The CALBC Silver Standard Corpus -

Harmonizing multiple semantic annotations in a large biomedical corpus

Dietrich Rebholz-Schuhmann
Antonio José Jimeno Yepes

EMBL Outstation - Hinxton
European Bioinformatics Institute
Hinxton, Cambridge, CB10 1SD, U.K.
{rebholz|yepes}@ebi.ac.uk

Erik M. van Mulligen
Ning Kang
Jan Kors
Department of Medical Informatics
Erasmus University Medical Center
NL-3000 Rotterdam, The Netherlands
{evanmulligen|j.kors}@erasmusmc.nl

David Milward
Peter Corbett
Linguamatics Ltd
St. John's Innovation Centre, Cowley Rd
Cambridge, CB4 0WS, U.K.
info@linguamatics.com

Udo Hahn

Language & Information Engineering (JULIE) Lab
Friedrich-Schiller-Universität
D-07743 Jena, Germany
udo.hahn@uni-jena.de

Abstract

The CALBC initiative aims to provide a large-scale biomedical text corpus that contains semantic annotations for tagged named entities of different kinds. The generation of this corpus requires that the annotations from different automatic annotation systems are harmonized.

In the first phase, the annotation systems from 5 participants (EMBL-EBI, EMC Rotterdam, NLM, JULIE Lab Jena, and Linguamatics) were gathered. All annotations were delivered in a common annotation format that included concept ids in the boundary assignments and that enabled comparison and alignment of the results.

During the harmonization phase, the produced results from different systems have been integrated into a single harmonised corpus ("silver standard" corpus) by applying a voting scheme. We give an overview of the processed data and the principles of harmonization – formal boundary reconciliation and semantic matching of named entities. Finally all submissions of the participants have been evaluated against the silver standard corpus. We found that species and disease annotations are better standardised amongst the partners than the annotations of genes and proteins.

The raw corpus is now available for additional named entity annotations. Part of the annotated corpus will be made available later for a public challenge. We expect that we can improve corpus building activities both in terms of the numbers of named entity classes being covered, as well as the size of the corpus in terms of annotated documents.

1 Introduction

The provision of gold standard annotated data is a time-consuming and costly process predominantly due to the manual curation work. We advocate the notion of a "silver standard" which results from the harmonization of annotations provided from automatic annotation systems. Different annotation groups deliver their meta data which, finally, is merged to form a compromise set. It has been shown in the past that a combination of annotation services can deliver a final result that exceeds the performance of any of the included solutions (Smith et al., 2008). We derive from this experiment that a combination of annotation solutions can deliver a large scale annotated corpus suitable to train text mining solutions for large-scale text mining tasks.

Current (biomedical) text mining experiments and challenges are based on relatively small corpora (usually in the order of 1,000 to 2,000 abstracts) in narrow sub-domains (e.g., human blood cells and transcription factors (Genia corpus),[1] inhibition of cytochrome P450 enzymes and oncology (Penn-BioIE corpus),[2] gene tagging and normalization (BioCreAtIvE I, II; Smith et al., 2008; Morgan

[1] http://www-tsujii.is.s.u-tokyo.ac.jp/~genia/topics/Corpus/

[2] http://bioie.ldc.upenn.edu/publications/~latest_release/data/

Proceedings of the 3rd International Symposium on Languages in Biology and Medicine,
Long Papers, pages 64-72, Jeju Island, South Korea, 8-10 November 2009.

et al., 2008),[3] protein/gene interactions (BioCreAtIvE II). These restrictions are mostly due to the fact that the manual annotation of such corpora is extremely time-consuming and costly. As a result, the annotated corpora are too small and too narrowly scoped to be useful for a large variety of other text mining themes and applications.

