Entity Recognition in Parallel Multi-lingual Biomedical Corpora: The CLEF-ER Laboratory Overview

Dietrich Rebholz-Schuhmann[1,2], Simon Clematide[1], Fabio Rinaldi[1], Senay Kafkas[2], Erik M. van Mulligen[3], Chinh Bui[3], Johannes Hellrich[4], Ian Lewin[5], David Milward[5], Michael Poprat[6], Antonio Jimeno-Yepes[7], Udo Hahn[4], and Jan A. Kors[3]

[1] Department of Computational Linguistics, University of Zürich, Ch
(rebholz,clematide,rinaldi)@ifi.uzh.ch

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

[3] Department of Medical Informatics, Erasmus University Medical Center,
Rotterdam
(kors,e.vanmulligen)@erasmusmc.nl, bqchinh@gmail.com

[4] Jena University Language & Information Engineering (JULIE) Lab,
Friedrich-Schiller-Universität Jena, Fürstengraben 30, D-07743 Jena
(udo.hahn,johannes.hellrich)@uni-jena.de

[5] Linguamatics Ltd, 324 Science Park, Milton Road, Cambridge CB4 0WG
(ian.lewin,david.milward)@linguamatics.com

[6] Averbis GmbH, Tennenbacher Strasse 11, D-79106 Freiburg
poprat@averbis.de

[7] National ICT Australia, Victoria Research Laboratory, Melbourne, Australia
antonio.jimeno@gmail.com

Abstract. The identification and normalisation of biomedical entities from the scientific literature has a long tradition and a number of challenges have contributed to the development of reliable solutions. Increasingly patient records are processed to align their content with other biomedical data resources, but this approach requires analysing documents in different languages across Europe [1,2].

The CLEF-ER challenge has been organized by the Mantra project partners to improve entity recognition (ER) in multilingual documents. Several corpora in different languages, i.e. Medline titles, EMEA documents and patent claims, have been prepared to enable ER in parallel documents. The participants have been ask to annotate entity mentions with concept unique identifiers (CUIs) in the documents of their preferred non-English language.

The evaluation determines the number of correctly identified entity mentions against a silver standard (Task A) and the performance measures for the identification of CUIs in the non-English corpora. The participants could make use of the prepared terminological resources for entity normalisation and of the English silver standard corpora (SSCs) as input for concept candidates in the non-English documents.

P. Forner et al. (Eds.): CLEF 2013, LNCS 8138, pp. 353–367, 2013.

The participants used different approaches including translation techniques and word or phrase alignments apart from lexical lookup and other text mining techniques. The performances for task A and B was lower for the patent corpus in comparison to Medline titles and EMEA documents. In the patent documents, chemical entities were identified at higher performance, whereas the other two document types cover a higher portion of medical terms. The number of novel terms provided from all corpora is currently under investigation.

Altogether, the CLEF-ER challenge demonstrates the performances of annotation solutions in different languages against an SSC.

1 Introduction

Advances in the research community are often driven by specific challenges, which are meant to benchmark the outcomes on a well defined task. Over recent years a number of challenges have been proposed that focus on different tasks for the development of innovative technologies: e.g. different CLEF challenges such as CLEFeHealth and CLEF-IP [3,4], the BioCreAtIve sequel [5,6], the bioNLP Shared Tasks [7], and the CALBC challenge [8,9].

Most challenges propose a gold standard corpus that is then used for the benchmarking of the proposed solutions. In addition, other challenges have been proposed that consider a silver standard corpus instead. This approach allows the processing of large corpora in contrast to the gold standard approaches.

The CLEF-ER challenge is unique in the sense that it combines different expectations and technologies, such as entity recognition in the biomedical domain with multilingual approaches and machine translation.

Furthermore, the CLEF-ER challenge anticipates the processing and management of large resources and will exploit the delivered results for the development of augmented terminological resources.

2 Background

The CLEF conference sequel has a long tradition in setting up challenges for the research community. The challenge tasks are concerned with information retrieval, covering different types of electronic data, e.g. images, texts, and their combinations, and also considering different domain knowledges, for example medical and clinical data in comparison to legal texts and patents. All challenges are organised as part of a CLEF laboratory and the overall conference serves the purpose of the exchange of information.

Other challenges in the biomedical research community are also focused to information retrieval, namedly in TREC Genomics [10], but tackle in addition other tasks such as information extraction, entity recognition and fact extraction. The BioCreAtIve challenges are tuned to develop solutions that would help biomedical curators to do their work in finding facts from the literature [11]. The BioNlp Shared Task serves the same purpose and increasingly seeks the integration between ontological resources and the text mining component. Recently the BioASQ[1] challenge has been introduced, which aims at the tasks of topic identification and question answering in the biomedical domain.

None of the challenges has been organized in a way to feed the results from the challenge into building resources as it is the case for the CLEF-ER challenge and the MANTRA[2] project.

Furthermore, most challenges make use of a gold standard corpus (GSC) to evaluate the contributions from the participants. There is no doubt that a GSC is a precious resource and forms the key means to determine novel standards for a specific task in the research community. On the other side, it has been shown that GSCs are selective in the sense that they limit the evaluation of the specific tasks to a relatively small number of samples as instances representing the standard. By contrast, it is important to develop resources and standards at a scale that are more representative for the underlying tasks and the long-term goals.

