The CALBC Silver Standard Corpus for Biomedical Named Entities – A Study in Harmonizing the Contributions from Four Independent Named Entity Taggers

Dietrich Rebholz-Schuhmann[a], **Antonio José Jimeno Yepes**[a], **Erik M. van Mulligen**[b], **Ning Kang**[b], **Jan Kors**[b], **David Milward**[c], **Peter Corbett**[c], **Ekaterina Buyko**[d], **Katrin Tomanek**[d], **Elena Beisswanger**[d], **Udo Hahn**[d]

[a] EMBL Outstation, European Bioinformatics Institute,
Hinxton, Cambridge, CB10 1SD, U.K.

[b] Department of Medical Informatics, Erasmus University Medical Center
NL-3000 Rotterdam, The Netherlands

[c] Linguamatics Ltd
St. John's Innovation Centre, Cowley Rd, Cambridge, CB4 0WS, U.K.

[d] Language & Information Engineering (JULIE) Lab, Friedrich-Schiller-Universität
D-07743 Jena, Germany

E-mail: {rebholz|yepes}@ebi.ac.uk, {evanmulligen|j.kors}@erasmusmc.nl, info@linguamatics.com,
{ekaterina.buyko|katrin.tomanek|elena.beisswanger|udo.hahn}@uni-jena.de

Abstract

The production of gold standard corpora is time-consuming and costly. We propose an alternative: the ‚*silver standard corpus*' (SSC), a corpus that has been generated by the harmonisation of the annotations that have been delivered from a selection of annotation systems. The systems have to share the type system for the annotations and the harmonisation solution has use a suitable similarity measure for the pair-wise comparison of the annotations. The annotation systems have been evaluated against the harmonised set (630.324 sentences, 15,956,841 tokens).

We can demonstrate that the annotation of proteins and genes shows higher diversity across all used annotation solutions leading to a lower agreement against the harmonised set in comparison to the annotations of diseases and species. An analysis of the most frequent annotations from all systems shows that a high agreement amongst systems leads to the selection of terms that are suitable to be kept in the harmonised set. This is the first large-scale approach to generate an annotated corpus from automated annotation systems. Further research is required to understand, how the annotations from different systems have to be combined to produce the best annotation result for a harmonised corpus.

1. Introduction

The evaluation of NLP systems requires benchmark corpora to measure their performance. Manually curated gold standard corpora have emerged as a special type of language resources to serve as ground truth for the evaluation of NLP systems' quality and to train and test such systems that rely on (semi-)supervised machine learning approaches (Morgan et al., 2008). Unfortunately, the production of gold standard language resources is time-consuming and thus costly, since major portions of these resources were manually built, up until now.

In the CALBC project (http://www.calbc.eu), the partners involved investigate in an alternative solution which is intended to serve as an approximation to a gold standard annotated corpus that we will hitherto call a *'silver standard corpus'* (SSC, Rebholz-Schuhmann et al., 2010). Our approach should provide a corpus that has a large number of annotated entities which are based on large-scale terminological resources. The key to such a large-scale and fully automatically annotated corpus is the exploitation of different NLP systems that have been developed completely independently from each other and possibly trained on different language resources available in the scientific community.

We stipulate that by gathering and combining the contributions from diverse annotation engines we can capitalize on the performance of all involved systems in such a way that the overall result exceeds the performance of any single system (see BioCreative Meta-Server)[1]. This hypothesis will be tested empirically throughout the run-time of the CALBC project.

1.1. The CALBC corpus

The CALBC project partners (*viz.* European Bioinformatics Institute, Erasmus Medical Center, JULIE Lab of FSU Jena, and Linguamatics) have collected a corpus of 150,000 Medline abstracts. The selection of semantic groups for the annotation of the corpus is compliant with the semantic group system described by Bodenreider and McCray (2003). The corpus has already been made available to the biomedical NLP community. In two CALBC challenges, we ask researchers to enhance this corpus by their own annotations. We will integrate these contributions into the CALBC corpus according to well defined harmonization rules and will produce the CALBC silver standard corpus.

2. Methods

Altogether the input from four different named enity taggers was used as a source for the harmonization efforts. The annotations are based on IeXML inline annotation[2].

EBI. This system considers the following sets of semantic groups: proteins and genes (Rebholz-Schuhmann et al.; 2007) and disease annotations (Jimeno-Yepes et al., 2008). The identification of species uses the content of the NCBI taxonomy and matches the terms taking morphological variability into consideration.