Within CALBC, we intend to create a much broader scoped and more diversely annotated corpus. We plan to have 150,000 Medline abstracts on immunology, a reasonably broad topic which is dealt with in more than 1M abstracts from the 18M abstract set in Medline, collaboratively annotated with 5 to 10 semantic types in the course of two public annotation challenges. Publicly available systems that we draw on are *Whatizit* (EBI: Rebholz et al., 2007), *Peregrine* (Erasmus Medical Center: Schuemie et al., 2007), *GeNo* (JULIE Lab, U Jena: Wermter et al., 2009), *MetaMap* (NLM: Aronson, 2001), *Abner* (U Wisconsin: Settles, 2005), *OSIRIS* (U Pompeu Fabra: Bonis et al, 2006) and *TerMine* (NaCTeM: Frantzi et al., 2000). Commercial systems using NER that will be encouraged to contribute include Collexis, Medscan (Ariadne Genomics), Tessi (Language & Computing), Temis, MedTAKMI (IBM), LingPipe, as well as Linguamatics, a project partner. Besides availability and costs, differences among these systems include coverage (e.g., entities used in genomic studies *vs*. all biomedical entities) and purpose (pure NER *vs*. integration in a specialized application, such as the identification of gene-disease relations). Hence, a performance comparison of these systems will be difficult to establish.

We anticipate that a total of 1 to 2M named entity annotations will be generated by running these systems on the CALBC corpus. Formal document metadata (e.g., sentence segmentation) will be added to this corpus prior to release for public annotation. Since all these systems have different application scopes – some will aim for high precision, whereas others will aim for better recall – the integrated corpus will have a broader scope compared to that of any individual system.

Using different NERs in such a collaborative effort will inevitably require a tremendous annotation integration effort and, thus, consistency issues will play a prominent role in corpus maintenance. The project's focus, from a methodological perspective, is to analyze the impact of various consensus models for generating a "silver standard" corpus (SSC). No manually annotated gold standard will be supplied. Individual systems can use the generated SSC to obtain analysis reports comparing their own annotations with the SSC.

A secondary goal of this project is to define a standardized format for representing the annotations contributed by the participants and comparing them effectively. Currently, the lack of such a format hinders progress in the evaluation of NERs.

1.1 CALBC's "Silver Standard" approach

The biomedical text mining research community has a long tradition of organizing text mining challenges (most notably the BioCreaTive I (Hirschman et al., 2005) and II (Krallinger et al., 2008) competitions) as a way to evaluate text mining techniques, sharing technical knowledge, and to improve the results from text mining systems. The CALBC corpus will be used to organize challenges where participants can download the corpus, annotate it with their own text mining solutions in a standard format, submit the annotated corpus to a central server and receive an assessment of their results through a fully automated analysis. The submissions can be contributed at any time during the challenge's open phase and comparative assessment reports can be obtained. At the end of that period, all submissions of annotated corpora will be used to generate the next fully annotated corpus which then will be used as an SSC for the next round of the challenge.

The first measurable result of these challenges is the number of annotations contributed to the corpus by the participants. This set of annotations can be analyzed by participant, by granularity of the annotation location (from individual sentences to groups of documents), by semantic categories (e.g., chemicals, proteins, diseases) and by reference terminology used for the annotation (e.g., using the Medical Subject Headings (MeSH)).

The second measurable outcome results from the comparison of the annotations of any given system to the SSC and their integration into the SSC. Consensus annotations are crucial for this phase and can be defined as annotations provided by a certain number (or proportion) of systems. These consensus annotations can then be used as a surrogate gold standard (hence, "silver standard")

[3] http://biocreative.sourceforge.net/biocreative_1_dataset.html and http://biocreative.sourceforge.net/biocreative_2_dataset.html

to compute the usual precision and recall metrics for each system.

Last but not the least, the organizers will make an effort to make the SSC available as part of the assessment infrastructure. This allows text mining system developers to automatically run their solutions against the available "silver standard". Where available, we will use publicly available gold standard annotations to assess annotations delivered to CALBC, and to evaluate the quality of the silver standard as a whole.

2 Experimental conditions for kick-off

In preparation for the collaborative annotation challenges, all institutions involved in the CALBC consortium conducted an experiment to test the feasibility of collaboratively annotating a corpus and comparing the annotations contributed by several partners. The institutions are: European Bioinformatics Institute (EBI), Erasmus Medical Center (EMC), Jena University Language & Information Engineering (JULIE) Lab, Linguamatics (LM), and the National Library of Medicine (NLM).

