority of the systems.

The third type of SSC is called the Selected-SSC and has been generated using only one submission from each participant [for each semantic group]. There are maximum of 14 systems contributing to the consensus. Participant systems have been selected based on their performances against the All-SSC. We examine the effect of different majority voting schemes on the SSC performance. The

Selected-SSC outperforms all partners' systems and the majority of the selected systems when a threshold value 3 for the agreement in the harmonisation procedure is used.

We also assessed the improvements of the different SSCs against the GSC to identify incremental improvements. We compared the All-SSC against the Selected-SSC and the Partner-SSC using the GSC as the performance benchmark. Results show that the All-SSC performs better than the Selected-SSC and the Partner-SSC, while the Selected-SSC on average performs better than the Partner-SSC.

Finally, we discuss the caveats and limitations of the measured results. For example, we do know that none of the solutions contributing to the Partner-SSC have been trained on the test datasets of the GSCs, but this cannot be fully excluded for the other contributing systems of the All-SSC.

In the future, we plan to expand our work on the SSC quality assessment against GSC to include other semantic groups. We also will analyze the effect of iterative participant solution increment on the SSC's performance. Results will help us to determine the best harmonization procedure that will be applied for SSC-III generation.

## **References**

- [1] Kulick S, Bies A, Liberman M, Mandel M, McDonald R, Palmer M, Schein A, Ungar L. Integrated annotation for biomedical information extraction. Proc BioLINK 2004, Association for Computational Linguistics. 2004. pp. 61–68.
- [2] Tanabe L, Xie N, Thom LH, Matten W, Wilbur WJ. GENETAG: a tagged corpus for gene/protein named entity recognition BMC Bioinformatics 2005;6(Suppl 1):S3

## A CRF-based approach to harmonize heterogeneous gene/protein annotations

David Campos<sup>1</sup>, Dietrich Rebholz-Schuhmann<sup>2</sup>, Sérgio Matos<sup>1</sup> and José Luís Oliveira<sup>1</sup>

{david.campos,aleixomatos,jlo}@ua.pt; rebholz@ebi.ac.uk

<sup>1</sup>University of Aveiro, DETI/IEETA  
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal

<sup>2</sup> European Bioinformatics Institute  
Wellcome Trust Genome Campus  
Hinxton, Cambridge CB10 1SD, UK

Named Entity Recognition (NER) is a task of Information Extraction (IE) that intends to extract names of specific entities. During the last years, there was a large research interest on this field, with the development of several systems using the most different approaches and techniques. The best results were achieved with solutions that combine annotations from several systems, exploiting their different characteristics and outcomes. However, the development of those solutions is not straightforward, due to the variability of the annotations provided by systems, human annotators and biological domains. We have used a machine learning-based solution to tackle the presented challenges, using Conditional Random Fields (CRFs) and several Gold Standard Corpora (GSC). At the end, the goal is to annotate any MEDLINE abstract with high accuracy.

The first version of the CALBC corpus results from a combination of the annotations provided by four well known NER and normalization systems, provided by EBI (European Bioinformatics Institute), EMC (Erasmus Medical Center), FSU (Friedrich-Schiller-Universitat) and LM (Linguamatics Lda.). The used harmonization technique was based on majority voting, requiring the vote of two systems to consider a chunk of text as an entity name. In case of overlapping, the shortest annotation was selected. This solution already presented a significantly performance improvement, in comparison with the results obtained individually by each system. The main goal of this work was to improve the harmonization results, providing a corpus-independent solution.

In order to accomplish this task, we have used four well-known GSC to train and test the machine learning model: GENETAG [1], JNLPBA [2], PennBioIE [3], and FSUPRGE. Overall, these corpora provided almost nine thousand abstracts (6566 for training and 2242 for testing). Afterwards, we defined a 2<sup>nd</sup> order CRF using the tokens as features and a f-1,1g window to model local context. With this solution, we can fix annotation boundaries based on human curated knowledge from several domains. Additionally, we can also create new annotations and ignore previous ones. However, to make this solution compatible with normalization systems, we have created an alternative version that does not generate new annotations, apart from the ones provided by the systems. Thus, we allow two distinct operation modes: a) creating new annotations; and b) without creating new annotations.

To check the performance, we have trained the CRF in the merged training corpus, and tested it in the merged test corpus, simulating a "MEDLINE environment". Moreover, we have also used the method on each test corpus, evaluating the behaviour on each specific domain. The obtained results (Figure 1) show that both CRF-based solutions present significant improvements (F1=66.1% and F1=69.5%) in comparison with the existing approach (F1=52.7%). However, the "vote 2" solution shows better precision in some individual corpora, namely PennBioIE and FSUPRGE. Comparing the

two CRF solutions, the technique that creates new annotations significantly improves the recall with a small drop of precision, presenting the best F-measures in all tests.

In this work we have presented an harmonization solution that uses knowledge from several GSC to fix the annotations' boundaries and, if desired, create new annotations. The ~70% F-measure achieved on the merged test corpus represents a highly positive outcome towards an automatic annotation of biomedical literature.

		<b>GENETAG</b>	<b>JNLPBA</b>	<b>PennBioIE</b>	<b>FSUPRGE</b>	<b>Merged</b>
<b>Vote 2</b>	R	34.65%	38.02%	62.32%	53.52%	43.20%
	P	62.56%	49.45%	↑ 86.54%	↑ 83.70%	67.72%
	<b>F1</b>	<b>44.60%</b>	<b>42.99%</b>	<b>72.46%</b>	<b>65.29%</b>	<b>52.75%</b>
<b>CRF without new annotations</b>	R	44.17%	52.15%	69.72%	65.10%	60.01%
	P	60.56%	60.22%	83.63%	79.19%	↑ 73.51%
	<b>F1</b>	<b>51.08%</b>	<b>55.89%</b>	<b>76.04%</b>	<b>71.45%</b>	<b>66.08%</b>
		(+6.48%)	(+12.91%)	(+3.58%)	(+6.16%)	(+13.33%)
<b>CRF with new annotations</b>	R ●	53.62%	● 67.93%	● 73.68%	● 69.20%	● 67.08%
	P ↑	64.74%	↑ 65.32%	81.98%	74.13%	72.13%
	<b>F1</b>	<b>58.66%</b>	<b>66.60%</b>	<b>77.60%</b>	<b>71.58%</b>	<b>69.51%</b>
	(+14.06%)	(+23.61%)	(+5.14%)	(+6.29%)	(+16.76%)	

**Fig1.** Results from each test corpus considering exact matching alignment. The shaded boxes highlight the F-measure of each system, where the bold font indicated the best solution. The arrow (↑) indicated the solution with highest Precision, and the circle (●) the one with better Recall.

## References

1. Tanabe, L., Xie, N., Thom, L., Matten, W., Wilbur, W.: GENETAG: a tagged corpus for gene/protein named entity recognition. BMC bioinformatics 6 (2005) S3
2. Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Association for Computational Linguistics (2004) 70-75
3. Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A., Ungar, L., Winters, S., White, P.: Integrated annotation for biomedical information extraction. In: Proceedings of the Human Language Technology Conference and the Annual Meeting of the North American Chapter. (2004)

## Use Cases for the Silver Standard Corpora

David Milward and Peter Corbett

Linguamatics, St John's Innovation Centre, Cowley Road, Cambridge, CB4 0WS, UK

The Silver Standard Approach provides the opportunity to create annotated corpora at a very large scale, based on a consensus between different automated systems. It is also possible to provide configurable Silver Standard Corpora where the balance between recall and precision can be adjusted.

Silver Standard Corpora (SSC) have wide applicability outside the CALBC challenges. In this talk we examine some reasons for using Silver rather than Gold standards, and look at various applications including:

1. improving named-entity recognitions systems e.g. by looking at false positives or false negatives relative to a configurable SSC
2. comparing terminologies e.g. contrasting coverage between rare and common term
3. machine learning from a SSC
4. relationship extraction from an SSC containing marked-up entities from different groups

## Normalisation of lexical entities: Jochem, LexEBI, cross-comparisons

Dietrich Rebholz-Schuhmann<sup>1</sup>, Caroline Friteyre<sup>1</sup>, Abhishek Dixit<sup>1</sup>, Jan Kors<sup>2</sup>,  
Erik van Mulligen<sup>2</sup>, Jee-Hyub Kim<sup>1</sup>, Ian Lewin<sup>1</sup>

<sup>1</sup>European Bioinformatics Institute  
Wellcome Trust Genome Campus  
Hinxton, Cambridge, CB10 1SD, U.K.  
{rebholz,lewin,jhkim}@ebi.ac.uk

<sup>2</sup>Erasmus University Medical Center Rotterdam  
's-Gravendijkwal 230  
3015 CE, Rotterdam  
[j.kors@erasmusmc.nl](mailto:j.kors@erasmusmc.nl), [mulligen@gmail.com](mailto:mulligen@gmail.com)

### Introduction

Terminological resources are important for information retrieval (IR) and information extraction (IE). Standardisation of the terminological resources enables better comparison of IR and IE methods and lead to long-term benefits such as the integration of the literature into the bioinformatics IT infrastructure [1, 2]. We distinguish different kinds of terminological resources: controlled vocabularies, taxonomies, and ontologies, but also biomedical databases such as UniProtKb which deliver annotations to named entities [3].

In the biomedical domain a lexical resource has to cover a number of semantic types that all are frequently used and relevant to data integration and data analysis. The most important semantic types are genes/proteins, chemical entities, species, diseases and others. Lexical resources have been proposed for the medical domain, i.e. resources such as the MetaThesaurus and the collection of resource in UMLS [4,5]. Other resources have developed into a standard in the biological domain such as the UniProtKb, InterPro or BioThesaurus for gene and protein terms, the NCBI taxonomy for species and ChEBI for chemical entities [3,6,7,8,9]. It is a common theme that a researcher in the text mining domain would collect terms from the described resource, extract the terminology and would use it in its own data analysis [10,11,12]. A common format to the different resources and the integration of the terms across the resources would contribute to the interpretation of research results, if the resource has been used in the research [13]. LexEBI contains the full scope of biomedical-chemical relevant terms. The terminological resource provides valuable services to the text mining community.

### Extraction of terms from the primary resource

The BioThesaurus versions 6.0 and 7.0 have been processed to extract the concept identifiers, terms, and the term variants. Non-sensical terms such “hypothetical gene”, “putative gene”, “probable gene”, “possible gene” and single numbers have been removed. The concept identifier of each term from each resource has been kept for later reference purposes. All term variants for a given concept have been organised in a single cluster, where the preferred term forms the label of the cluster. In the same way, the terms from ChEBI, JoChem, IntEnz, and the NCBI taxonomy have been processed [14].

The UMLS terminological resource has been filtered using the tag assignments to the terms. Terms belonging to the following categories have been extracted: (1) Antibiotic and neuroreactive substances, (2) Biologically active substances, (3) Enzymes, (4) Lipids and Carbohydrates, (5) pharmacological active substances, and (6) vitamins and hormones. Other categories such as disease and disorder and immunological factors have been ignored.

## Representations of the terms the lexical resource

LexEBI uses an XML format for the representation and storage of the terminological resource. Explicit reference are implemented to the preferred term, the term variants, concept ids, term frequency in the British National Corpus, in Medline, and the frequency of the term variants. An additional table makes reference to the nestedness of the terms in the resources.

The structure of the a single entry is shown as a piece of XML code:

```
<Cluster clsId="CHEBI_CHEBI:32" semType="CHEBI">
<Entry entryId="CHEBI_CHEBI:32_1" baseForm="(+) -N-methylconiine" type="PREFERRED">
  <PosDC posName="POS" pos="N"/>
  <SourceDC sourceName="CHEBI" sourceId="CHEBI:32"/>
  <Variant WRITTENFORM="(+) -N-Methylconiine" type="orthographic"/>
  <Variant WRITTENFORM="(2S)-1-methyl-2-propylpiperidine" type="orthographic"/>
  <Variant WRITTENFORM="Methylconiine" type="orthographic"/>
  <Variant WRITTENFORM="C9H19N" type="orthographic"/>
  <Variant WRITTENFORM="CCC[C@H]1CCCCN1C" type="orthographic"/>
  <Variant WRITTENFORM="InChI=1/C9H19N/c1-3-6-9-7-4-5-8-10(9)2/h9H,3-8H2,1-2H3/t9-/m0/s1" type="orthographic"/>
  <DC att="KEGG COMPOUND accession (KEGG COMPOUND" val="C10159)"/>
  <DC att="CAS Registry Number (KEGG COMPOUND" val="35305-13-6)"/>
  <DC att="CAS Registry Number (ChemIDplus" val="35305-13-6)"/>
  <DC att="Beilstein Registry Number (Beilstein" val="79936)"/>
  <Relation type="has_parent_hydrate" target="CHEBI_CHEBI:28322"/>
</Entry>
</Cluster>
```

## Results

LexEBI is a compilation of terms from different biological, chemical and chemical resources. The lexical resource fulfills requirements for text mining and term normalisation in the context of data retrieval.

### *Distribution of terms in LexEBI*

The terminological resource LexEBI contains about 2,729,134 clusters that make reference to a label (aka. Baseform), to 13,598,649 term variants (including redundancy) and to 5,791,531 unique terms in total, where redundancy between resources has not been removed.