[1] http://www.biocreative.org/news/chapter/metaserver/

[2] http://www.ebi.ac.uk/Rebholz-srv/CALBC/challenge_guideline.pdf

EMC. Annotations are generated by EMC's Peregrine concept recognizer (Schuemie et al., 2007). The annotation groups are the 135 semantic types from UMLS grouped according to NLM's definition into semantic groups. Peregrine uses a dictionary for term identification and then associates the concept with that term, including homonym disambiguation. Peregrine makes use of UMLS and other resources such as SwissProt, MESH, etc.

FSU Jena. This solution is based on the Jena Component Repository (JCoRe; Hahn et al., 2008). For named entity annotation, FSU Jena uses its GeNO (Wermter et al., 2009) which incorporates the JNET tagger (JCoRe, Hahn et al., 2008). Furthermore, FSU Jena matches MeSH terms using the Lingpipe chunker[3] that incorporates acronym detection results (Schwartz and Hearst, 2003) for disambiguation.

LGM. Linguamatics' solution is based on I2E that uses fuzzy matching techniques to recognize terms from ontologies in text, and disambiguation to remove false positives. The following resources contribute to the NER tasks of CALBC: MeSH, NCI Thesaurus, UniProt Tissues List, SNOMED-CT, ChEBI, MedDRA and Entrez Gene.

All systems share that they exploit the same terminological resources: UniProt for genes/proteins, NCBI taxonomy for species, and UMLS for disease annotations. The annotation systems still differ in the way how the text is processed. The size of this manuscript does not allow a more detailed comparison of the different systems.

2.1. Corpus harmonization and evaluation

Combining the annotations from different named entity taggers requires subsequent alignment of the annotations, for example pair-wise alignment of all tokens for a certain stretch of text. Several methods can be applied, once the alignment has been achieved: (1) exact matching of the annotation boundaries for entities of the same semantic group ("pair-wise exact matching"), (2) nested matching of one annotation inside of the second annotation of the same semantic group ("pair-wise nested matching"), and (3) a continuous similarity measure that scores the token similarity of the annotations over the stretch of annotated text ("cosine similarity matching").

For example, under the condition that the annotator A_1 annotates the phrase P_a = "T_1 T_2" and the annotator A_2 the phrase P_b = "T_1 T_2 T_3", there is no exact match between the two annotations. Since P_a is nested P_b, this would count as a nested match. For the cosine similarity we consider the inverse document frequency f_x = idf(T_x) of each token as a measure of relevance and calculate the cosine similarity between the vectors $v_1 = \langle f_1, f_2, 0\rangle$ and $v_2 = \langle f_1, f_2, f_3\rangle$. A match is accepted, if $\cos(v_1, v_2) \geq 0.98$, for example "lung cancer" as part of "rare lung cancer".

In the first step of the harmonization process, the results of two annotators are compared to determine the number of shared annotated tokens (e.g. $c(\langle T_1, T_2, T_3\rangle) = \langle 2, 2, 1\rangle$). In the next step the harmonised corpus is compared against the next annotator. In the last step, a stretch of text is accepted under the condition that at least two annotators from all four annotators have jointly annotated the same part of a given phrase.

The harmonised set generated with the 98% cosine similarity measure contains 15,956,841 tokens in 630.324 sentences.

[3] http://alias-i.com/lingpipe/

2.2. Most frequent shared annotations

For our analysis of the distribution of annotations, we selected the most frequent annotations for all three semantic types. Every mention in the corpus was annotated with the systems that have identified the given mention, i.e. the single token or sequence of tokens in the text, based on 98% cosine similarity matching. We then gathered all mentions together with the system profile and sorted them by their frequency. Finally we selected the one hundred most frequent pairs of a mention with its system profile and analysed the findings (see result section).

3. Results

A random selection of 150,000 abstracts ("CALBC Corpus") was taken from the complete hit set that resulted from running the query "immunology" on Medline[4] (1 million documents overall). Three semantic groups were then annotated by four different systems of the CALBC consortium on the CALBC Corpus: proteins and genes (PGN), diseases (DIS) and organisms (SPE). Every mention of an entity of these groups had to be annotated in the documents (named entity recognition task), but it was not necessary to identify the correct reference to any data entry in a reference data resource (name entity normalization task), although all participants contributed concept ids for the identified terms.