A small corpus (1,485 Medline abstracts) was created and annotated by each partner independently. This corpus prefigures the much larger annotated corpus (150,000 documents) to be delivered as the outcome of the kick-off phase.

2.1 Annotation guidelines

The following principles were applied for the annotation of the named entities:

- All the annotations will be based on XML, so that it is both machine and human readable.
- Inline annotation is preferred, although stand-off annotation is also supported.
- A namespace is used to identify the concept in the original knowledge source.
- The exact boundaries of the entity have to be specified.
- Annotation of the largest span of text within the same semantic type is preferred. In the case of *"lung cancer"*, e.g., only *"lung cancer"* is annotated rather than *"cancer"*. This means that nested entities in the same semantic group will not be annotated (but overlapping entities are still allowed).
- Given a semantic entity, if the system cannot decide on the identifier of that entity in a knowledge source all identifiers are provided.

The annotation of entities was done using the *e element* that encloses the text where entity/entities can be found (Rebholz-Schuhmann et al., 2006). The entities are identified within the knowledge source using the *id attribute* in the *e element.* The identifier of a given entity in a given data source is composed of the namespace of the knowledge source (e.g., "UMLS"), the identifier of the entity in this source (e.g., "C0001403"), the semantic type and the semantic group. If multiple identifiers may be assigned to the same text boundary (e.g., in cases of ambiguity), the pipe symbol is used to separate them.

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

After the entity identifier, specified above as (namespace:id:semantic type:semantic group), a colon indicates that there is a comma separated list of token identifiers. The following example illustrates this point:

```
<e id="UMLS:C0222601:T023:1,2|UMLS:C0006142:T191:
2,3"><w id="1">left</w> <w id="2">breast</w> <w
id="3">cancer</w></e>
```

In this example, *left breast* (i.e., tokens 1 and 2) is identified by UMLS:C0222601:T023, while *breast cancer* (tokens 2 and 3) is identified by UMLS:C0006142:T191.[4]

2.2 Evaluation procedure

As evaluation measures for the comparison of named entities in the annotated corpora we use standard precision, recall and F-score. We employ two different types of boundary alignments:

- **Exact match:** For each semantic group, the system's assignments of the named entity boundaries have to match exactly the entity boundaries in the SSC. This evaluation is meaningful in the sense that the participant has to achieve a high agreement concerning the boundaries of the entities annotated in the SSC.

[4] Note that there is no concept for *left breast cancer* in the UMLS

- **Nested match:** The boundary assignments from the participant's system ("evaluation set") have to include the boundaries of the SSC ("reference set"), i.e., the boundaries in the evaluation set cover an equal or larger span than the boundary assignments from the SSC in addition to including the SSC assignments of the concept id. This evaluation is meaningful in the sense that the system identifies the complete location of the entity corresponding with the semantic group.

3 Initial experimental results

All five partners relied on the Unified Medical Language System (UMLS) as a reference for the semantic categorization of their annotations and home-grown solutions for NER (e.g., for gene and protein name identification). It can be expected that the semantic types and groups from the UMLS are the most commonly used categorization framework. Yet, alternatives can be explored throughout the project.

The initial comparison discussed below is mainly focused on the partners EBI, EMC and NLM only, since they had already adapted their systems according to the annotation guidelines during a pre-project test phase. EMC delivered two types of annotations for the gene/protein entities. The other two partners (JULIE and LM) delivered their annotations at a later stage, once all the modifications to the annotation pipeline had been finalized and tested. Their contributions were incorporated in the harmonization phase (see Section 4.6).

In the next subsections the results for several annotations are shown. From these experiments, the best solution for the harmonization of the corpora was identified using a simple consensus model. In this consensus model, the annotations of a minimum of two partners had to agree on the location of an entity (nested boundaries) and the semantic type. This approach leads to a corpus with a large number of annotations, since these two requirements put only weak restrictions on the agreement between the annotation systems.

3.1 Initial annotation comparison

The initial assessment takes into consideration two major parameters, *viz*. boundaries and their reconciliation, on the one hand, and semantic type matching of recognized named entities, on the other hand.