For the terminology linked to genes and proteins, two different resources of the same origin were used: Biothesaurus 6.0 (here called "GP-6.0") and the next version, i.e. Biothesaurus 7.0 (called "GP-7.0"). GP-6.0 gives access to 1,564,436 terms and GP-7.0 to 1,726,853 terms. 1,444,247 are shared between both resources using exact matching (ref. to tbl. 1). This results to 92.3% of the unique terms in GP-6.0 and to 83.6% in GP-7.0, showing that the new version contains a larger number of terms. In total, GP-7.0 contains 27,536 additional clusters or baseforms that account for 162,417 additional unique terms and 643,260 overall term variants (including redundancy).

The terminological resources for genes and proteins show a high number of term variants per cluster, i.e. 8.76 and 7.94 for GP-7.0 and GP-6.0, respectively, and also high numbers of term variants for chemical entities, i.e. 7.07 and 5.82 for JoChem and for ChEBI. Term variability is only of minor importance for species terms (1.31) and for the other resources.

		# Labels	# Variants	Total	Total / Labels	# Unique terms	Uniq. T. / Labels
Gene/ Prot.	GP 7.0	516,113	4,005,040	<b>4,521,153</b>	8.76	1,726,853	3.35
	GP 6.0	488,577	3,389,316	<b>3,877,893</b>	7.94	1,564,436	3.20
	Interpro	20,671	0	<b>20,671</b>	1.00	20,671	1.00
	Enzymes	4,905	8,082	<b>12,987</b>	2.65	12,377	2.52
Chemicals	Jochem	278,578	1,691,980	<b>1,970,558</b>	7.07	1,527,752	5.48
	ChEBI	19,645	94,748	<b>114,393</b>	5.82	101,307	5.16
	ChEBI (all)	549,838	1,187,322	<b>1,737,160</b>	3.16	863,227	1.57
Other	Diseases	56,010	165,581	<b>221,591</b>	3.96	186,555	3.33
	Species	643,280	199,130	<b>842,410</b>	1.31	838,135	1.30
UMLS	Neoplast.	4,718	6,488	<b>11,206</b>	2.38	11,196	2.37
	Bio. Act.	54,148	87,209	<b>141,357</b>	2.61	141,121	2.61
	Enzymes	26,065	56,332	<b>82,397</b>	3.16	82,033	3.15
	Lipid, Carb.	11,518	9,770	<b>21,288</b>	1.85	21,281	1.85
	Pharm. Act.	104,201	123,840	<b>228,041</b>	2.19	227,799	2.19
	Vit., Horm.	6,877	10,258	<b>17,135</b>	2.49	17,007	2.47
	<b>Total</b>	<b>2,765,499</b>	<b>10,940,348</b>	<b>13,705,847</b>		<b>7,240,443</b>	
	Total PGN	1,030,266	7,402,438	8,432,704		3,324,337	
	Total Chem	828,416	2,879,302	3,707,718		2,390,979	

Tbl. 1: The table shows the distribution of terms from LexEBI sorted according to the resource that delivered the terms. The biggest portions of the terms contained in LexEBI result from BioThesaurus (GP 6.0 and GP 7.0), from Jochem and ChEBI and from the NCBI taxonomy.

### Cross-Reference within LexEBI

Several resources have been compared against GP-6.0 and GP-7.0. 150,104 enzyme labels from the IntEnz database are already covered in GP-6.0 and this number increases to 173,994 for the GP-7.0 (data not shown). Morphological variability only adds little to the identification of terms (157,099 and 180,829), whereas nestedness adds a bigger portion to the number of matched terms leading to now 178,155 and 202,484 terms for exact matching and 200,921 and 224,877 terms for variable matching of contained terms. By contrast, terms from Interpro occur in the GP-6.0 and GP-7.0 at lower numbers, 88,613 and 93,979 for both resources respectively, but the number increases to more than twice the size, if variable matching or nestedness is considered (see fig. 1). This shows that the generic terms from Interpro form parts of the terms in GP-6.0 and GP-7.0 in contrast to the terms denoting enzymes.

### Conclusion

LexEBI provides access to a large set of terms that have been organised in a standardised format. The resource will enable researchers in the biomedical text mining community to better standardise their results and to cross-compare the outcome of their text mining solutions.

## References

1. Sasaki Y, McNaught J, Ananiadou S: The value of an in-domain lexicon in genomics qa. *J Bioinform Comput Biol* 2010, 8(1):147-161.
2. Nenadic G, Spasic I, Ananiadou S: Terminology-Driven Mining of Biomedical Literature. *Bioinformatics* 2003, 19(8):938-943.
3. UniProt Consortium: The Universal Protein Resource (UniProt) in 2010. *Nucleic Acids Res* 2010, 38(Database issue):D142-D148.
4. Bodenreider O: The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res* 2004, 32(Database issue):D267-270.
5. Browne AC, Divita G, Aronson AR, McCray AT: UMLS language and vocabulary tools. In: *Proceedings of the AMIA Annual Symposium*. Washington DC, USA; 2003: 798.
6. Degtyarenko K, Hastings J, de Matos P, Ennis M: ChEBI: an open bioinformatics and cheminformatics resource. *Curr Protoc Bioinformatics* 2009, Chapter 14:Unit 14 19.
7. de Matos P, Alcantara R, Dekker A, Ennis M, Hastings J, Haug K, Spiteri I, Turner S, Steinbeck C: Chemical Entities of Biological Interest: an update. *Nucleic Acids Res*, 38(Database issue):D249-254.
8. Liu H, Hu ZZ, Zhang J, Wu C: BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 2006, 22(1):103-105.
9. InterPro [<http://www.ebi.ac.uk/interpro/>]
10. Krallinger M, Valencia A, Hirschman L: Linking genes to literature: text mining, information extraction, and retrieval applications for biology. *Genome Biol* 2008, 9 Suppl 2:S8.
11. Tsuruoka Y, McNaught J, Tsujii J, Ananiadou S: Learning string similarity measures for gene/protein name dictionary look-up using logistic regression. *Bioinformatics* 2007, 23(20):2768-2774.
12. Tsuruoka Y, McNaught J, Ananiadou S: Normalizing biomedical terms by minimizing ambiguity and variability. *BMC Bioinformatics* 2008, 9 Suppl 3:S2.
13. Pezik P, Jimeno-Yepes A, Lee V, Rebholz-Schuhmann D: Static dictionary features for term polysemy identification. In: *Proceedings of the LREC Workshop on Building and Evaluating Resources for Biomedical Text Mining*. 2008: 35-41.
14. Hettne KM, Williams AJ, van Mulligen EM, Kleinjans J, Tkachenko V, Kors JA: Automatic vs. manual curation of a multi-source chemical dictionary: the impact on text mining. *J Cheminform*, 2(1):3.

## Normalisation of the CALBC silver standard corpus using normalised terminological resources

E. Jimenez-Ruiz<sup>1</sup>, I. Lewin<sup>2</sup>, and D. Rebholz-Schuhmann<sup>2</sup>

<sup>1</sup>Oxford University Computing Laboratory, ernesto@comlab.ox.ac.uk

<sup>2</sup>European Bioinformatics Institute, lewin,rebholz@ebi.ac.uk

### Normalised terminological resources

CALBC systems use different terminological sources in order to provide concept identifiers of the found entities. For example, when annotating chemical entities, project partners' systems provide identifiers from the following sources:

- Linguamatics: Mesh tree (30%), NCI (37%) and ChEBI (33%)
- EBI : Chemlist
- ErasmusMC: UniProt (47%), Entrez (0.02%), ChemidPlus (18%), ChEBI (31%), MeSH (2%), Pubchem Substance (2%), Pubchem Compound (0.01%), Chemical Abstracts Service Registry Number (0.3%).
- Jena: UMLS (56%), UniProt (44%)

We have exploited available terminological resources such as UMLS [1], JoChem [2], Biothesaurus [3], PIR [4] or InterPro [5] (see Figure 1) to provide correspondences between the given concept identifiers. Note that extracted correspondences have been classified in three types: synonyms (e.g. UMLS-Chemlist mappings), family groups (e.g. UniProt-InterPro mappings) and taxonomic relationships (e.g. direct broader and narrower relationships within UMLS).

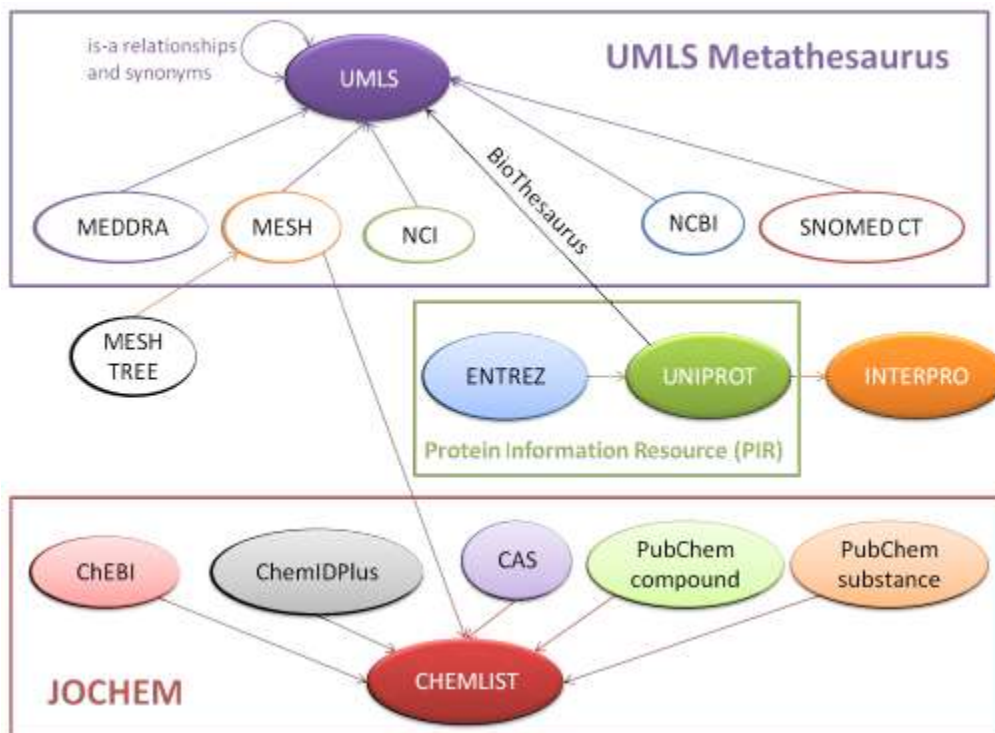


Fig1. Available correspondences between terminological resources

## Normalization of CALBC corpus I

We have evaluated the concept identifier agreement among project partners' systems over an excerpt of the CALBC corpus containing 2113 abstracts.

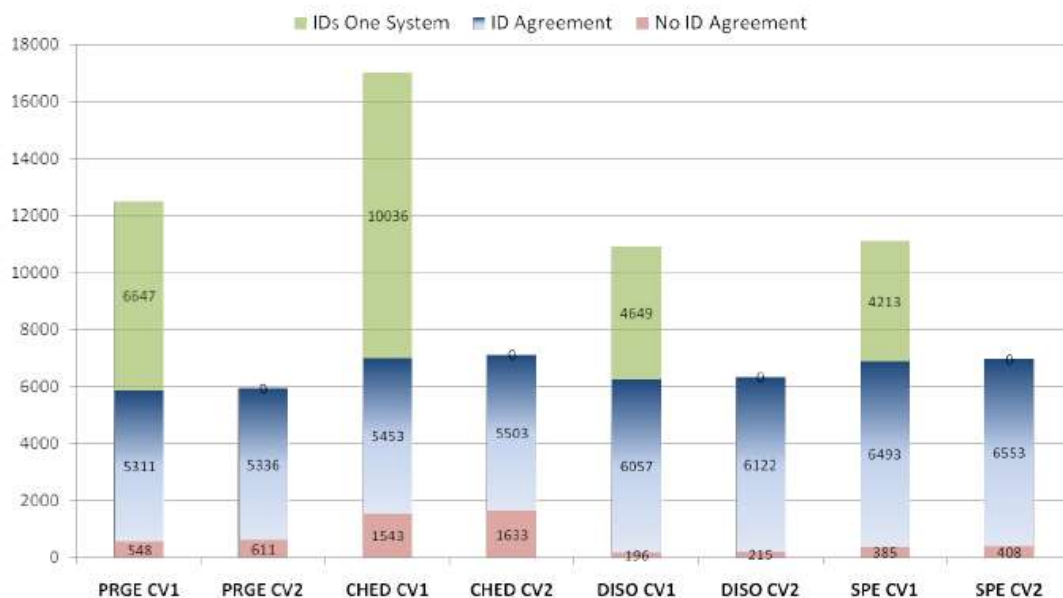


Fig. 2. Concept identifier agreement for each semantic group. CV1 and CV2 stands for CharVote-1 and CharVote-2, respectively, and refer to the character-based multiple alignment using 1-vote and 2-vote agreement.

Figure 2 summarizes the concept identifier agreement achieved for each semantic group. For CharVote-1 an important number (above 50% for protein-gene and chemical entities) of single assignments (i.e. only one of the systems provided an annotation) of concepts ids can be found (green part). For CharVote-23, the boundaries with concept id agreement almost reach the 97% of the cases for diseases, it is over 94% for species and around 90% for protein-genes. For chemical entities, however, the agreement only reaches the 77%. This is mainly due to the complexity in the identification of these kinds of entities and also to the diversity of used sources by the partners.