This data was then further harmonized. The harmonized CALBC Corpus based on pair-wise agreement for nested annotations contains in total (a) 780,836 annotations for 40,913 unique lexical items for diseases, (b) 1,251,374 annotations for 81,654 unique lexical items for PGNs, and (c) 715,043 annotations for 18,964 lexical items of species.

		DIS	PGN	SPE
All	s01234	11,303	4,924	6,633
Leave out one -	s0123	14,156	7,421	7,266
	s0124	12,118	6,421	8,911
	s0134	15,933	7,561	8,734
	s0234	11,717	8,145	13,266
	s1234	13,305	6,198	13,266

Table 1: In the first line the number of unique terms is listed, where all systems agreed on the annotation (DIS for disease, PGN for gene/protein, SPE for species, s01234 represents system 0 to system 4)[5]. Unique terms denote a unique stretch of text that can have several occurrences in the corpus ("mention"). The part below the first line shows the total figures of annotations, where one system did not have to agree on the annotation ("Leave out one"). Systems 0 and 1 share the same annotation solution for species and have been submitted by the same participant (EMC).

In the next step, the annotations between the different systems were harmonized to produce the silver standard corpus. After the harmonization, all annotation systems involved were assessed against the harmonized corpus to determine precision and recall, and to calculate the F-measure of the systems. All analyses were produced on the

[4] http://www.ncbi.nlm.nih.gov/pubmed/

[5] In the following, system P0, P1 = two submissions from EMC, P2=JULIE, P3=EBI, P4=LGM.

complete silver standard corpus. Table 1 gives an overview on the number of shared annotations for the disease (DIS), the gene and protein (PGN), and the species annotations (SPE).
Table 1 shows the overall numbers for the annotation of terms. The agreement between all systems is higher for the annotations of diseases (11,303) than for the PGNs (4,924) or for the species (6,633). We can conclude that the terminological resources for diseases are better standardised, or the other way around, the term variability linked to PGNs is much higher than for diseases.
If not all systems have to agree, then we gain an additional set of annotated entities. For diseases, the agreement is highest, if system 2 is left out (additional 4,630 annotations) and for PGNs, if system 1 is left out.

	DISO	PGN	SPE	DISO	PGN	SPE	DISO	PGN	SPE	DISO	PGN	SPE
	p1			p2			p3			p4		
p0	1,877	4,550	1,751	72	1,925	357	96	**5,598**	190	130	**2,213**	1,375
p1				96	945	756	327	**1,291**	331	285	820	1,375
p2							302	**2,789**	583	442	1,501	845
p3										1,313	1,552	352

Table 2: The table shows the cross-comparison between two systems for the different semantic groups. The counts report on unique terms (i.e. not mentions) that have been identified only by the referenced systems (e.g., S1 vs. S0 on top left). The table shows the agreement between two systems, if terms are not considered where additional systems provided the same annotation.

According to Table 2, pair-wise agreements for PGNs exist (that are different from annotations from other partners) between system 0 on the one side and system 1 or 3 on the other side. For diseases and for species the best pair-wise agreement for the outliers is between P0 and P1.

Disease	EMC (s1)	Jena (s2)	EBI (s3)	LGM (s4)
Pw. nested	430,025			
F-Measure	0.74	0.63	0.76	0.68
Cosine	420,892			
F-Measure	0.76	0.66	0.80	0.68

Table 3: Results from the different harmonisation solutions measured on the annotations for diseases. The F-measure refers to the measurement of the participant's system (see header row) against the harmonised set as the reference corpus. The performance for pair-wise nested comparison ("Pw. nested") has a similar performance in comparison to the 98% cosine score ("Cosine"). The reported F-measures are below state of the art solutions measured against gold standard corpora.

The harmonization of the corpus uses two different methods (see methods section): a pair-wise voting scheme of nested annotations between two systems (one could be the harmonized set) and the second being a cosine similarity comparison of the two annotation systems. Table 3 gives on overview on the results: the annotations from the participants' solutions have been evaluated against the harmonized corpus.
Examples of agreements from this cross-comparison step for diseases only are shown in Table 3. The minimum F-measure for the partners' solutions against the harmonised corpus has been measured at 0.63 (i.e. is the lowest value in this experiment). The best performing system achieves an F-measure of 0.80. The two different measures, i.e. pair-wise nested matching and 98% cosine similarity agreement, show similar performances. In the case of 98% cosine similarity, the average performance of all systems is higher (73.2% instead of 71.4%) in comparison to the alternative case of using pair-wise nested matching.