Formal boundaries of named entities

The number of boundary assignments delivered from the three sites differed markedly (see Table 1): 14,955 (EBI), 59,934 (EMC) and 56,585 (NLM). The number of semantic annotations was even higher, since multiple assignments could be rendered for a single boundary assignment: 16,142 (EBI), 60,958 (EMC), and 88,442 (NLM). The comparatively low number of annotations for the EBI resulted from the fact that disease annotations were added only at a later stage of the analysis.

With regards to boundary assignments, 8,846 boundary assignments formed the core of exact agreement between EBI (59.2%) and EMC (14.8%). EBI and EMC share the same number of exact boundary agreements (8,846), but the percentage varies since EBI annotated a smaller set of entities leading to a bigger percentage of its annotations matching to EMC, and vice versa for EMC.

12,294 boundary assignments from EBI are nested in the EMC annotations (82.2%), and a smaller number of annotations, only 9,340 boundary assignments, from EMC are nested in EBI (15.6%). This shows that EBI selects narrower boundaries assignments for its annotations where the assignments are tightly coupled to the lexical resource. EMC exploits contextual information to identify genes/proteins even if the terms in the text deviate from the lexical resource.

Reference	EBI	EBI	EMC	EMC	NLM	NLM
Evaluation	EMC	NLM	EBI	NLM	EBI	EMC
Boundaries	14,955	14,955	59,934	59,934	56,585	56,585
exact	8,846	1,329	8,846	8,619	1,329	8,619
nested	12,294	14,094	9,340	54,561	1,358	9,684
exact [%]	59.2	8.9	14.8	14.4	2.3	15.2
nested [%]	82.2	94.2	15.6	91.0	2.4	17.1
Annotations	16,142	16,142	60,958	60,958	88,442	88,442
exact	7,620	1,128	6,889	5,115	1,044	5,128
nested	9,178	10,860	7,204	36,343	1,063	5,372
exact [%]	51.0	7.5	11.5	8.5	1.8	9.1
nested [%]	61.4	72.6	12.0	60.6	1.9	9.5

Table 1: Comparison of different annotation solutions on the initial corpus (1,500 Medline ab-

stracts). The annotations from one partner ("Evaluation", e.g. EMC) are compared against the annotations from another partner ("Reference", e.g. EBI). The comparison included the boundary assignments and the group assignments. A boundary assignment from the evaluation set has to nest the boundary assignment from the reference set to be counted as a nested match. The semantic group in the evaluation set has to be listed in the reference set to give a positive count (see Annotations section).

8,619 boundary assignments exactly agreed between EMC (14.4%) and NLM (15.2%), which is a very small number in comparison to the large set of annotations that was delivered from the EMC and NLM. 54,561 boundary assignments from EMC (91.0%) are nested in NLM. Less than this, only 9,684 boundary assignments from NLM (17.1%) are nested in EMC's data. This leads to the conclusion that EMC's annotations are nested in NLM's annotations, the latter using MetaMap (Aronson, 2001) for boundary segmentation and MeSH mapping.

Semantic types of named entities

In addition to the formal issue of boundary reconciliation we also compared the semantic categories assigned to each boundary.

7,620 semantic annotations in exact annotations are consistent between EBI and EMC (51.0%), and only 6,889 semantic annotations in exact annotations are consistent between EMC and EBI (11.5%). It shows that EBI has only annotated a smaller set of semantic types in the corpus. More types were added at a later stage.

9,178 semantic annotations from EBI are in agreement for nested annotations with EMC (61.4%). Less than this, only 7,204 semantic annotations from EMC, are in agreement for nested annotations with EBI (12.0%). 82% of matches of nested boundary assignments (EMC against the reference EBI) yield only 61.4% of agreed semantic annotations. EBI annotates a given boundary with several concept ids (16,142 semantic annotations over 14,955 boundary assignments), whereas EMC assigns a single concept id (60,958 semantic types for 59,934 boundaries). It is evident that mismatches in the semantic type of the concept id will inevitably lead to mismatches in the alignment of the semantic annotations.