It is worth mentioning that, when id disagreement, in the 86% of cases only two partners were able to annotate the entity, and only in the 17% of cases there was an agreement in the used source. This fact shows that the set of correspondences between sources is not totally complete. For example, the entity varicella-zoster virus is annotated by EBI as UMLS:C0740380 whereas Linguamatics provides MedDRA:10046980 as identifier. In other cases, an inherent disagreement in the identified entity was also detected. For example, the entity Coxsackie virus is annotated by EMC system as UMLS:C0010246:Coxsackie virus whereas JENA system only provides UMLS:C0003060:Virus as annotation. In this case, a direct broader/narrower relationship between concepts C0010246 and C0003060 was not found within UMLS either.

Figure 3 represent the agreement and ambiguity (i.e. boundaries with two or more associated semantic groups) when pairs of semantic groups are considered. Note that we have split the cases with no id agreement in order to detect those boundaries with also a disagreement in the given semantic group. The ambiguity for chemicals and protein-genes is specially high and is around the 55% of the cases with id disagreement.

<sup>3</sup>With CharVote-2 single annotations are removed since 2-vote agreement is required for every character an transition in the annotations.

This is not surprising since chemical names are usually contained in protein-gene names. The ambiguity for the other semantic group pairs is around 30% for PRGE-DISO, CHED-DISO and DISO-SPE, and under 20% for PRGE-SPE and CHED-SPE.

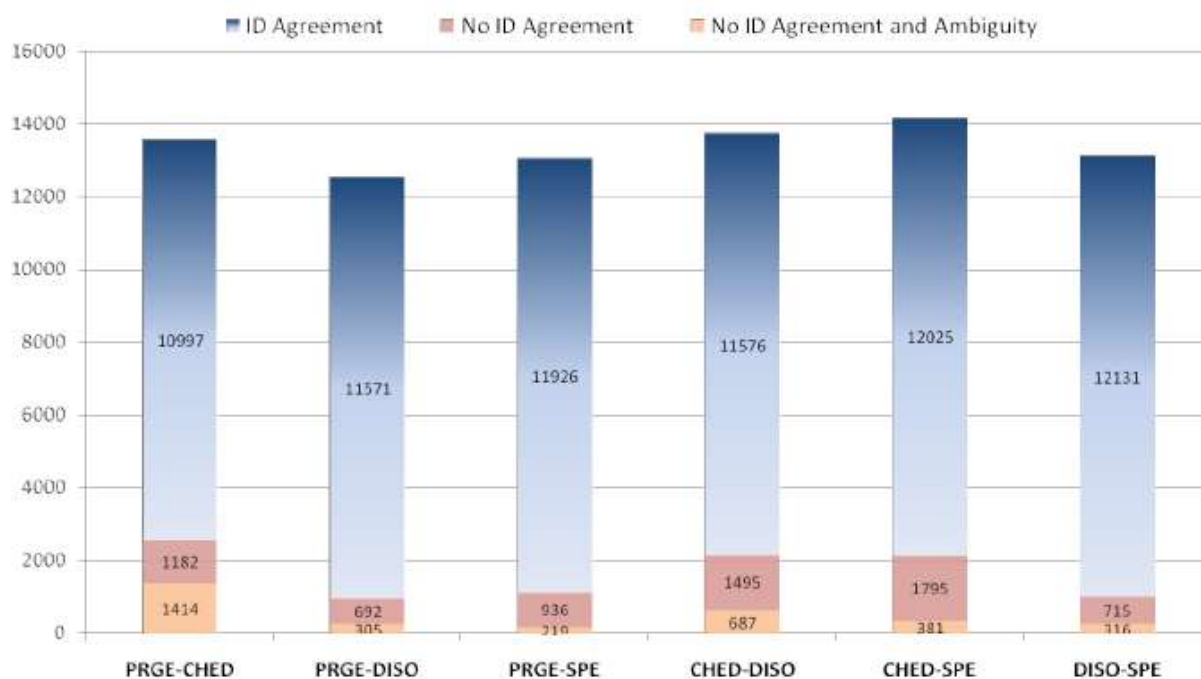


Fig. 3. Agreement and ambiguity for pairs of semantic groups (char vote 2)  
References

## References

1. Bodenreider O. The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic acids research* 32, 2004.
2. Hettne KM et al. A Dictionary to Identify Small Molecules and Drugs in Free Text. *Bioinformatics*. 25(22):2983-91, 2009.
3. Liu HF et al. BioThesaurus: a web-based thesaurus of protein and gene names. *Bioinformatics* 22:103-105, 2006.
4. Barker WC et al. The Protein Information Resource (PIR). *Nucleic Acids research*, 28, 2000.
5. Hunter S et al. InterPro: the integrative protein signature database. *Nucleic Acids research*, 37,2009.

## Integration of literature with biomedical data resources - the SESL project

Dietrich Rebholz-Schuhmann<sup>1</sup>, Christoph Grabmüller<sup>1</sup>,  
Samuel Croset<sup>1</sup>, Silvestras Kavaliauskas<sup>1</sup>

<sup>1</sup>European Bioinformatics Institute  
Wellcome Trust Genome Campus  
Hinxton, Cambridge, CB10 1SD, U.K.  
{rebholz,grabmuel,samuel.croset,kavalia}@ebi.ac.uk

The SESL pilot project explores the technical feasibility for federated querying across full text literature and bioinformatics databases. Five Life Science and Pharmaceutical companies have collaborated with four publishers and the Rebholz group (EMBL-EBI) to extract selected data from bioinformatics databases (Uniprot, OMIM and ArrayExpress) and full text literature with focus on human diseases related to Type 2 diabetes mellitus. Gene to disease related assertions have been delivered through a single point of query to the scientist users.

The pilot implements the integration of content from public resources and extracted information from the scientific literature into a shared infrastructure based on Semantic Web technology. The SPARQL endpoint is hosted at the EBI and can be accessed remotely through SPARQL queries, a Web browser based graphical user interface or through a SOAP Web services client. The project delivers a preliminary set of standards describing the minimal infrastructure necessary to support a biology brokering service and the provision of a prototype instance of that infrastructure as a public demonstrator.

### **Acknowledgements:**

The SESL project was funded by the Pistoia Alliance (<http://www.pistoiaalliance.org/>). Core assistance to the project was contributed by Pfizer Inc. (Ian Harrow), UniLever (Wendy Filsell, Ian Stotts), AstraZeneca (Mike Westaway, Ian Dix), and GlaxoSmithKline (Peter Woollard). Scientific literature was contributed by Nature Publishing Group (David Hoole), Oxford University Press (Richard o'Beirne, Claire Bird), Royal Society of Chemistry (Richard Kidd), and Elsevier (Jabe Wilson).

## Abstracts – CALBC Challenge II: system descriptions – project partners

### Annotation of the CALBC corpus using biomedical terminological resources (Whatizit)

Dietrich Rebholz-Schuhmann<sup>1</sup>, Şenay Kafkas<sup>1</sup>, Chen Li<sup>1</sup>

<sup>1</sup>European Bioinformatics Institute,  
Wellcome Trust Genome Campus,  
Hinxton, Cambridge, CB10 1SD, U.K.  
rebholz@ebi.ac.uk

Named entity recognition (NER) is a complex task. It requires exploiting terminological resources, defining the extraction methods for the identification of the terms from the literature and ideally in the last step, normalising the entities to the correct entry of the reference data resource. In principle, dictionary-based or rule-based solutions can be applied to identify the mention of the entity or as an alternative a machine-learning approach can be trained on an annotated corpus to reproduce the annotations on a new set of documents. In the latter case, the NER solution profits from contextual information but also requires a significant set of annotated documents.

The Whatizit Web services provide access to an IT infrastructure that analyses text delivered by the user or retrieved from EBI's Medline installation. The user profits from modules for named entity recognition of selected semantic types or from combinations of such modules. These services satisfy the need for terminology driven feature extraction from text for document classification and relation extraction. Furthermore, the modules automatically integrate links to database concepts to offer additional support to readers (for example like CiteXplore or [www.hubmed.org](http://www.hubmed.org)) and will be extended in the future with novel modules for information extraction. All modules are implemented in Java partly based on special libraries for the matching of large terminology sets and one installation of the infrastructure uses the leXML formatting guidelines (Kirsch et al., 2006; Rebholz-Schuhmann et al., 2006b). Terms are matched to the text taking morphological variability into consideration.

The following modules have been applied to the CALBC corpus:

**whatizitSwissprot:** The annotation of proteins is based on the identification of their names in the text considering morphological variability (Kirsch et al., 2006). Ambiguous acronyms representing proteins and general English terms are assigned to a protein, if the long-form of the acronym is mentioned in the text or on frequency parameters of the term in general English based on the British National Corpus (Rebholz-Schuhmann et al., 2006a; Rebholz-Schuhmann et al., 2007).

**whatizitDiseaseUmlsDict:** This module identifies disease terms using a controlled vocabulary (CV) extracted from UMLS (Jimeno Yepes et al., 2008).

**whatizitChebiDict:** searches for chemical entities based on the terminology from ChEBI (Rebholz-Schuhmann et al., 2008).

**whatizitiOrganisms:** The terminology used for this module has been extracted from the NCBI taxonomy (Rebholz-Schuhmann et al., 2008).

Alternative modules have been tested to measure their performance against the Gold Standard corpora and against the Silver Standard Corpora. For example, Abner has been measured and an annotation solution that combines false positive filtering of annotations based on a machine learning NER solution (called "Chang2") with a dictionary-based solution (e.g., WhatizitiSwissProt).

All terminological resources are also available from the term repository<sup>1</sup> and the BioLexion (see ELRA).

Other solutions have been applied too, for example a solution that has been optimized for the annotation of human genes/proteins and that has been trained on the BioCreative II gene mention data, but this solution did not show any improvements to the performance in comparison to the "whatizitiSwissprot" module (Pezik et al., 2008). Also a solution for the identification of chemical entities has been applied that has been trained on patent documents (Grego et al., 2009), but showed again inferior performance against the silver standard corpus.

### **Acknowledgements**

This work is funded by the EU FP7 Support Action grant 231727 (ICT 2007.4.2).

### **References**

Grego, T., Pezik, P., Couto, F.M. and Rebholz-Schuhmann, D. (2009) Identification of Chemical Entities in Patent Documents. Springer Verlag Berlin / Heidelberg. Lecture Notes in Computer Science, 5518: 942-949.

Jimeno Yepes, A., Jimenez-Ruiz, E., Lee, V., Gaudan, S., Berlanga, R., and Rebholz-Schuhmann, D. (2008) Assessment of Disease Named Entity Recognition on a Corpus of Annotated Sentences. BMC Bioinformatics 9, no. SUPPL. 3 (2008): Article S3.

Kirsch, H., et al. (2006) Distributed modules for text annotation and IE applied to the biomedical domain. Int. J. Med. Inform. 75(6):496-500.

Pezik, P., Jimeno, A., Lee, V., and Rebholz-Schuhmann, D. (2008) Static Dictionary Features for Term Polysemy Identification. Proc Lang Res Eval, Conf (LREC-2008), workshop on "Building and evaluating resources for biomedical text mining", Marrakech (Morocco), 28-30 May 2008

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., and Jimeno, A. Text Processing through Web Services: Calling Whatiziti. Bioinformatics 24, no. 2 (2008): 296-98.

Rebholz-Schuhmann, D., et al. (2006a) Annotation and Disambiguation of Semantic Types in Biomedical Text: a Cascaded Approach to Named Entity Recognition. Workshop on "Multi-Dimensional Markup in NLP", EACL 2006, Trento, Italy.

---

<sup>1</sup> <http://www.ebi.ac.uk/Rebholz-srv/BioLexicon/biolexicon.html>

Rebholz-Schuhmann,D., Kirsch,H., and Nenadic,G. (2006b) leXML: towards an annotation framework for biomedical semantic types enabling interoperability of text processing modules. SIG BioLink, ISMB 2006, Fortaleza, Brasil.

Rebholz-Schuhmann,D., Kirsch,H. Arregui,M., Gaudan,S., Rynbeek,M., and Stoehr,P. (2007) EBIMed – Text crunching to gather facts for proteins from Medline. *Bioinformatics*, 23(2):e237-44.

## Annotating the second CALBC corpus with Peregrine

Erik M. van Mulligen, Kang Ning, Martijn J. Schuemie, Jan A. Kors

Department of Medical Informatics, Erasmus University Medical Center, Rotterdam, The Netherlands

Peregrine is a fast dictionary-based concept recognition tool [1]. The Peregrine system translates the terms in a dictionary into sequences of tokens. When such a sequence of tokens is found in a document, the term, and thus the concept in the dictionary associated with that term, is recognized in the text. Some tokens are ignored, since they are considered to be non-informative (“of”, “the”, “and”, and “in”). To allow for small lexical variations, gene/protein ‘long forms’ (i.e., terms that contains a space or are longer than six characters) and all UMLS terms are first reduced to their stem using NLM’s Lexical Variant Generator program.

In CALBC we used as a dictionary a combination of the Unified Medical Language System (UMLS), a gene/protein dictionary, and a chemical dictionary of small molecules and drugs (Jochem). For UMLS, we used version 2010AB restricted to level 0 and 1 combined with SNOMED CT. The gene/protein dictionary was constructed by combining information from different genetic databases [2]. Similarly, the chemical dictionary combines information from a large number of chemical resources [3]. To enhance the precision and recall of the dictionaries, we applied a number of rewrite and suppress rules to the terms in the dictionaries [4].

Peregrine incorporates several disambiguation rules:

1. We first determine whether a term is ambiguous. A term is considered ambiguous if it refers to more than one concept in the dictionary, or when it is shorter than six characters and does not contain a number. A non-ambiguous term will automatically be assigned;
2. An ambiguous term will only be assigned if the term is the ‘preferred name’ of the gene or chemical, if a synonym is found in the same document, or if a specific part of a synonym is found (we only used words part of synonyms that occur less than 1000 times in the dictionary as a whole).