PGNs	EMC (s1)	Jena (s2)	EBI (s3)	LGM (s4)
Pw. nested	488,466			
F-Measure	0.52	0.58	0.60	0.62
Cosine	482,935			
F-Measure	0.52	0.58	0.60	0.62

Table 4: Results from the different harmonisation solutions measured on the annotations for genes and proteins (PGNs). For details concerning the labels in the table, please refer to table 3. The performance of the different systems on PGN annotations is lower than for disease annotations (see table 3).

When comparing the annotations for PGNs of all systems against the harmonised corpus (see table 4), it becomes obvious that the performance is lower. This signifies that the harmonised set shows high diversity of annotated entities for PGNs. The F-measure ranges from 0.52 to 0.62 independent from the method used for the harmonisation of the annotations, but the average F-measure is slightly higher for 98% cosine similarity over pair-wise nested voting (57.4% in comparison to 57.0%).

Specie	EMC (s1)	Jena (s2)	EBI (s3)	LGM (s4)
Pw. nested	483,853			
F-Measure	0.65	0.65	0.63	0.80
Cosine	469,347			
F-Measure	0.67	0.70	0.66	0.81

Table 5: Results from the different harmonisation solutions measured on the annotations for species (see table 3 for details on the labels). The performance of the different systems on specie annotations is similar to the one for disease annotations (see table 3).

Comparing the annotators for species against the harmonised sets, we find performances that are similar to the annotation of diseases against the harmonised set (see table 5). The average performance against the harmonised set based on pair-wise nested voting reaches 69.9% and for 98% cosine similarity it reaches 72.1% on average. This annotation set contains is larger than the one for disease entities.

3.1. Most frequent shared annotations

We analysed the distribution of annotations across the different annotators to better understand, how the different annotators contribute to the harmonised set (see method section). In our statistical analysis (see table 6), we identified the 100 most frequent mentions together with their annotation profile composed of the systems that attributed the same semantic type to the mention. This analysis gives an overview on the distribution of the

annotations across the most frequent mentions in the corpus.

	Disease		PGNs		Species	
	# Syst.	# Mention	# Syst.	# Mention	# Syst.	# Mention
Total	5	268,305	5	352,808	4	351,983
Maxim.	4	21,807	1	17,842	2	43,946
Minim.	5	1,062	2	1,287	3	864
	# Terms	# Occurr.	# Terms	# Occurr.	# Terms	# Occurr.
5 agree	20	2,173	4	2,905		
4 agree	14	3,636	9	2,979	16	3,187
3 agree	9	2,995	9	3,477	24	2,962
2 agree	11	2,962	25	2,581	41	4,653
1 agrees	46	2,487	53	4,124	18	2,174
Median/ Average	**14**	**2,851**	**9**	**3,213**	**21**	**3,244**

Table 6: The overview summarizes the findings across all annotations from all systems that have been gathered for the corpus. A single term in a given location (a "mention") can be annotated by one system or by the full number of systems (e.g. 5, see "# Syst."). For each type we have gathered the 100 most frequent combinations of a term mention together with the set of systems that have annotated this mention (see Method section). The table indicates that the annotations for diseases are more homogeneous, i.e. the systems agree better than for the other two types. The most frequent species mention ("as") has been excluded to show less distorted statistical figures, since this annotation is an outlier and a faulty annotation.

		s4	s3	s0	s1	s2	Sum
Count	Term Total =>	82	48	46	42	33	
7,245	tumor	4	3	0	1	2	5
3,597	disease	4	3	0	1	2	5
3,164	asthma	4	3	0	1	2	5
2,636	rheumatoid arthritis	4	3	0	1	2	5
2,314	tuberculosis	4	3	0	1	2	5
2,222	lupus erythematosus	4	3	0	1	2	5
2,121	arthritis	4	3	0	1	2	5
2,064	lymphoma	4	3	0	1	2	5
21,807	infection	4	3	0	1		4
4,375	tumors		3	0	1	2	4
4,255	infections	4	3	0	1		4
3,208	diseases	4		0	1	2	4
2,528	SLE	4		0	1	2	4
2,405	AIDS	4		0	1	2	4
14,290	disease	4		0	1		3
2,395	syndrome	4		0	1		3
2,351	cancer	4	3			2	3
6,830	LPS			0	1		2
5,397	absence			0	1		2
5,358	tumor		3			2	2
2,876	immunodeficiency	4	3				2
2,744	HBV			0	1		2
2,180	PHA			0	1		2
2,149	RA	4				2	2
9,322	strains	4					1
6,856	strain	4					1
5,522	infected	4					1
5,229	membrane	4					1
4,843	apoptosis	4					1
4,629	adhesion		3				1
4,510	HIV-1	4					1
3,486	disease	4					1
3,477	exposure	4					1
2,916	neurons	4					1
2,875	mucosal	4					1
2,824	secondary	4					1
2,549	phagocytosis	4					1