The overall conclusion is that EBI's semantic annotations are covered in nested annotations from EMC. Accordingly, one harmonization step could use the nested annotations of EMC over EBI and could use the semantic type shared between both partners. This solution was chosen at a later stage as one of the harmonization rules.

5,115 semantic annotations in exact annotations are in agreement between EMC and NLM (8.5%), and 5,128 semantic annotations in exact annotations are in agreement between NLM and EMC (9.1%). This result is not surprising when taking into consideration that NLM uses other boundary assignments than EMC.

36,343 semantic annotations from EMC are contained as nested annotations of NLM (60.6%). Less than this, only 5,372 semantic annotations from NLM, are found as nested annotations of EMC (9.5%). Again it becomes obvious that EMC's annotations are in general contained in annotations from NLM.

3.2 Boundary reconciliation: Removal of stop words

The differences in the boundary assignments between EMC and EBI led us to the hypothesis that the annotations might be due, to some degree, to "uninformative" words that do not modify the semantic type of the annotations. We therefore removed these "uninformative" stop words (e.g., about, every, since) from the annotations which should yield better normalizations. The stop words found at the left or the right border of the entity annotations were removed from the annotated span. If this processing step generated empty annotations indicating that only stop words were contained in the annotation span, the full annotation span was removed. The stop word removal has several effects.

The overall number of boundary assignments remained unchanged for EBI and decreased for EMC (59,934 compared to 58,420, respectively) and NLM (56,585 compared to 54,930, respectively). About 1,500 annotations in either corpus were composed of stop words only and so were completely removed after this reconciliation step.

The number of agreements (exact or nested) between EBI and EMC did not change significantly (less than 1% change). The number of disagreements decreased by the number of removed boundaries in the EMC corpus.

The biggest improvement was seen in the agreement with EBI on the NLM corpus and with EMC on the NLM corpus. This was no surprise, since the stop word removal reduced the boundary assignments of the NLM more to the size that was used by EBI and EMC.

Summarizing, the stop word removal harmonized the annotations with the NLM, but did not much improve the harmonization between EBI and EMC.

3.3 Evaluation of noun phrase boundaries

Apart from stop word removal, we analyzed the compliance of the annotations with a noun phrase chunker. We expected that annotations that are embedded in well-formed noun phrases (NPs) are more likely to represent entities than more complex syntactic structures.

Reference	NP	NP	NP	EBI	EMC	NLM
Evaluation	EBI	EMC	NLM	NP	NP	NP
Boundaries	61,327	61,327	61,327	18,306	59,934	56,585
Agreement	4,722	16,862	51,564	13,529	48,206	16,496
Disagreem.	56,605	44,465	9,763	1,426	11,728	40,089
Recall	0.09	0.27	0.84	0.86	0.80	0.29
Precision	0.32	0.28	0.91	0.26	0.79	0.27
F-measure	0.15	0.28	0.87	0.40	0.80	0.28

Table 2: Comparison of the boundary assignments in the evaluation set against the noun phrase boundaries (exact match)

The relation between the syntactic structure of the sentences and the identification of entities was also part of the analysis. The first hypothesis is that the chunks are placed within an annotated named entity boundary. To test this hypothesis, the LT-Chunk shallow parser (Mikheev, 1996) was used to detect noun and verb phrases in text. As can be seen from Table 2, EBI and EMC annotations are largely contained in the NP annotation, while NLM ones are not. In a significant number of cases, the NP boundary assignment is linked to an entity that has not been annotated by EBI or EMC. Furthermore, the analysis shows that the NP boundaries often deviate from the boundary assignments of EBI and EMC. This is partly due to syntactical structures using prepositional attachments and coordination that have not been considered by the NP chunker, but could also be the result of NP chunking mistakes of an imperfect NP chunker.

3.4 Annotations for proteins and genes

The following analysis focuses on the annotations that were delivered for genes and proteins with the intention to get an overview of the semantic normalization that would be required at a later stage.