These rules were only used for genes, proteins and chemicals, since these are known to have high ambiguity.

All terms recognized by Peregrine in the CALBC corpus were assigned a concept identifier and a semantic type based on the 135 semantic types in the UMLS.

Peregrine can easily be applied to large corpora. The indexing of the corpus for the second CALBC challenge (~850k Medline abstracts) took less than an hour on a moderate server.

1. Schuemie MJ, Jelier R, Kors JA. Peregrine: lightweight gene name normalization by dictionary lookup. Proceedings of the Biocreative 2 workshop; 2007 April 23-25; Madrid. Available from: <http://concept.biosemantics.org/uploads/Biocreative2.pdf>.
2. Schuemie MJ, Mons B, Weeber M, Kors JA. Evaluation of techniques for increasing recall in a dictionary approach to gene and protein name identification. J Biomed Inform 2007;40:316-24.

3. Hettne KM, Stierum RH, Schuemie MJ, Hendriksen PJ, Schijvenaars BJ, van Mulligen EM, Kleinjans J, Kors JA. A dictionary to identify small molecules and drugs in free text. *Bioinformatics* 2009;25:2983-91.
4. Hettne KM, van Mulligen EM, Schuemie MJ, Schijvenaars BJ, Kors JA. Rewriting and suppressing UMLS terms for improved biomedical term identification. *J Biomed Semantics* 2010;1:1.

## Robust Annotation of BioMedical Entities for CALBC using I2E

Himanshu Agrawal and Peter Corbett

Linguamatics, St John's Innovation Centre, Cowley Road, Cambridge, CB4 0WS, UK

For the CALBC Challenge II we annotated two datasets of 175K abstracts and 715K abstracts each with four semantic types CHED (chemicals and diseases), DISO (diseases and disorders), PRGE (proteins and genes) and SPE (species and living beings). For this task, 125K abstracts of Silver Standard Corpus, developed in the first round of the project, was made available as training data.

For annotating the entities we used the interactive text mining system, I2E, without any modification other than a layer of post-processing to create the leXML format. I2E is typically used to extract relationships between entities of bio-medical interest. It uses ontologies developed from terminology resources such as Entrez Gene, MeSH etc. to recognise entities in text. It employs text normalization techniques to improve recall of entities, and performs cosine-based disambiguation to eliminate false positives. It is applicable to structured and unstructured text sources and is commonly used with abstracts, full-text documents, patents etc.

For the CALBC annotation, the source text is indexed with I2E using appropriate ontologies, and the index is queried for the four entity types. The query results are then post-processed to annotate the discovered entities in the text in-line. We have used well-established and widely used ontologies like MeSH, MedDRA, Entrez Gene, NCI thesaurus, SnoMED CT, ChEBI, Uniprot and JoChem. When using the general purpose ontologies like MeSH, NCI etc, we checked different branches of the ontologies to see which gave better results, and excluded the branches which were least relevant. The ontologies are used with the recommended Linguamatics settings which govern specific text normalization techniques such as fuzzy matching of synonyms.

Since the previous round of CALBC, we have made improvements to both the system and the process of annotation. I2E uses an improved disambiguation process together with ontologies enriched with contextual information to support this. The major improvements to the process of annotation include the use of the silver standard training training data and the extended use of I2E queries. We have used the silver standard training data to create lexicons of family-terms of entities which are absent in the ontologies that we were using. Apart from this, we have used the training data to customize the disambiguation thresholds. Varying the disambiguation thresholds in I2E can vastly affect the scope and coverage of entity identification and thus it provides a good way to balance the precision and recall in a supervised/semi-supervised entity recognition task. Apart from using the training data, in this round, we have also used I2E querying patterns to discover entities which do not exist in the ontologies that we have used. This accounts for approximately 4% of the entities we recognise.

The results on gold standard data-sets like PennBioIE, FSUPRGE and BioCreative show an improvement of 7-14% in overall F-measure versus our previous results, with a particular improvement in recall. For PennBioIE, I2E was the top performing system while it registered significant improvements on FSUPRGE and BioCreative.

## Annotating the CALBC Challenge II Corpora with the JULIE Lab Tools

Ekaterina Buyko, Elena Beißwanger, Katrin Tomanek & Udo Hahn

Jena University Language & Information Engineering (JULIE) Lab  
Friedrich-Schiller-Universität Jena  
Jena, Germany  
<http://www.julielab.de>

The JULIE Lab team from FSU Jena processed the CALBC Challenge II data with the JULIE Lab pipeline based on JCoRe, the JULIE Lab Components' Repository (Hahn et al., 2008). Tokenization was performed by the JULIE Lab tokenizer (Tomanek et al., 2007), which is especially adapted to the intricacies of the biomedical sublanguage and was extensively evaluated on various biomedical corpora (Tomanek et al., 2007). For part-of-speech tagging, the OpenNLP POS tagger was applied (Buyko et al., 2006), which was re-trained on the GENIA treebank corpus (Ohta et al., 2002). For named entity recognition (NER) proper, FSU Jena considered various configurations of the JCoRe NER pipeline sub-part (see below). In general, FSU Jena applied two basic NER components, *viz.* GeNo (Wermter et al., 2009) and the Lingpipe chunker.<sup>2</sup> GeNo is a dedicated gene tagger and normalizer developed at the JULIE Lab that contains a variety of processing components for high-performance gene name recognition and normalization. The Lingpipe chunker implements a purely dictionary-based NER approach and was modified in such a way that it incorporates results from the acronym detection (Schwartz and Hearst, 2003) for the disambiguation of terms which were found in the texts.

FSU Jena submitted for the second CALBC Challenge five variants of the FSU system. Each of them differs in essential system parameters, *viz.* the dictionaries and the machine learning models being used:

*FSU system 1* applied GeNo in its original parameter settings as presented by Wermter et al. (2009) and as used without any changes in the first CALBC Challenge 2010. The gene dictionary contains entries from the BioCreAtivE II Entrez Gene dictionary (utilized in the BioCreative II gene normalization task)<sup>3</sup> and additional entries extracted from the UniProt database. Two additional dictionaries were incorporated in the Lingpipe chunker. First, ChemList (dictionary of small molecules and drugs) provided by EMC and filtered for chemicals with names no longer than 30 characters. Second, MeSH<sup>4</sup> headings (MH) and alternative labels (ENTRY, PRINT ENTRY) extracted from the MeSH (2008 version, in ASCII format).

*FSU system 2a* applied GeNo as in its *FSU system 1* version. Yet, the gene dictionary was extended by those annotations from the CALBC SSC II that were tagged with the semantic group 'PRGE' and that were not present in the previous version of the gene dictionary. Furthermore, the Lingpipe chunker used the unfiltered ChemList dictionary extended by those annotations from the CALBC SSC II that

---

<sup>2</sup> <http://alias-i.com/lingpipe/>

<sup>3</sup> [http://biocreative.sourceforge.net/biocreative\\_2\\_dataset.html](http://biocreative.sourceforge.net/biocreative_2_dataset.html)

<sup>4</sup> <http://www.nlm.nih.gov/mesh/>

were tagged with the semantic group 'CHED' and that were not present in the previous version of the chemicals dictionary. The MeSH dictionary was collected from MeSH headings (MH) and alternative labels (ENTRY, PRINT ENTRY) extracted from the MeSH (2011 version, in ASCII format) plus those annotations from the CALBC SSC II that were tagged with one of the semantic groups 'DISO' or 'SPE' and were not present in the previous version of the MeSH dictionary.

*FSU system 2b* differed from *FSU system 2a* only in the coverage of the applied MeSH dictionary. For *FSU system 2b* we extended the MeSH dictionary by names of substance (NM) and synonym (SY) entries extracted from the MeSH SCR (2011 version, in ASCII format).

*FSU system 3* used the same dictionaries as *FSU system 1*. The only difference between both systems is in the training data used for the GeNo's JNET (Jena Named Entity Tagger) model, which was originally trained on various publicly available gene-annotated corpora (see Wermter et al, 2009, for details). For *FSU system 3*, GeNo's JNET model was trained on the CALBC Challenge II training corpus.

*FSU system 4* used the same dictionaries as *FSU system 2b*, plus *FSU system 3* GeNo's JNET model trained on the CALBC Challenge II training corpus.

The rationale underlying the submission of different FSU systems with different parameter settings was to make explicit and also measure effects of various applications of the previous CALBC SSC corpus (SSC II) for the prediction on unseen data. The systems differ only in the coverage of dictionaries and in the training data used for the creation of the ML-based GeNo models.

## References

E. Buyko and **U. Hahn**, 2008. Fully embedded type systems for the semantic annotation layer. In *ICGL 2008 – Proceedings of the 1st International Conference on Global Interoperability for Language Resources*. Hong Kong, SAR, January 9-11, 2008, pp.26-33.

**E. Buyko**, J. Wermter, M. Poprat, and U. Hahn, 2006. Automatically adapting an NLP core engine to the biology domain. In *Proceedings of the Joint BioLINK-Bio-Ontologies Meeting. A Joint Meeting of the ISMB Special Interest Group on Bio-Ontologies and the BioLINK Special Interest Group on Text Data Mining in Association with ISMB*. Fortaleza, Brazil, August 2006, pp.65-68.

U. Hahn, E. Buyko, R. Landefeld, M. Mühlhausen, M. Poprat, K. Tomanek, and J. Wermter, 2008. An overview of JCoRe, the JULIE Lab UIMA Component Repository, In *Proceedings of the LREC'08 Workshop 'Towards Enhanced Interoperability for Large HLT Systems: UIMA for NLP'*, Marrakech, Morocco, May 2008, pp.1-7, 2008.

A. Schwartz and M. A. Hearst, 2003. A simple algorithm for identifying abbreviation definitions in biomedical text, In: *PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing 2003*, Kauai, Hawaii, USA, January 2003, pp.451–462.

T. Ohta, Y. Tateisi and J. Kim, 2002. The GENIA corpus: An annotated research abstract corpus in molecular biology domain, In: *HLT 2002 -- Human Language Technology Conference. Proceedings of the 2nd International Conference on Human Language Technology Research*, San Diego, CA, USA, March 2002, pp. 82–86.

K. Tomanek, J. Wermter and **U. Hahn**, 2007. A reappraisal of sentence and token splitting for life sciences documents, In: *MedInfo 2007 - Proceedings of the 12th World Congress on Health (Medical)*

*Informatics. Building Sustainable Health Systems*. Brisbane, Australia, August 2007. IOS Press, 2007, pp.524-528.

J. Wermter, K. Tomanek and U. Hahn, 2009. High-performance gene name normalization with GeNo. *Bioinformatics*, 25(6), 2009: 815-21.

## Abstracts – CALBC Challenge II: system descriptions – participants

### Concept identification by machine learning aided dictionary-based named entity recognition and rule-based entity normalisation

György Móra

Department of Informatics, University of Szeged, Hungary  
gymora@inf.u-szeged.hu

**Introduction:** We developed a concept identification UIMA pipeline to annotate genes, proteins and species in the CALBC corpus and participated in the Concept Identification task of the CALBC Challenge. The recognition of entities was performed via dictionary lookup using normalised string matching. The candidate expressions were filtered by a machine learning model trained on the BiocreativeII Gene Name Recognition task and lastly a rule-based PRGE normalisation was applied.

**Mention detection:** First, gene and protein concepts from the Entrez Gene<sup>5</sup>, Uniprot<sup>6</sup> and UMLS<sup>7</sup> databases and species from the NCBI Taxonomy<sup>8</sup> database were annotated. The synonyms collected from these sources were indexed by the Lucene<sup>9</sup> framework. Here we used the Specialist Lexical Tools<sup>10</sup> for normalising the tokens of the text. It assigns the possible normalised forms to each token. Then the punctuation characters were omitted and the Lucene search was conducted based on tokens containing alphanumerical characters.

Species recognition was performed by combining the labelling of the LINNAEUS (Gerner et al.) species tagger and the result of the dictionary matching. LINNAEUS is able to detect species with only their genera given, but in this case it automatically maps genera to their most frequent species names. We identified genera names and other references to specific species with our rule-based Taxonomical Entity (TE) mapper (Móra and Farkas 2010). The NCBI Taxonomy terms found by dictionary lookup were checked for compatibility with terms tagged by LINNAEUS and matching terms were linked together. This also eliminates the erroneous annotation of rare genera and other TEs which have no distinctive names. These types of TEs only occurred together with an exact species name from the taxonomical group in a document and are not used to refer to species alone.

**Mention filtering:** Matched protein and gene (PRGE) mentions were filtered by binary classification (we employed the Mallet implementation of the Maximum Entropy model). The training was carried out on the dataset of the BioCreative II gene mention detections task. The output of our dictionary-based matcher was used as training and testing instances, and their positive versus negative classes were set according to whether their boundaries matched one of the gold standard gene boundaries. The features were extracted from the words in a fixed window of four tokens before and after the candidate term as well as from the term itself. The features included the term type (gene or

---

5 <ftp://ftp.ncbi.nih.gov/gene/>

6 <ftp://ftp.uniprot.org/pub/databases/uniprot>

7 <http://www.nlm.nih.gov/research/umls/licensedcontent/umlsknowledgesources.html>

8 <ftp://ftp.ncbi.nih.gov/pub/taxonomy/>

9 <http://lucene.apache.org/java/docs/index.html>

10 <http://lexsrv3.nlm.nih.gov/Specialist/Summary/lexicalTools.html>

protein), word form, stem, term and token lengths, orthographic features, and character 1-3-grams of the token patterns. A set of stop words was applied as a feature and was used to filter out PRGE mentions.