Table 7: The 100 most frequent term mentions together with the annotations from the 4 contributing partners have been analysed. The table shows to the left the number of mentions of the given term in combination with the profile of annotators that have marked this term (systems 0 to 4).

Terms occur twice in the list (e.g., tumor) if they have been annotated by a different set of annotators in different locations. This result happens if an annotator uses contextual information to select or deselect a term, or chooses to annotate as part of a larger more specific concept. Only mentions with at least 2,000 counts are show, except for the list of single annotations which has been truncated.

It becomes obvious that disease mentions have been identified with a higher agreement amongst the annotators in comparison to the other entity types: all five systems can agree on a significant portion of the mentions (20% of the mentions in the top 100 selected mentions). For Species and PGNs, only a smaller number of agreements can be found amongst the most frequent annotated mentions.

		s0	s3	s4	s2	s1	Sum
Count	Term Total =>	70	31	30	27	23	
4,410	IL-6	0	3	4	2	1	5
3,156	CD4	0	3	4	2	1	5
2,597	TNF-alpha	0	3	4	2	1	5
7,081	IL-2	0	3	4	2		4
6,841	IFN	0	3	4	2		4
3,163	IL-10		3	4	2	1	4
2,350	CD4	0	3	4	2		4
5,971	CD4	0	3		2		3
5,123	genes	0		4		1	3
5,094	gene	0		4		1	3
4,459	IL-4		3	4	2		3
4,071	IFN	0	3		2		3
8,801	IgE	0		4			2
5,258	receptor	0				1	2
3,900	sequence	0				1	2
3,831	features	0				1	2
3,812	receptors	0				1	2
3,464	protein	0	3				2
3,395	domain	0				1	2
2,920	CD4		3		2		2
2,515	sequences	0				1	2
2,415	glycoprotein	0	3				2
2,307	suppressor	0				1	2
2,083	homologous	0				1	2
17,842	IgG	0					1
13,309	protein	0					1
13,267	sera			4			1
9,571	proteins	0					1
8,994	monoclonal antibodies	0					1
8,516	peptide	0					1
8,404	IgM	0					1
7,721	monoclonal antibody	0					1
7,447	cytokines	0					1
6,489	IgA	0					1
6,162	peptides	0					1
6,046	CTL		3				1
5,859	cytokine	0					1
5,722	per	0					1
5,025	Il			4			1
4,951	complement	0					1
4,763	autoantibodies	0					1
4,715	PCR		3				1
3,909	components	0					1
3,724	antiserum	0					1
3,652	antisera				2	0	1
3,471	Th1		3				1
3,103	HBV		3				1
3,102	component	0					1

Table 8: The most frequent mentions of PGNs have been analysed to identify systematic errors between the annotation systems. The table has been generated according to the same selection procedure as table 7. The term mentions show the high variability of the mentions of PGNs. The 5 different annotators agree only on a small set of PGN mentions.

We have further analysed the most frequent mentions together with their annotations from the different annotators. Table 7 shows the high agreement of all five annotators for the first eight mentions of terms. "absence" is the only term with at least two agreements that

represents a faulty entry. The presence of the two separate entries for "tumor" (all five agree over only three agree) indicates that some annotators take context-sensitive decisions that are not followed by others. The lower part of the list shows the annotations that are only supported by system 4 (strains, strain, membrane, adhesion, exposure, secondary) and that could be considered to be of little relevance to the disease named entities.