The categorization of named entities in the semantic groups proposed by (Bodenreider and McCray, 2003) requires revision before they can be used in the CALBC project. The semantic types "Amino Acid, Peptide, or Protein" and "Enzyme" are grouped in the semantic group "Chemicals & Drugs". This coarse categorization is not very supportive to the CALBC challenge, since the pilot partners and the text mining community at large distinguish between chemicals and proteins. On the other hand, the semantic group system provides a separate category for genes. From an information extraction perspective, it is unrealistic to distinguish protein named entities from gene named entities. In conclusion, genes and proteins should be grouped together. The pilot partners agreed to use the category "CHED" for "Chemical & Drugs", excluding the two before-mentioned semantic types for proteins and enzymes. On the other hand, the novel category "PRGE" consists of the semantic group "GENE" and will include the semantic types that are not anymore included in "CHEM".

Semantic Type	**Description**
T028	Gene or Genome
T086	Nucleotide Sequence
T087	Amino Acid Sequence
T116	Amino Acid, Peptide, or Protein
T126	Enzyme
T192	Receptor

Table 3: Semantic types defining the PRGE group. The left column lists the codes used by UMLS for the different semantic types described in the right column.

In our experiment, the annotations of proteins and genes were compared. Annotations were selected either based on the name space (using Uni-

Prot) for the EBI annotation or on semantic types denoting a gene or protein (see Table 3).

Table 4 shows the results of the assessment. Only the analysis for the nested boundary agreement is shown. The agreement on the genes/proteins is better than the agreement overall (see Table 1). Obviously, the resources for the genes/proteins as well as the mapping of acronyms to the scientific literature and the assignment to semantic types seem to be better standardized than the annotation of text passages with general UMLS concepts. Since EBI annotates a large number of concept identifiers with the same span of text, it is again obvious that the overall performance of the alignment of two annotated corpora is reduced.

Nested

Reference	EBI	EBI	EMC	EMC	NLM	NLM
Evaluation	EMC	NLM	EBI	NLM	EBI	EMC
Boundaries	4,933	4,933	8,215	8,215	8,312	8,312
Agreement	2,650	3,150	2,643	6,200	590	1,028
Disagreem.	2,283	1,783	5,572	2,015	7,722	7,284
Recall	0.54	0.64	0.32	0.75	0.07	0.12
Precision	0.32	0.38	0.54	0.75	0.12	0.13
F-measure	0.40	0.48	0.40	0.75	0.09	0.12

Table 4: PRGE pair-wise comparison. The boundary assignments of the evaluation set have to nest the boundaries from the reference set.

3.5 Annotations for Diseases

The semantic types selected for the analysis of diseases are enumerated in Table 5:

Semantic Types	**Description**
T047	Disease or Syndrome
T191	Neoplastic Process
T019	Congenital Abnormality
T048	Mental or behavioral Dysfunction
T050	Experimental Model of Disease
T190	Acquired Abnormality

Table 5: Semantic types defining the “Disease” group

In Table 6 we find that the different annotations unveil a large agreement between EBI and EMC, larger than the one found with proteins and genes (cf. Table 4)

Nested

Reference	EBI	EBI	EMC	EMC	NLM	NLM
Evaluation	EMC	NLM	EBI	NLM	EBI	EMC
Boundaries	4,091	4,091	3,881	3,881	4,021	4,021
Agreement	2,769	2,936	2,580	2,946	401	410
Disagreement	1,322	1,155	1,301	935	3,620	3,611
Recall	0.68	0.72	0.66	0.76	0.10	0.10
Precision	0.71	0.73	0.63	0.73	0.10	0.11
F-measure	0.69	0.72	0.65	0.75	0.10	0.10

Table 6: Disease pair-wise comparison. Again the boundary assignments of the evaluation set have to nest the boundaries from the reference set.

3.6 Corpus harmonization

The objective of the harmonization is to provide a corpus with acceptable quality obtained by the combination of the annotations provided by the various partners. The harmonization requires an understanding of the annotations provided by the partners and described above, as well as a comparison of the annotations and defining heuristics that can be used for harmonization.