We adopted a string-shape feature similar to the one described by Tsai et al. (2005) and distinguished between the character classes of upper and lower case Latin and Greek letters, Roman numerals, white spaces, punctuations and numbers.

Character sequences of the same character class were marked by the character associated with the class and character 2-3-grams of the pattern were also generated. With this feature it was possible to capture the structure of gene and protein names independently of the actual word form. This feature was also useful for filtering out common English words marked as PRGE names by the dictionary mapper.

**PRGE normalisation:** Inter-species protein and gene normalisation were done by using rules similar to those for TE normalisation. The nearest term to the gene or protein referring to species was determined and the compatibility of the two entities was checked. The Entrez and Uniprot databases contain species information in a formalised way, but the entries in UMLS may contain references to species only in their names and descriptions. Following the CALBC annotation guidelines in cases where we were unable to detect which species the gene or protein belongs to, we annotated all of the entities associated.

If all of the PRGE entities are from the UMLS database (therefore no species information is available), we annotate all of them. The nearest species mentions to the PRGE mention were determined and if the PRGE mention had entities associated with the same species id, then only those PRGE mentions were kept. If no entities matched the species mention, then the next nearest species mention was taken. If no species mention was found which was compatible with at least one PRGE entity, then all of the PRGE entities were kept.

The nearest species mentions were looked for before the term in the sentence and if none was found, we looked for them after the term in the same sentence. If there was no species term inside the sentence, the search scope was extended to the paragraph, section and the full document, respectively.

**Results:** Our system achieved 0.66 F-score (0.75 precision and 0.59 recall) in the PRGE (protein and gene) category and 0.68 F-score (0.64 precision and 0.72 recall) in the LIVB (species) category. These results compared to the other submissions were ranked 7<sup>th</sup> by F-score (5<sup>th</sup> precision and 6<sup>th</sup> recall) in the PRGE category and 8<sup>th</sup> by F-score (9<sup>th</sup> precision and 7<sup>th</sup> recall) in the LIVB category.

**Conclusion:** We developed an integrated gene, protein and species concept identification system which utilizes normalised dictionary lookup for named entity recognition, machine learning and rule-based systems for entity normalisation. In the future we plan to improve the performance of the normalisation by utilising information available in UMLS and other knowledge bases.

### **Acknowledgement**

This research was supported by the TÁMOP-4.2.1/B-09/1/KONV-2010-0005 program of the Hungarian National Development Agency.

## References

- Martin Gerner , Goran Nenadic and Casey M Bergman, 2010, LINNAEUS: A species name identification system for biomedical literature, *BMC Bioinformatics*
- McCallum and Andrew Kachites,"MALLET: A Machine Learning for Language Toolkit."<http://mallet.cs.umass.edu>. 2002.
- György Móra and Richárd Farkas, 2010 Species taxonomy for gene normalization, Fourth *International Symposium on Semantic Mining in Biomedicine*
- Tzong-Han Tsai, Chia-Wei Wu, and Wen-Lian Hsu. 2005. Using Maximum Entropy to Extract Biomedical Named Entities without Dictionaries. In *Second International Joint Conference on Natural Language Processing*, pages 268–273.

## Annotating large corpora with concept retrieval

Rafael Berlanga<sup>1</sup>, Victoria Nebot<sup>1</sup> and Ernesto Jimenez-Ruiz<sup>2</sup>

<sup>1</sup>Computer Languages and Systems, Universitat Jaume I (Spain)  
{berlangaromerom}@lsi.uji.es

<sup>2</sup>Computing Laboratory, Oxford University (UK)  
ernesto@comlab.ox.ac.uk

**Motivation.** For the second workshop of the CALBC challenge we have focused on the optimization of the Concept Retrieval approach we presented in the First CALBC workshop [1]. Basically, our approach consists of regarding concepts as documents and text chunks as queries, so that the problem of semantic annotation is viewed as an information retrieval (IR) task [2]. Thus, the annotation system must first find the most relevant concepts w.r.t. the text chunk words, and then select those concepts that best fit with the text chunk. In the First CALBC workshop this system took around one week to annotate the SSC-I corpus, even restricting the semantic types to those regarded in the competition. The optimizations we have introduced to scale-up our approach have been mainly focused on the indexing data structures used for concept retrieval, which aim at reducing the generation of concept candidates for the final semantic annotation. As a result we have been able to annotate the whole SSC-II corpus in less than a day, regarding all the UMLS semantic types.

**Concept Retrieval Method.** In our system, concept retrieval relies on the similarity between a given query (i.e. a text fragment) and each document of the collection (i.e. concept) in order to give a conceptual cover of the query. Currently, this similarity is estimated through an IDF-based measure, namely:

$$\text{sim}(C, T) = \max_{S \in \text{lex}(C)} \left( \frac{\text{info}(cw(S, T)) - (\text{info}(S) - \text{info}(cw(S, T)))}{\text{info}(S)} \right)$$

Where  $\text{info}(S)$  measures the relevance of the terms in the string  $S$ , and  $cw(S, T)$  is the set of terms in common between the concept string  $S$  and the text fragment  $T$ . It is defined as:

$$\text{info}(S) = - \sum_{w \in S} \log(P(w|UMLS))$$

A similar method has been also applied for detecting chemical expressions. We compare a text chunk  $T$  that potentially represents a chemical expression to each semantic group DG (e.g. CHED, PRGE, DISO, etc.) as follows:

$$\text{sim}(SG, T) = \sum_{w \in T} \log(P(w|SG))$$

**Optimizations.** As traditional IR systems, the implementation of the concept retrieval system relies on inverted files. Thus, each normalized word has a unique entry that contains the occurrences of the word in each concept string  $S$  (hitlist). In the hitlist, we also store the final score of each concept string (i.e.  $\text{info}(S)$ ) in order to speed up the calculation of the similarity function. Given a text chunk  $T$ , the system retrieves the involved hitlists of the words of  $T$ , generates the concept candidates

appearing in the hit lists, and finally calculates their similarity. As a result, a ranked list of concepts is obtained. With this ranking, the system obtains a concept cover of  $T$  so that it maximizes the global score of all the annotations associated to that cover. One of the critical points of this approach is the management of large hitlists, which implies huge lists of concept candidates, even for short text chunks. Although in our first approach we adopted a top-k strategy for processing hitlists, the number of generated candidates is still very high. This problem has been solved by adopting the following strategy: highly frequent terms do not generate candidates but just contribute to those generated by non-highly frequent ones. In this way, highly frequent terms are not associated to hitlists but to some look-up index (e.g. a trie) much cheaper to maintain. However, there are two cases in which these terms are required to generate candidates, namely: if a highly frequent term represents an interesting entity (e.g. mouse), and if a concept is represented only with highly frequent terms (e.g. muscle tumour).

For these terms, a mixed index structure is created: one hitlist for the previous exceptions and one trie for the rest. Fortunately, the number of highly frequent terms is very low, and therefore the mixed data structure does not imply any overhead over the whole index. In the system presented in the first CALBC workshop, we did not use any kind of chunker to generate the queries. Sentences were just split by stop words, and the resulting chunks were considered queries. In the current system we first attempted to detect noun phrases through some standard POS-tagging tool.

However, this produced a considerable overhead over the annotation process. Moreover, resulting chunks were not good enough due to the complexity of SSC-II abstracts. We must also point out that this issue is specially critical for avoiding the generation of large candidate lists due to large text chunks. For the SSC-II we have developed an ad hoc chunker that splits sentences by verb groups and connecting words.

**Preliminary Results.** We have used the UMLS Metathesaurus 2010AA as the concept inventory. The following table shows the number of annotations obtained for each corpora and semantic group. SSC I and SSC II contains the same abstracts. From this table we can conclude that the new system achieves a good coverage for most semantic groups, improving in all of them the results of the first workshop. This is also due to the change of the UMLS version used for the concept inventory (from 2009AB to 2010AA).

Semantic Group	SSC I	SSC II (175k)	SSC II (714k)
CHED	602,317	940,354	4,080,577
PRGE	531,729	2,813,972	12,230,662
DISO	332,413	789,236	3,566,787
OBJC	-	31,315	144,945
PHYS	-	460,309	1,976,062
OCCU	-	19,475	95,489
ORGA	-	10,705	53,455
MISC	-	325,667	1,497,960
ANAT	-	1,083,706	4,779,907
PHEN	-	147,460	642,270
CONC	-	8,973	42,998
LIVB (SPE)	310,591	999,255	4,378,881
PROC	-	545,896	2,474,235
GEOG	-	22,017	106,118
<b>Total</b>	<b>1,777,050</b>	<b>4,982,675</b>	<b>21,929,106</b>
<b>Distinct</b>	-	<b>91,676</b>	<b>144,319</b>

## References

[1] R. Berlanga, E. Jimenez, and V. Nebot. Semantic annotation of texts through concept retrieval. In Proc First CALBC Workshop, pages 1{3, 2010.

[2] R. Berlanga, V. Nebot, and E. Jimenez. Semantic annotation of biomedical texts through concept retrieval. *Procesamiento del Lenguaje Natural*, 45(A):247{250, 2010.

## Annotating the CALBC corpus with a machine learning harmonization approach

David Campos, Sérgio Matos, and José Luís Oliveira

University of Aveiro, DETI/IEETA  
Campus Universitário de Santiago, 3810-193 Aveiro, Portugal  
{david.campos,aleixomatos,jlo}@ua.pt

Named Entity Recognition (NER) is a field of Natural Language Processing (NLP) that intends to identify chunks of text referring to entities of interest. The main goal of the CALBC project [1] is to provide a large-scale biomedical corpus that contains annotations of several biological concepts (diseases, species, chemicals and genes/proteins), by combining annotations from several NER systems. In order to participate in this challenge, we built a system that uses Conditional Random Fields (CRFs) [2, 3] as the core component, to recognize gene/protein names and harmonize heterogeneous annotations. Nowadays, there is an increasing research interest in combining annotations from several NER systems, in order to take advantage of the annotations' variability from systems, human curators and corpora. To create the NER systems with different characteristics, we used three well known Gold Standard Corpora (GSC): GENETAG [4], JNLPBA [5], and PennBioIE [6]. Considering the several types of features described in the literature, we trained one model per corpus, selecting the features that optimize the results for each one. At the end, three different models were obtained with the following characteristics:

- GENETAG: a 2nd Order CRF model with lemma, POS, chunk, orthographic and morphological features. It also uses a f-2,2g window of features to model local context, and dictionary matching with verbs and gene/protein names as features;
- JNLPBA: a 1st Order CRF model with lemma, POS, chunk, orthographic and morphological features. It uses conjunctions of features to model local context, and dictionary matching with gene/protein names and domain concepts as features;
- PennBioIE: a 2nd Order CRF model with chunk and morphological features. It also uses conjunctions of features to model local context.

To combine the heterogeneous annotations from the three systems, we developed a machine learning system based on CRFs. The systems' annotations were used as token features and the gold standard annotations as labels, both encoded using the BIO tagging format. With this strategy, the CRF model uses a rich knowledge base, learning the correct tags for the tokens and making the final result more precise and reasoned. In comparison with traditional harmonization solutions, which only allow fixing the annotations boundaries (by adding or removing tokens), our solution also allows creating new annotations or removing incorrect ones with high precision, extending the traditional behaviour and improving the global recall.

Based on the methods previously described, we propose three different strategies to annotate the CALBC corpus (Figure 1): 1) single model trained on GSC; 2) combination of models using GSC; and 3) combination of models using Silver Standard Corpora (SSC).

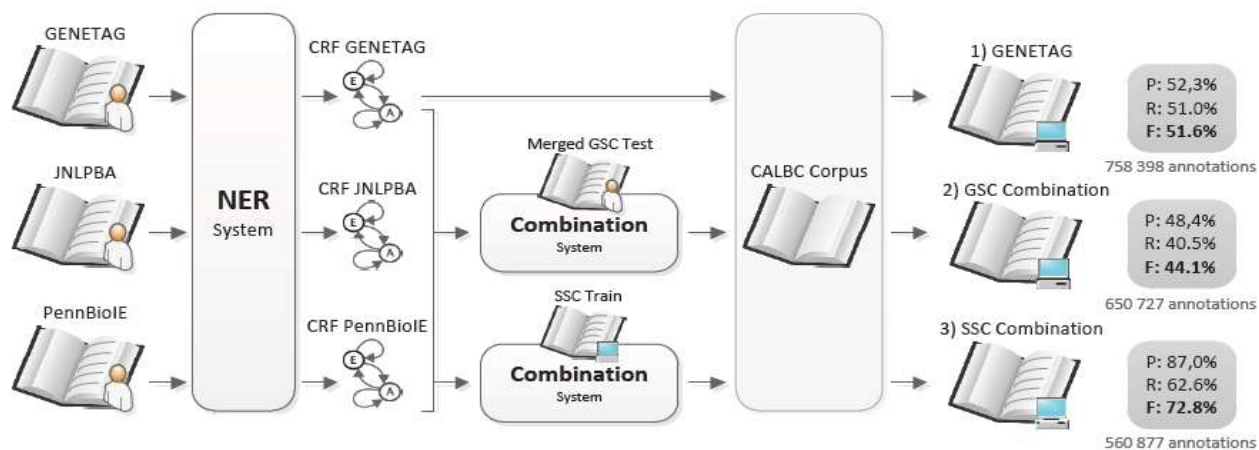


Fig. 1. Illustration of the different strategies used to annotate the CALBC corpus.