		s0	s1	s3	s2	Sum
Count	Term Total =>	70	66	45	52	
10,504	rats	0	1	3	2	4
7,783	rat	0	1	3	2	4
5,181	HIV-1	0	1	3	2	4
4,995	HIV	0	1	3	2	4
3,499	bacteria	0	1	3	2	4
3,251	humans	0	1	3	2	4
2,765	viruses	0	1	3	2	4
2,303	Escherichia coli	0	1	3	2	4
2,095	pigs	0	1	3	2	4
13,681	virus	0	1		2	3
10,212	animals	0	1		2	3
8,812	murine	0	1	3		3
3,452	sheep	0	1	3		3
3,386	rabbits	0		3	2	3
3,350	bovine	0	1	3		3
3,337	thymus	0	1	3		3
2,921	HIV		1	3	2	3
2,794	parasite	0	1		2	3
2,400	cattle	0	1	3		3
2,035	E. coli	0	1	3		3
89,044	as	0	1			2
43,946	mice			3	2	2
40,465	human	0	1			2
8,970	strains	0	1			2
8,596	strain	0	1			2
6,993	individuals	0	1			2
6,458	recombinant	0	1			2
5,466	host	0	1			2
4,888	rabbit	0		3		2
4,843	recipients	0	1			2
3,762	bearing	0	1			2
3,739	BALB/c	0	1			2
3,647	HCV			3	2	2
3,465	individual	0	1			2
3,297	As	0	1			2
2,924	areas	0	1			2
2,875	area	0	1			2
2,451	Human	0	1			2
2,359	dogs			3	2	2
2,305	rat			3	2	2
12,271	mouse			3		1
3,066	EBV				2	1
2,355	human	0				1

Table 9: The table shows the distribution of specie mentions in the annotated sets. The table has been generated according to the same selection procedure as for table 7 and 8. Although there seems to be a high agreement amongst the systems, it is the case that system 0 and system 1 introduce general terms since they are not independent from each and produce similar results (strain(s), area(s), as/As, others).

The diversity of annotations for PGNs is high (see table 8). Only a few mentions are shared amongst all 5 systems and even a lower agreement amongst the annotators does not lead to a significant increase in mentions that are selected. Again, a few annotations are annotated according to contextual information (see CD4, IFN). A number of terms are too unspecific (gene, genes, components) to be shared over all annotators, other terms could be considered to be specific or unspecific (IgE, CD4) and finally, a number of terms form clearly mistakes (features, homologous, per, II, antiserum, component).

In the case of the annotation of species mentions, it is remarkable that the set of annotations shared amongst at least 3 annotators seems to contain only a few mistakes (thymus, parasite). A few terms can be considered to be unspecific (virus, animals). System 0 and 1 introduce unspecific terms since they use similar resources (strain(s), host, recipients, individual). A few annotated terms should be removed (as, area(s), recombinant, bearing). For the harmonised set, only system 1 has been considered thereafter which resolved systematic errors.

4. Conclusions

We have generated the first corpus that contains a very large number of annotations, that contains the annotations from several annotation systems and that has been generated fully automatically. We expect high benefits from the corpus for systems that train a NER solution against the corpus and then identify a large number of semantic groups from similar types of text. Participants from the general public can contribute different types of annotations (more specific ones or more general ones), can receive an assessment against the SSC with automatic text mining means from the EBI's Web site and can contribute to the next SSC.

5. Acknowledgements

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

6. References

Bodenreider, O., and A.McCray, A. (2003) Exploring semantic groups through visual approaches. J Biomed Inform, 36(6): 414-432.

Hahn, U., et al., (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, pages 1-7.

Hettne, K., et al. (2009) Rewriting and suppressing UMLS terms for improved biomedical term identification, submitted to AMIA 2009.

Jimeno Yepes, A., et al.. (2008) Assessment of Disease Named Entity Recognition on a Corpus of Annotated Sentences. BMC Bioinformatics 9, no. SUPPL. 3: Article S3.

Morgan, A.A., et al. (2008) Overview of BioCreative II gene normalization. Genome Biol. 9(S3).

Rebholz-Schuhmann, D., et al. (2007) EBIMed – Text crunching to gather facts for proteins from Medline. Bioinformatics 23(2): e237-e244.

Rebholz-Schuhmann, D., et al. (2010) 'CALBC Silver Standard Corpus.' J Bioinform Comput Biol. Feb;8(1):163-79.

Schuemie, M., et al.. (2007) Peregrine: Lightweight gene name normalization by dictionary lookup, *Proceedings of the Biocreative 2 workshop* 2007 April 23-25, Madrid, 131-140

Schwartz, A., and Hearst, M.A. (2003) A simple algorithm for identifying abbreviation definitions in biomedical

text, In: PSB 2003 – Proceedings of the Pacific Symposium on Biocomputing p. 451—462.
Wermter, J., et al. (2009) High-performance gene name normalization with GeNo. Bioinformatics, 25(6): 815—821.