The heuristics used in this first harmonization benefitted from the experiences gathered from the comparison of the corpora between EBI, EMC and NLM. The data now includes also the annotations from JULIE Lab and Linguamatics (LM). All annotated spans of text were selected whenever the annotations from at least two partners were in agreement. Annotations were considered to be in agreement, if (1) the annotations of one partner were the same or nested in the other partner’s annotation and, if (2) they both agreed on the semantic type. The results from this analysis were then used to assess the annotations from the different partners against it.

The following Tables 7, 8 and 9 show the comparison of the annotation sets from the different partners (evaluation sets) against the harmonized reference set for the different semantic types: genes/proteins (PRGEs), diseases and species, respectively. The PRGE and disease group follows the definitions presented previously. The species group is defined by a subset of UMLS semantic types defining organisms that are not related to population studies. All available annotated corpora were assessed against the SSC. EMC provided two corpora, one relying on the UMLS (EMC-U) while the second one is based on other resources (EMC-O).

Evaluation	NLM	EMC-U	EMC-O	JULIE	EBI	LM
Boundaries	485,260	485,260	485,260	485,260	485,260	485,260
Agreement	328,827	344,720	357,532	230,659	310,753	348,742
Disagreem.	156,433	140,540	127,728	254,601	174,507	136,518
Recall	0.68	0.71	0.74	0.48	0.64	0.72
Precision	0.80	0.85	0.90	0.74	0.75	0.58
F-measure	0.74	0.77	0.81	0.58	0.69	0.64

Table 7: SSC “Disease” nested comparison

Evaluation	NLM	EMC-U	EMC-O	JULIE	EBI	LM
Boundaries	646,971	646,971	646,971	646,971	646,971	646,971
Agreement	461,259	261,888	443,945	309,746	322,870	291,186
Disagreem.	185,712	385,083	203,026	337,225	324,101	355,785
Recall	0.71	0.40	0.69	0.48	0.50	0.45
Precision	0.53	0.77	0.51	0.82	0.62	0.68
F-measure	0.61	0.53	0.59	0.60	0.55	0.54

Table 8: SSC PRGE nested comparison

Corpus	NLM	EMC-U	EMC-O	JULIE	EBI	LM
Boundaries	618,221	618,221	618,221	618,221	618,221	618,221
Agreement	360,642	507,349	491,597	282,342	259,835	360,812
Disagrem.	257,579	110,872	126,624	335,879	358,386	257,409
Recall	0.58	0.82	0.80	0.46	0.42	0.58
Precision	0.90	0.93	0.97	0.84	0.78	0.71
F-meas.	0.71	0.87	0.87	0.59	0.55	0.64

Table 9: SSC species nested comparison

The harmonized corpus contains, for each category, about half a million named entity boundaries (485,260 for Disease, 646,971 for PRGE, and 618,221 for species.

The evaluation of the annotations from the different participants shows that the participants achieved the best agreement with the SSC for the identification of species (F-measure from 0.55 to 0.87, average is 0.71). We assume that this result is due to the fact that species names are used in a standardized way in the scientific literature.

For the identification of diseases, the best performing annotation solutions showed lower performance in comparison to the results from the species identification. On the other side, the spread of F-measure was smaller and on average the performance of all systems was similar to the species annotation (F-measure from 0.58 to 0.81, average is 0.71). Although all partners used UMLS as primary resource, their applied methods produced different results, which leads to the conclusion that none of the partner’s annotations fully agrees with the SSC.

The results for the annotations of the PRGEs show the lowest performance (F-measure from 0.53 to 0.66, average is 0.59). The performance is also significantly lower than the results report from gene annotation competitions (BioCreative I and II). It is known that the variability of the representation of PRGEs in the literature is high. Furthermore, the participants use different solutions to map family names of proteins to concept ids and therefore chose different solutions to accept or ignore such terms.

Altogether, this analysis shows that the harmonized corpus allows comparing annotation solutions with different origins.

4 Conclusions and Future work

The production of a gold standard corpus requires a significant amount of manual curation work. We advocate the notion of a “silver standard” which results from the harmonization of annotations provided from automatic annotation systems. Different annotation groups deliver their meta data which, finally, is merged to form a compromise set.