The first solution does not use any combination strategy, using only one CRF model trained on the GSC. To choose the best model, each model was trained on each corpus. Afterwards, their performance was evaluated against the CALBC training data (75 thousand abstracts automatically annotated), with the model trained on the GENETAG corpus presenting the best results. The second approach takes advantage of the combination solution, using GSC to train the combination module. To make this possible, the three corpora were divided in training and testing parts. The three models were trained on the respective training parts, and the combination system was trained on a corpus that merges the three test parts. Finally, the third solution uses part of the SSC training data (the first 28 thousand abstracts) to train the combination model. In this approach, the GSC models were trained using the complete corpus.

#### Annotating the CALBC corpus 3

The final results of the challenge (Figure 1) showed that the third solution achieved the better performance results. Overall, it was the solution with the 3rd best F-measure in the recognition of gene/protein names. Surprisingly, the first solution presented a better performance than the GSC combination, which may be a consequence of the small amount of data used to train the combination model. Moreover, the best solution also presents the smaller amount of annotations, removing a large number of false positives.

As a future work, the combination model could be trained using the complete training SSC (75k abstracts), in order to improve the solutions recall. With these results, we can also make a deeper analysis regarding the usage of GSC against SSC.

#### References

1. Rebholz-Schuhmann, D., Yepes, A., Van Mulligen, E., Kang, N., Kors, J., Milward, D., Corbett, P., Buyko, E., Beisswanger, E., Hahn, U.: CALBC silver standard corpus. *Journal of bioinformatics and computational biology* 8 (2010) 163{179
2. Lafierty, J., McCallum, A., Pereira, F.: Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In: *Proceedings of the International Conference on Machine Learning, Citeseer* (2001) 282{289
3. McCallum, A.K.: MALLETT: A machine learning for language toolkit. 2002. (<http://mallet.cs.umass.edu>)
4. Tanabe, L., Xie, N., Thom, L., Matten, W., Wilbur, W.: GENETAG: a tagged corpus for gene/protein named entity recognition. *BMC bioinformatics* 6 (2005) S3

5. Kim, J., Ohta, T., Tsuruoka, Y., Tateisi, Y., Collier, N.: Introduction to the bio-entity recognition task at JNLPBA. In: Proceedings of the International Joint Workshop on Natural Language Processing in Biomedicine and its Applications, Association for Computational Linguistics (2004) 70-75
6. Kulick, S., Bies, A., Liberman, M., Mandel, M., McDonald, R., Palmer, M., Schein, A., Ungar, L., Winters, S., White, P.: Integrated annotation for biomedical information extraction. In: Proceedings of the Human Language Technology Conference and the Annual Meeting of the North American Chapter. (2004)

## Dictionary-based concept identification with UMLS

Max De Wilde <sup>a,b</sup>, Roser Morante <sup>a</sup>, Walter Daelemans <sup>a</sup>

<sup>a</sup> CLiPS Computational Linguistics Group, University of Antwerp, Belgium

<sup>b</sup> ReSIC Information and Communication Research Center, Free University of Brussels (ULB), Belgium

The concept identification and annotation task was performed by means of a dictionary-based system combined with low-level linguistic processing, with no previous training.

Our system combines Python scripts with a dictionary-based approach, taking full advantage of the UMLS metathesaurus which contains over 5 million concepts from the biomedical domain. Some basic linguistic processing (mainly stemming and stopwords filtering) was used in order to improve coverage, but had to be kept at a minimum in order to optimize the efficiency of the algorithm.

As no training on the previous Silver Standard Corpora (SSC) was used, the system is largely dependent on what is to be found in the UMLS database, which raised a number of quality issues. Therefore, a preliminary data cleansing step was undertaken with the CASPER tool (CleAn, SuPprEss and Rewrite UMLS data),<sup>11</sup> using default options.

After loading the UMLS information into Python dictionaries, we parsed the PubMed XML files with xml2dict<sup>12</sup> and tokenized the abstracts and titles with the MBSP tokenizer specialized for biomedical texts.<sup>13</sup> Our algorithm then checked every term (up to five words long, as recommended by Casper) against the dictionary and associated each match with its UMLS unique identifier.

Short common words such as *an* and *I* were systematically ruled out in order to avoid being mistaken for medical abbreviations. Unrecognized words were further stemmed with a Python implementation of the Porter Stemmer in order to identify more terms.<sup>14</sup> It is important to understand that lemmatization was not a option here as it would have increased the processing time dramatically. Each UMLS ID was subsequently mapped to a semantic type, and each group in turn to a broader semantic group, as recommended in the challenge guidelines.

Thanks to the Python dictionary data structure, we were able to process the smaller corpus (175,000 abstracts) in under 30 minutes, and the bigger one (714,000 abstracts) in little more than two hours, using only 2GB of computer memory. As a reference point, a processing time of one second per abstract with the same computer would have taken over two days for the small corpus.

First results show high recall scores for semantic types having good UMLS coverage (e.g. 93.58% for geographic areas) and an average precision of almost 70%, with peaks over 85% for certain types. The system was best at identifying phenomena and physiological concepts, with F-scores of 73.48% and 74.91% respectively, and miscellaneous objects were also well detected. These results,

---

<sup>11</sup> <http://biosemantics.org/casper>

<sup>12</sup> <http://code.google.com/p/xml2dict>

<sup>13</sup> <http://www.clips.ua.ac.be/pages/MBSP#tokenizer>

<sup>14</sup> <http://snowball.tartarus.org/algorithms/english/stemmer>

combined with the fact that our system scored better for the less developed types that were not used in the first challenge, confirm our expectations that it is a good baseline approach that could serve as a basis for more specialised systems.

## References

T. De Smedt, V. Van Asch and W. Daelemans, Sep. 2010. "Memory-based Shallow Parser for Python". Tech. Rep. 2, CLiPS Technical Report Series (CTRS).

K. Hettne, E. van Mulligen, M. Schuemie et al., 2010. "Rewriting and suppressing UMLS terms for improved biomedical term identification". *Journal of Biomedical Semantics*, 1(1).

V. Van Asch, R. Morante and W. Daelemans, 2010. "Towards improving the precision of a relation extraction system by processing negation and speculation". In *Proceedings of the 4th Symposium on Semantic Mining in Biomedicine (SMBM)*, 164–165.

# OntoGene at CALBC II and Some Thoughts on the Need of Document-Wide Harmonization

Simon Clematide, Fabio Rinaldi, Gerold Schneider

Institute of Computational Linguistics, University of Zurich  
{siclemat,rinaldi,gschneid}@cl.uzh.ch

## Introduction

The OntoGene group has developed several syntax-based approaches for relation mining in the molecular biology domain, especially for the detection of mentions of protein-protein interactions. The effectiveness of these approaches has been validated by participation to shared evaluations, such as BioCreative II.5 [1] and III [2], or BioNLP event extraction task [3]. For the first CALBC challenge [4], the dictionary-based term recognizer originally developed for BioCreative II.5 has been adapted to the needs of large scale annotation. This system makes use of an efficient dictionary-based longest-match lookup procedure for the annotation of token sequences, which includes a flexible normalization to deal with surface variants of the terms stored in the dictionaries. The text tokens undergo the same normalizations as the original dictionary terms, thus allowing direct comparison of the normalized version of a textual candidate term with the normalized version of a reference term.

For the second CALBC challenge, we retained this dictionary-based engine for candidate term generation. However, as our results for protein and gene recognition of the first CALBC challenge showed, the bias towards high recall and lower precision, which works well for protein-protein interaction detection, needs remedy. Therefore, we filtered the candidate terms of our dictionary-based engine by a statistical hidden Markov Model (HMM). In particular, we trained a “First-Best Named Entity Chunking Model” using the LingPipe framework [5] for each of the 4 basic semantic types in the training corpora. For this supervised learning step we worked with the training corpora of CALBC I and II, as well as with the GENETAG corpus [6]. In order to benefit from the combination of a dictionary-based longest-match recognizer and a statistical chunker, we filtered the candidate terms for the final submission by the following rule: discard all candidate terms where the HMM chunker does not predict the begin of a named entity. We did not require exact correspondence of the end of terms deliberately, because we observed a slight bias towards shorter terms in the case of the HMM chunker.

## Technical Details and Evaluations for Our Participation in Task A of CALBC II

First, we split the huge corpora into smaller slices in order to speed up the processing by massive parallelization. Then, we tokenized using Lingpipe's biomedical HMM tokenizer. The HMM chunker was applied separately for each semantic type we had trained it for. From this step, only a “begin-of-term” marker was retained for each term chunk, which ensures that the dictionary-based term recognizer excludes matches containing them. The term recognizer used the following external dictionaries in addition to the terms we extracted from the CALBC training corpora (I and II):

- UniProt for proteins and genes (prge): 826,901 terms (incl. CALBC terms)
- PharmGKB for diseases (diso): 36,080 terms (incl. CALBC terms)
- PharmGKB for chemical substances and drugs (ched): 43,997 terms<sup>1</sup> (incl. CALBC terms)
- NCBI Taxonomy (and own resources) for species (spe): 903,880 terms (incl. CALBC terms)

Table 1 contains an overview of the filtering effect of the HMM chunker on the output of the dictionary-based term recognizer. The mismatch between both methods is rather high for all categories, except for the recognition of species. However, in a 10-fold cross-validation experiment with the HMM chunker on the CALBC II training corpus using an exact boundary recognition criterion, recall, precision, and F1-measure were not as high as one would expect (see Table 2). One reason may be that a simple HMM model generally performs worse than what can be expected from techniques as Conditional Random Fields [7].

1 The ChEBI 3 star level database was not used for the submissions because of an unfortunate configuration error of our system.

Another reason may be that the term annotation in the harmonized corpus is not as consistent as it should be. We tried to roughly quantify this effect by using the following one-sense-per-abstract hypothesis [8]: for each abstract in the CALBC II training corpus, all token sequences of annotated terms were collected. Then, in the same (!) abstract we searched for occurrences of term token sequences that were not annotated as terms of the same type. Under the one-sense-per-abstract hypothesis each unannotated occurrence counts as a false negative. Table 3 gives an overview of how many missing terms we can expect for each semantic type. This corresponds roughly to the results in Table 2.

Type	DTR	HMM Chunker	Validated	in %
prge	869,550	535,959	493,572	57%
ched	741,059	472,097	405,104	55%
diso	644,717	416,654	347,711	54%
spe	494,183	393,040	365,470	74%

Table 1: Amount of validated terms of dictionary-based term recognizer (DTR) by HMM NER chunker on 175k article set

Type	Recall	Precision	F1-Measure
spe	88%	82%	85%
diso	79%	73%	76%
prge	75%	57%	65%
ched	72%	50%	59%

Table 2: 10-fold cross-validation experiment on CALBC II training corpus using only the HMM NER chunk

Even if the one-sense-one-abstract hypothesis does not hold strictly (e.g. “in” should not be considered a protein in the whole abstract if it appears once as one), there is still room for improvement of document-wide harmonization of annotations (e.g. in one abstract “immunoglobulin” appears 6 times as a term, yet it is missed another 4 times).

Table 4 contains our official results evaluated against the consensus annotation from the CALBC project partners. These results show that simple HMM filtering in combination with dictionary-based term recognition improves in particular in the case of the more difficult problems, namely the recognition of protein/genes and chemical substances.

Type	Total Terms	Missing	in %	Miss. Types
prge	327,090	84,727	26%	6621
ched	165,098	36,604	22%	2355
diso	230,459	17,143	7%	987
spe	247,529	13,940	6%	718

Table 3: Total terms in CALBC II training corpus and missing terms according to one-sense-one-abstract hypothesis

Type	R (rank)	P (rank)	F (rank)	S
spe	88% (2)	86% (2)	87% (2)	14
diso	74% (5)	80% (5)	77% (4)	15
prge	83% (1)	67% (11)	74% (2)	18
ched	77% (2)	69% (7)	72% (1)	14

Table 4: Official results for recall (R), precision (P) and F1-measure (F) with the corresponding rank in parens. Column S contains the total number of submissions for each semantic type.

**Funding:** The OntoGene group is supported by the Swiss National Science Foundation (grant 105315-130558/1) and by NITAS/TMS, Text Mining Services, Novartis Pharma AG, Basel, Switzerland.

## References

- [1] F. Rinaldi, G. Schneider, K. Kaljurand, S. Clemenide, T. Vachon, and M. Romacker, “OntoGene in BioCreative II.5,” *IEEE/ACM Transactions on Computational Biology and Bioinformatics*, vol. 7, 2010, pp. 472-480.
- [2] F. Rinaldi, G. Schneider, S. Clemenide, M. Romacker, and T. Vachon, “OntoGene (Team 65): preliminary analysis of participation in BioCreative III,” *BioCreative III workshop*, 2010.
- [3] K. Kaljurand, G. Schneider, and F. Rinaldi, “UZurich in the BioNLP 2009 Shared Task,” *Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task*, Boulder, Colorado: Association for Computational Linguistics, 2009, pp. 28-36.
- [4] S. Clemenide, F. Rinaldi, and G. Schneider, “OntoGene in CALBC,” *First CALBC Workshop*, 2010, pp. 30-31.
- [5] B. Carpenter, “LingPipe for 99.99 % Recall of Gene Mentions,” *Proceedings of the Second BioCreative Challenge*, 2007, pp. 2-4.

[6] L. Tanabe, N. Xie, L.H. Thom, W. Matten, and W.J. Wilbur, "GENETAG: a tagged corpus for gene/protein named entity recognition," *BMC Bioinformatics*, vol. 6, 2005, p. S3.

[7] K. Hara, "Towards automatic biomedical entity annotation by reducing error propagation," *First CALBC Workshop*, 2010, pp. 35-37.

[8] W.A. Gale, K.W. Church, and D. Yarowsky, "One sense per discourse," *Proceedings of the workshop on Speech and Natural Language*, Association for Computational Linguistics, 1992, pp. 233-237.

# Biomedical Named Entity Recognition in CALBC Challenge II using Dictionaries and SVMs

Aliyu Kabir Musa<sup>1</sup>, Ekrem Varoğlu<sup>1</sup>, Nazife Dimililer<sup>2</sup>

<sup>1</sup>Computer Engineering Department, Eastern Mediterranean University, North Cyprus.

<sup>2</sup>School of Computing and Technology, Eastern Mediterranean University, North Cyprus.  
{aliyukabir.musa, ekrem.varoglu, nazife.dimililer}@emu.edu.tr

## 1-Introduction

Named Entity Recognition (NER) or Entity Boundary Identification is one of the first and most crucial steps in text mining. Many efforts have been put into this task using gold standard data such as those made available by the BioCreative NER challenges and the JNLPBA competition. A large number of systems using similar features, but different learning approaches have been used with satisfactory performance for this purpose.

In this work we discuss the systems used in the CALBC Challenge II, Named Entity Recognition Task (Task A). Two different systems are proposed to annotate the Silver Standard Corpus III (SSC-III). The first approach is a dictionary based system using an in-domain dictionary formed solely from the training data (SSC-II). The second approach is a multi-classifier system where ten Support Vector Machines (SVM) were trained on non-overlapping sections of the training data and combined using simple majority voting. In both approaches, external resources were not used and the data was tagged with the 4 semantic groups, namely, *ched*, *diso*, *livb*, and *prge* without any reference to specific UMLS semantic types.

## 2-Systems Used

### 2.1 Dictionary Based System

The success of a dictionary based system is determined by the coverage of the dictionaries used and the efficiency of lookup mechanism together with the criteria employed for finding a match. The approach used in this work was to use the word stems computed by the Genia Tagger [1] in order to minimize the effect of inflected variants and different derivations of a given word.

In order to form the dictionaries, a separate semantic group dictionary that included the stems of the most frequently occurring tokens in that semantic group was created for each semantic group. If more than 30% of all occurrences of a word stem were tagged in a certain semantic group in the training data, that word stem was added to the dictionary of the aforementioned semantic group. The 30% cut off point was used in order to allow a token to be assigned to more than one semantic group as is the case in the training data and also to minimize noise by allowing a maximum of three semantic groups per token. A final dictionary was formed for the OUT category using the same procedure with a 65% cut off point

During the testing phase, the semantic group dictionaries were utilized in order to tag each token in the test data with up to three semantic groups. If the stem of a word was listed in a semantic group dictionary, the word received the corresponding semantic group tag(s), otherwise it was tagged as OUT class. However, no word was assigned as belonging to the OUT class together with the other semantic groups. The dictionary for the OUT category was not used in this approach. The dictionary based system was applied to both the small (175K) and big (714K) test sets.

### 2.2 SVM Based System

In the machine learning approach proposed in this work, the train data and the test data were tokenized and each token was tagged with features representing clues for categorizing the token into the four semantic groups. The original training data contained multiple annotations for a token but our SVM based system did not support multiple annotations. Therefore some pre-processing was done in order to transform the training data into a suitable format. As the first step, multiple entity annotation of a token was replaced by a single entity tag corresponding to the most frequently used semantic group, determined from the training data, for the given token. In the second step, the tag representation was transformed into the IOB2 format.

Previous work in this domain has shown that orthographic features such as the use of upper case letters and Greek characters, morphological feature such as suffix and prefix information, as well as POS and phrase tags are useful for the recognition of biomedical entities [2,3]. In this system, orthographic and morphological features as discussed in [3] and a novel vocabulary feature were used in conjunction with POS and phrase tags produced by the GENIA tagger. The orthographic feature was based on a vector representation of the most commonly occurring orthographic properties of biomedical entities in the JNLPBA corpus [3,4] The use of this feature was shown to improve the recognition performance of an entity tagger in comparison to the use of single orthographic features in isolation [3]. Additionally morphological features comprised from 1,2, and 3 n-grams of the tokens both as prefixes and suffixes were used. Finally a 5-bit vector representation of the vocabulary feature was added such that each bit of the vector represented a semantic group and the last bit corresponded to the OUT class. The dictionaries constructed for the dictionary based system proposed above were used to encode the vector.

Ten different SVMs were trained using the same set of features and properties on 10 non-overlapping segments comprising the complete SSC-II training data set. This approach was also expected to produce classifiers with different recognition performance for the cases where the training data is not uniform.

During testing, each SVM was tested with the complete test data and the individual SVM outputs were combined using simple majority voting. In this approach, each token was assigned a single annotation that belonged to one of the semantic groups in the IOB2 format which was then converted to the format required by the challenge. The SVM based multi-classifier system was applied only to the small (175K) test set mainly due to time constraints.

### 3-Results and Discussions

We have annotated the small (175K) and the big (714K) test sets using the dictionary based system and the small set only using the SVM based system, both described above. Since we have used the train data supplied as the only source in designing and training of our systems, we have annotated the test data sets with the four semantic groups provided in the train data; namely *ched*, *diso*, *livb*, and *prge*. Our results for the small data set against a "consensus" annotation derived from the partner submissions are as given in Table 1 below.

Table 1. Performance on the 175K test set.

System	Semantic Group	Precision(%)	Recall(%)	F-score(%)
Dictionary based	<i>ched</i>	75.7	47.8	58.6
	<i>diso</i>	56.7	57.9	57.3
	<i>livb</i>	73.5	61.6	67.0
	<i>prge</i>	73.8	61.1	66.9
SVM based	<i>ched</i>	81.0	56.6	66.6
	<i>diso</i>	60.2	41.0	48.8
	<i>livb</i>	78.2	64.4	70.6

	<i>prge</i>	73.4	32.3	44.9
--	-------------	------	------	------

In general our systems seem to be lacking good recall properties. For the dictionary based system, this an expected behaviour from the dictionary based system considering the fact that the dictionary was constructed using only stem words from the train data. Performance can be improved using external sources for dictionary formation. For the SVM based system, one can always increase the recall by training systems with different features. The overall performance can then be improved using classifiers with varying precision-recall properties in the ensemble.

#### 4-Future Work

The use of external dictionaries will be investigated in order to increase the recall property of the dictionary based system. Since we use the dictionary constructed as a vocabulary feature in our SVM system, it is believed that the use of such a dictionary will improve the performance of the SVM based system.

Each SVM used in this study was trained using the same set of features and parameters. As future work, SVMs with different feature sets and/or parameters will be trained on the complete training data in order to introduce diversity among the classifier members in the classifier ensemble. Another improvement in the initial setup of the SVM based approach proposed in this work is in the majority voting algorithm which was used for selecting a single entity tag as the output of classification. Alternatively, it will be redesigned to choose two or three labels that receive the highest votes.

A final improvement that is planned for the machine learning approach will be use of a genetic algorithm to choose the SVMs that participate in the final decision of the ensemble as described in [4]. Here, the individual strengths of the classifiers will be incorporated into the ensemble by assigning a separate vote, representing the reliability of the estimation, for each entity tag estimated by each classifier member of the ensemble. For the case of  $n$  classifiers in the ensemble and with  $m$  entity tags, the proposed approach results in a total of  $nxm$  votes for the majority voting algorithm. Finally, we plan to apply the SVM based system for the annotation of the big test set.

#### References

- [1] <http://www-tsujii.is.s.u-tokyo.ac.jp/GENIA/tagger>
- [2] G. D. Zhou, and J. Su, "Exploring deep knowledge resources in biomedical name recognition," in Proc. of the Joint Workshop on Natural Language Processing in Biomedicine and its Applications (JNLPBA-2004), pp. 96-99, 2004.
- [3] N.Dimililer, E.Varoğlu, "Recognizing Bio-medical named entities using Support Vector Machines: Improving recognition performance with a minimal set of features", Lecture Notes in Computer Science(LNCS), vol. 3886, Springer-Verlag, pp. 53-67, April 2006.
- [4] N. Dimililer, E. Varoğlu, H. Altınçay, "Classifier Subset Selection for Biomedical Named Entity Recognition", Applied Intelligence Journal, vol. 31, pp. 267-282, December 2009.

## Brief Description of ITNLP System for CALBC II

Yaming Sun, Chengjie Sun, Lei Lin

School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China  
{ymsun, cjsun, [linl](mailto:linl@insun.hit.edu.cn)}@insun.hit.edu.cn

We participated in the task A of CALBC challenge II: Named Entity Recognition (NER). A maximum entropy classifier with eight effective features was designed to complete this task. The system could deal with regular named entities, but no special process for nested NER. During the system development, we split the afforded training file (CALBC.SSCII.Train.75k\_01-12-10.xml) into two parts which have the same sentence amounts. One part served as training file, and the other as test file. Experiment results show that ITNLP system could obtain 70.71% F-measure with the dataset described before (evaluation tool is conllevl.pl<sup>1</sup>).

NER is an extremely important and fundamental task of biomedical text mining. But due to the complex nature of biomedical NERs, biomedical named entity recognition is still a challenging task. The CALBC project aims to annotate named entities in a large biomedical corpus with a variety of semantic categories. The CALBC challenge II consists of two tasks, they are challenge task A (NER) and challenge task B (Concept identification).

In consideration of the complexity of biomedical named entity, such as nestification, ambiguity and so on, and CALBC organizer provided us abundant labeled data, so we used machine-learning method to complete this task. Machine learning methods like Conditional Random Field (CRF), Maximum Entropy (MaxEnt), Hidden Markov Model (HMM) and Support Vector Machine(SVM) have been widely used for learning to recognize named entities in biomedical literatures. At first, we tried to use CRF, but owing to the restriction of our hardware condition, the CRF++ toolkit will say out of memory if the training file was larger than 100MB. The training file we used was larger than 400MB, so CRF was given up. What's worse, the training time of CRF was too long. Compared with CRF, MaxEnt could overcome these two shortcomings. Firstly, MaxEnt could make use of all the training data; secondly, the training time of MaxEnt was far below that of CRF, so we could have enough time to conduct experiments to find effective features. For example, given a training file that contains 18815 sentences, we used CRF++<sup>2</sup> and maxent toolkit<sup>3</sup> to train model with the training file separately. The training time of CRF was 1518.71 seconds, but the training time of maxent was only about 5.33 seconds (4G RAM, 2.66G\*4 core). Therefore, MaxEnt was adopted in our system.

In order to describe the complex language phenomenon in biomedical literatures, we used eight features, they are: the current word, word base, word morphology, word-formation, Part of Speech (POS) information, chunk and the previous two NE tags. We didn't use any external resources such as dictionaries. We used GeniaTagger<sup>4</sup> to acquire word base, POS and chunk information. The words of the same form are likely to have the same properties, so we introduced word-formation. Word-formation refers to how a word is formed. In view of the particularity of the morphology of biomedical named entities, we introduced morphology features. In our system, there are six kinds of morphology features, which are single-char, single-digit, punctuation, all-letter, all-digit and mix. In

the process of training and annotation, we used B/I/O format. “B-ne” refers to that the word is at the beginning of a named entity of type “ne”, and “I-ne” indicates the rest words that constitute the named entity of type “ne”, while “O” suggests that the word does not belong to any type.

Following is the work flow of our system. Firstly we selected all the sentences from the original training file in XML format and extracted the entity type. When we extracted the entity type, we mainly dealt with the simplest situation which means that a named entity only belongs to one type. If a named entity belongs to several types, for example, in one sentence of the training file, a named entity Cytochromes belongs to both type :::ched and type :::prge, then we treated :::ched|:::prge as a new entity type. Then we extracted the features mentioned above to get the well-formatted training data for maximum entropy tool kit to train model. Thirdly, we dealt with the test file in a similar way, and used the model we had trained to annotate the test file. Fourthly, some post-processes were conducted to the result of the annotation. The last step was to restore the result file. The training file we used is CALBC.SSCII.Train.75k\_01-12-10.xml, and the test file we used is CALBC.SSCII.Test.175k\_24-11-10.xml.

According to the preliminary challenge results released by CALBC team, our system can recognize four semantic groups, they are CHED, DISO, LIVB and PRGE. The results are shown in Table 1, and the number of valid\_annotation is 15134.