We elaborated on this idea, and merged annotations for genes/proteins, diseases and species from five contributing teams. Assuming, however, an even higher number of contributors, we certainly have to cope with more sophisticated consensus metrics than the one we used up until now (cf., e.g.; Rahman and Fairhurst, 2003) for a comprehensive discussion of combination strategies).

Previous research work has shown that the combination of annotation services can deliver results that are superior to any integrated solution (Smith et al., 2008). Another conclusion from this work is that the upper limit for the gene/protein mention recognition could be around 90% F-measure. The silver standard corpus integrates a larger number of semantic types. Achieving high accuracy on all semantic types seems to be difficult to reach, but we can also expect that the SSC enables new approaches to disambiguate all semantic types that are used in the same context.

As we invite the community to contribute to this effort in the course of two waves of annotation challenges (cf. the Call for Participation @ bionlp list), we will certainly have to shape our ideas how to achieve fair consensus, how to exclude or include outliers, and, also, how to eliminate malevolent contributors (spam annotations). Our forthcoming challenges and alignment experiments will show whether automatically supplying a consensus-based silver standard might really constitute a reasonable and qualitatively acceptable advancement over manually supplied gold standards given their costs in terms of training and supervising human expert annotators.

Acknowledgments

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

References

Aronson, A.R. Effective mapping of biomedical text to the UMLS Metathesaurus: The MetaMap program. In: *Proceedings of the 2001 AMIA Symposium*, pp.17-21 (2001).

Bodenreider, O., McCray, A.T. (2003). Exploring semantic groups through visual approaches, *Journal of Biomedical Informatics* 36(6): 414-432.

Bonis, J., Furlong, L.I., Sanz, F. (2006). OSIRIS: A tool for retrieving literature about sequence variants. *Bioinformatics* 22(20): 2567-69.

Frantzi, K., Ananiadou, S., Mima, H. (2000). Automatic recognition of multi-word terms. *International Journal of Digital Libraries* 3(2): 117-132.

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

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

Morgan, A.A., Lu, Z., Wang, X., Cohen, A.M., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.H., Torres, R., Krauthammer, M., Lau, W.W., Liu, H., Hsu, C.N., Schuemie, M., Cohen, K.B., Hirschman, L. (2008) Overview of BioCreative II gene normalization. *Genome Biol*. 9(S3).

Mikheev, A. Learning part-of-speech guessing rules from lexicon: extension to non-concatenative operations, In: *Proceedings of the 16th International Conference on Computational Linguistics*, August 5-9, 1996, Copenhagen, Denmark

Rahman, A., Fairhurst, M.C. (2003). Multiple classifier decision combination strategies for character recognition: A survey, *International Journal of Document Analysis and Recognition*, 5: 166-94.

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

Rebholz-Schuhmann, D., Arregui, M., Gaudan, S., Kirsch, H., and Jimeno, A. (2008) Text Processing through Web Services: Calling Whatizit. *Bioinformatics* 24(**2**): 296-98.

Schuemie, M.J., Jelier, R., Kors, J.A. (2007). Peregrine: Lightweight gene name normalization by dictionary lookup. In: *Proceedings of the Biocreative 2 Workshop*. Madrid, Spain, April 23-25, 2007, pp.131-140.

Settles, B. (2005), Abner: An Open Source tool for automatically tagging genes, proteins, and other entity names in text. *Bioinformatics*, 21(14): 3191-92.

Smith, L., Tanabe, L.K., Ando, R.J., Kuo, C.J., Chung, I.F., Hsu, C.N., Lin, Y.S., Klinger, R., Friedrich, C.M., Ganchev, K., Torii, M., Liu, H., Haddow, B., Struble, C.A., Povinelli, R.J., Vlachos, A., Baumgartner, W.A, Hunter, L., Carpenter, B., Tsai, R.T., Dai, H.J., Liu, F., Chen, Y., Sun, C., Katrenko, S., Adriaans, P., Blaschke, C., Torres, R., Neves, M., Nakov, P., Divoli, A., Maña-López, M., Mata, J., Wilbur, W.J. (2008) Overview of BioCreative II gene mention recognition. *Genome Biol*. 9(S2).

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