Table 1: Official evaluation results for the CALBC SSCII test data

Semantic group	precision	recall	f-measure	p-rank	r-rank	f-rank
ched	0.839	0.560	0.671	2	6	2
livb	0.789	0.815	0.802	4	4	3
prge	0.869	0.570	0.688	3	7	5
diso	0.558	0.549	0.553	12	10	11

<sup>1</sup><http://www.cnts.ua.ac.be/conll2000/chunking/output.html>

<sup>2</sup><http://crfpp.sourceforge.net/#download>

<sup>3</sup> <http://homepages.inf.ed.ac.uk/lzhang10/pmwiki/pmwiki.php>

<sup>4</sup> <http://www-tsujii.is.s.u-tokyo.ac.jp/>

## MetaMap in the CALBC Workshop II

James Mork, Lee Peters, Antonio Jimeno-Yepes, Alan R. Aronson, Olivier Bodenreider

National Library of Medicine  
8600 Rockville Pike, Bethesda, 20894, MD, USA

### 1 Introduction

MetaMap [4] is a tool which maps biomedical text to UMLS<sup>®</sup> Metathesauri<sup>®</sup> concepts. In MetaMap, input text undergoes a lexical/syntactic analysis consisting of a first analysis in which tokens, sentence boundaries and acronyms or abbreviations are identified and each token is assigned a part of speech. Input words are mapped to the SPECIALIST lexicon [8] using lexical lookup and then the SPECIALIST minimal commitment parser [8] identifies phrases and their lexical heads. The identified phrases are processed to generate variants (normally by table lookup), then candidates (Metathesaurus strings) are identified computing and evaluating their match to the input text. Then mapping constructions are produced in which candidates found in the previous step are combined and evaluated to produce a final result that best matches the phrase text. Finally word sense disambiguation (WSD) might optionally be used, in which mappings involving concepts that are semantically consistent with surrounding text are favored. MetaMap is available from [2]. Downloads are restricted and require a valid UMLS user account. A 2011 MEDLINE baseline annotation with MetaMap is available from [3].

### 2 Methods

From the two available CALBC sets, we have performed the annotation of the 175K citations set. MetaMap has been configured to run using the 2010AA version of the UMLS. Preprocessing of the UMLS Metathesaurus is described in [6]. Several customizations of MetaMap were required to adjust to the requirements from the challenge guideline [1]. MetaMap has been adapted to run on the sentence boundary annotation provided in the citation set. The output of MetaMap has been turned into the leXML format. The 2010 version of the UMLS Semantic Network has been mapped back to an earlier version specified in the CALBC guidelines, in which recent changes to the Organismhierarchy of the Semantic Network are not reflected. In addition, CALBC also uses its own mappings between Semantic Types and Semantic Groups. A mapping table has been generated to produce the set of Semantic Group annotations required in the guidelines.

MetaMap is a highly configurable tool. We have produced three runs of increasing complexity with the following options:

1. MetaMap with default options (metamap10 -Z 10 -% format -E). The default options use the strict model of the Metathesaurus [6]. This model should produce the highest precision of the annotation when compared to the moderate or relaxed model. It contains 2,424,017 (44.99%) of the 5,394,495 English Metathesaurus strings.
2. MetaMap with default options combined with Word Sense Disambiguation (WSD) (metamap10 -Z 10 -y -% format -E). Ambiguity is a main concern since in the UMLS Metathesaurus [7]. The WSD algorithm used in MetaMap is based on the JDI method [5].
3. MetaMap with default options combined with WSD and quick composite phrases (metamap10 -Z 10 -yQ -%format -E). A composite phrase is a simple phrase followed by any prepositional phrase optionally followed by one or more of prepositional phrases. An example is “pain on the left side of the chest” which will map to “Left sided chest pain” rather than separate concepts as it would without the option. The quick composite phrases option is still experimental.

## References

- [1] CALBC challenge II guidelines. [http://www.ebi.ac.uk/Rebholzsrv/CALBC/challenge\\_guideline.pdf](http://www.ebi.ac.uk/Rebholzsrv/CALBC/challenge_guideline.pdf).
- [2] MetaMap website. <http://metamap.nlm.nih.gov>.
- [3] NLM LHCBC Semantic Knowledge Representation website. <http://skr.nlm.nih.gov>.
- [4] A.R. Aronson and F.M. Lang. An overview of MetaMap: historical perspective and recent advances. *Journal of the American Medical Informatics Association*, 17(3):229, 2010.
- [5] S.M. Humphrey, W.J. Rogers, H. Kilicoglu, D. Demner-Fushman, and T.C. Rindfleisch. Word sense disambiguation by selecting the best semantic type based on Journal Descriptor Indexing: Preliminary experiment. *Journal of the American Society for Information Science and Technology*, 57(1):96, 2006.
- [6] F.M. Lang and A.R. Aronson. Filtering the UMLS Metathesaurus for MetaMap. <http://skr.nlm.nih.gov/papers/references/filtering10.pdf>.
- [7] F.M. Lang, S.E. Shooshan, J.G.Mork, and A. R. Aronson. Ambiguity in the UMLS Metathesaurus. <http://skr.nlm.nih.gov/papers/references/ambiguity10.pdf>.
- [8] A.T. McCray, A.R. Aronson, A.C. Browne, T.C. Rindfleisch, A. Razi, and S. Srinivasan. UMLS knowledge for biomedical language processing. *Bulletin of the Medical Library Association*, 81(2):184, 1993.

# Scalable Interlinking of Bio-Medical Entities and Scientific Literature in Linked Life Data

Georgi Georgiev\*, Konstantin Pentchev, Andrey Avramov, Todor Primov and Vassil Momtchev

{georgi.georgiev;konstantin.pentchev;andrey.avramov;todor.primov;vassil.momtchev}@ontotext.com

## Introduction

Delivering high quality gold standard data for any semantic annotation task is a time consuming process mainly because the manual curation work has to be consistent and semantically correct. In order to achieve usable quality, the different participants in the annotation process should agree in advance on each entity annotation.

Targeting the normalization of the entities, i.e., linking a particular chunk in a text to its canonical form or identifier in a database, is even harder in terms of software infrastructure, knowledge representation and defining a standardized format for representing the annotations. This task is important especially because the annotated texts should be integrated with other web resources, the Semantic Web and the linked data cloud<sup>15</sup>. The latter allows better search and ability to explore the data by different schemas and ontologies.

The CALBC challenge<sup>16</sup> advocates an alternative approach to reaching a high quality in the annotations that is a “silver standard”. It may result from harmonizing the automatically provided annotations. The automatic annotations are delivered by different groups and finally merged to form a compromise set of annotations.

Our objective in participating is to show that high quality dictionaries, along with software infrastructure that allows filtering, rewriting over dictionary instances, and efficient matching strategies over the text can be competitive to other more sophisticated approaches, particularly in large scale annotation tasks. Another aim is to integrate the semantic annotations created by the system with Linked Life Data<sup>17</sup> (LLD) and thus with the linked data cloud. This integration gives many benefits, including a powerful semantic and keyword search due to the linking of science literature to semantic annotations, as well as a fast access to the concept metadata, which on the other hand is already linked to other sources in the linked data cloud.

## Results And Discussion

LLD is a semantic repository for biomedical information that integrates data from more than 20 different public databases and more than 140 biomedical ontologies, vocabularies and thesauri. The data in the repository is represented in RDF triples, which allows the transformation of heterogeneous data into a common abstract data model. This abstract data model has been used to

---

\* corresponding author

<sup>15</sup> <http://linkeddata.org/>

<sup>16</sup> <http://www.calbc.eu/>

<sup>17</sup> <http://linkedlifedata.com/>

resolve semantic redundancy across different data sets and it provides a flexible mechanism for identifying heterogeneous biomedical concepts – via concepts URIs.

Our approach to semantic annotation is based on vocabularies from the UMLS MetaMap Strict Data Model (MetaMap SDM). In MetaMap SDM additional rules were applied over the set of names to (i) re-write the literals (to enrich the vocabulary with new literal forms) and (ii) suppress some literal forms (to filter low quality and noisy literals) (we extended the work of Hettne et. al. 2010). Various strategies to match the literal forms (e.i., the names and aliases of the concepts in MetaMap SDM) have been studied including direct or case insensitive as well as lemma or word stem matching,, normalization of names by ignoring numbers and punctuation, and different approximate string matching metrics. The best results from our experiments were observed with a combination of case insensitive matching against the word stem.

For each semantic annotation (or concept in UMLS) our annotator provides (i) the instance URI – a stable identifier of the UMLS concept that is also resolvable on the web, (ii) the class URI – a stable identifier of the UMLS Semantic Type or Semantic Group (Bodenreider and McCray 2003, McCray et. al. 2001), which is again resolvable on the web, and (iii) the Literal Value – the particular literal form that matches in this text offset.

The abstract texts subject to annotation are provided by the challenge organizers in a big xml file that may contain 714k documents. Our technology efficiently maps individual abstracts to a number of annotation applications developed in GATE<sup>18</sup>. This approach allows parallelizing the annotation process and annotating 714k docs (6144 b average size) for 02 hours 02 sec. on a desktop machine with 8 processor (we use 8 threads) at 2.5 Ghz and 15 GB RAM.

In addition to the submitted xml files, our annotation process also generates RDF triples, which represent semantic relations between the document URI and the instance URIs. Since the instance URIs are unique and resolvable on the web (and in LLD), they can be easily used to retrieve additional meta data (coming either from UMLS data sets or from any other data sets aligned to UMLS in LLD) with keyword searches and SPARQL<sup>19</sup> queries in LLD. We should mention that the document URIs are also present in LLD and this allows efficient searches in a space containing interlinked bio-medical text, semantic annotations, and their respective meta data.

## Conclusion

Since all of the concept URIs are resolvable in the LLD namespace, <http://linkedlifedata.com/resource/umls/id/>, the LLD integration platform can be used as a data provider. Third party annotation efforts can be easily supported by the LLD service. All that is required is to reuse the namespace, defined by LLD. For example, the additional annotation meta data of the concept C0024114 (Chronic Obstructive Airway Disease) can be retrieved from LLD using the following URI - <http://linkedlifedata.com/resource/umls/id/C0024114>. To help the efficient integration and enable a more powerful bio-medical search, we integrated the complete NCBI PubMed set of abstracts (including the CALBC challenge silver standard subset of 714k documents) in the LLD system. We also provided a text annotation service that accepts any web text resource or document and links them to the respective entities in LLD.

---

<sup>18</sup> <http://gate.ac.uk/>

<sup>19</sup> <http://www.w3.org/TR/rdf-sparql-query/>

## References

Kristina M Hettne, Erik M van Mulligen, Martijn J Schuemie, Bob JA Schijvenaars, Jan A Kors, Rewriting and suppressing UMLS terms for improved biomedical term identification, *Journal of Biomedical Semantics* (2010) 1:5.

Olivier Bodenreider and Alexa T. McCray, Exploring semantic groups through visual approaches, *Journal of Biomedical Informatics* 36 (2003) 414–432.

Alexa T. McCray, A Burgun, Olivier Bodenreider, Aggregating UMLS semantic types for reducing conceptual complexity, *Proceedings of Medinfo* 10 (2001) 216-20.

## List of attendees

Himanshu Agrawal	Linguamatics, United Kingdom hagrawal@linguamatics.com
David Campos	Universidade de Aveiro, Portugal david.campos@ua.pt
Simon Clematide	University of Zurich, Switzerland simon.clematide@cl.uzh.ch
Peter Corbett	Linguamatics, United Kingdom peter.corbett@linguamatics.com
Laura Croft	L.Croft@nature.com
Samuel Croset	EBI, United Kingdom samuel.croset@gmail.com
Max De Wilde	University of Antwerp, Belgium madewild@ulb.ac.be
Juliane Fluck	Fraunhofer Institute SCAI, Germany juliane.fluck@scai.fraunhofer.de
Martin Gerner	University of Manchester, United Kingdom martin.gerner@gmail.com
Christoph Grabmüller	EBI, United Kingdom grabmuel@ebi.ac.uk
Udo Hahn	Friedrich-Schiller-Universität Jena, Germany udo.hahn@uni-jena.de
Timo Hannay	Digital Science / Nature Publishing Group, London, U.K. timo@digital-science.com
Ian Harrow	Pfizer, United Kingdom ian.harrow@pfizer.com
Lynette Hirshman	The MITRE Corporation, U.S.A. lynette@mitre.org
Kerstin Hornbostel	Jena University Language & Information Engineering (JULIE) Lab, Germany kerstin.hornbostel@uni-jena.de

Ernesto Jimenez-Ruiz	Oxford University Computing Laboratory, United Kingdom ernesto.jimenez.ruiz@gmail.com
Senay Kafkas	EBI, United Kingdom kafkas@ebi.ac.uk
Jee-Hyub Kim	EBI, United Kingdom jhkim@ebi.ac.uk
Jan Kors	Erasmus University Medical Center, Netherlands j.kors@erasmusmc.nl
Ian Lewin	EBI, United Kingdom ian.lewin@ebi.ac.uk
Chen Li	EMBL-EBI, United Kingdom chenli@ebi.ac.uk
Sergio Matos	IEETA, Universidade de Aveiro, Portugal aleixomatos@ua.pt
David Milward	Linguamatics, United Kingdom david.milward@linguamatics.com
Gyorgy Mora	University of Szeged, Hungary gymora@inf.u-szeged.hu
Yves Moreau	Katholieke Universiteit, Leuven, Be yves.moreau@esat.kuleuven.be
Aliyu Kabir Musa	EMU, Cyprus aliyukabir.musa@emu.edu.tr
Victoria Nebot Romero	Universitat Jaume I, Spain romerom@lsi.uji.es
Andrew Needham	
Dietrich Rebholz-Schuhmann	EMBL-EBI, United Kingdom rebholz@ebi.ac.uk
Fabio Rinaldi	University of Zurich, Switzerland fabio@ontogene.org

Therese Vachon

Novartis Institutes for Biomedical Research, Basel, Ch

[therese.vachon@novartis.com](mailto:therese.vachon@novartis.com)

Erik van Mulligen

Erasmus Medical Center Rotterdam, Netherlands

[e.vanmulligen@erasmusmc.nl](mailto:e.vanmulligen@erasmusmc.nl)

Ying Yan

EBI, United Kingdom

[yan@ebi.ac.uk](mailto:yan@ebi.ac.uk)