The CALBC challenge has been such an initiative that was tackling the annotation of a large-scale corpus in the biomedical domain with a significant number of named entities for the benefits of long-term development of entity recognition solutions. The project partners have prepared a lexical resource, a large-scale annotated corpus, and a triple store containing the facts from the scientific literature covering the information in the annotated corpus.

The MANTRA project and the CLEF-ER challenge extend the work from the CALBC challenge into the development of multilingual resources for the medical domain. With the help of parallel corpora and a multilingual terminological resource, the project partners motivate the participants in the CLEF-ER challenge to contribute annotations in an English and a non-English corpus. The final goal is the annotation of medical entity mentions in the non-English corpus

2.1 Overview

This manuscript gives an overview on the setup of the CLEF-ER challenge including the resources that have been developed, the evaluation parameters and the outcomes of the challenge. The next section ("Material and Method") explains the provided resources, i.e. the terminological resources and the parallel corpora, as well as the evaluation metrics and the generation of the SSCs. Towards the end of the section, an overview on the contributing systems by the participants is given. In the results section, the performances of the systems overall is shown and the performances in dependence of the available corpora, the semantic groups from UMLS, and the different approaches from the participants. In the conclusion section, we will give views on the outcome of the challenge overall.

[1] http://www.bioasq.org/

[2] http://www.mantra-project.eu/

3 Material and Method

3.1 Terminologies

The MANTRA Terminological Resources (MTR) [12] used for the CLEF-ER challenge were derived from the Unified Medical Language System (UMLS) Metathesaurus [13]. The UMLS Metathesaurus is an umbrella system combining over 100 biomedical terminologies, e.g. the Medical Subject Headings (MeSH), the Medical Dictionary for Regulatory Activities Terminology (MedDRA, [14]) or the Systematized Nomenclature Of Medicine Clinical Terms (SNOMED-CT, [15]). The UMLS Metathesaurus contains both hierarchical (e.g. 'isa') and associative (e.g. 'caused by') relations between its entries, called *concepts*. Each concept is identified by a Concept Unique Identifier (CUI) and can have multiple names per language, called *synonyms*. Concepts are organized by semantic types (e.g. 'steroid'), which are themselves organized into semantic groups (e.g. 'chemicals & drugs'). To derive the MTR from the UMLS Metathesaurus we selected a subset containing only entries from selected semantic groups, e.g. anatomy (ANAT). This was done both due to the lower frequency and perceived irrelevance of the other semantic groups. The MTR contain 531,466 concepts with 2,839,277 synonyms (cf. tbl. 1 for details).

The MTR were distributed to the participants as a single file in the OBO format [16], which was selected both due to existing tooling and its readability for humans. The MTR is provided through the submission site of the CLEF-ER challenge[3] and requires a proper UMLS license from the participants.

Table 1. (Terminological resource): The English part of the TR contains most terms. Only Spanish is covered in SNOMED-CT. MedDRA terms have been translated in all languages.

Terms	MeSH	SNOMED-CT	MedDRA
en	764,000	1,184,005	56,061
de	77,249	-	50,128
fr	105,758	-	49,586
es	59,678	1,089,723	49,499
nl	40,808	-	50,932

3.2 Selection of Parallel Corpora

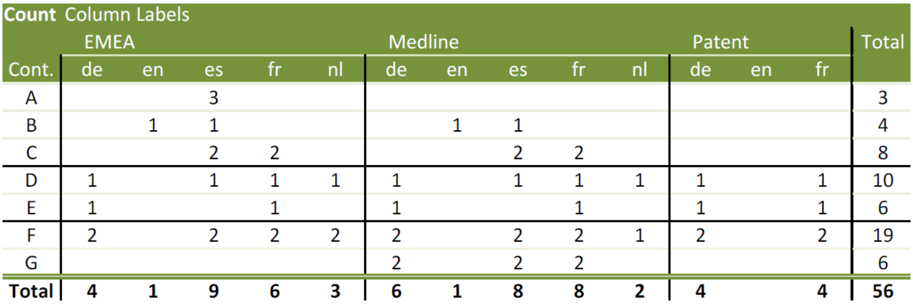
Different corpora have been selected and tested as input to the CLEF-ER challenge [12]. The parallel corpora have to be available in different (European) languages, should be available in languages that are shared between the different corpora, should have a reasonable size, and should deal with biomedical topics. The selection of Medline abstracts and EMEA drug labels fulfills the requirements. In addition, patent claims have been selected from patents that cover

[3] https://sites.google.com/site/mantraeu/terminology

Table 2. (Units counts, all corpora): The number of units is highest in English for Medline. German and French are evenly well covered in all three corpora, and Spanish shows similar coverage, except that Spanish (and Dutch) are not represented for patent texts.

Units	EMEA	Medline	Patent
en	140,552	1,593,546	120,638
de	140,552	719,232	120,637
fr	140,552	572,176	120,636
es	140,552	247,655	
nl	140,552	54,483	

Table 3. (Submissions to the CLEF-ER challenge): The Table gives an overview on the submissions to the CLEF-ER challenge. For all corpora and for all languages at least one annotated corpus has been contributed.



Count	Column Labels													
	EMEA					Medline					Patent			Total
Cont.	de	en	es	fr	nl	de	en	es	fr	nl	de	en	fr	
A			3											3
B		1	1				1	1						4
C			2	2				2	2					8
D	1		1	1	1	1		1	1	1	1		1	10
E	1			1		1			1		1		1	6
F	2		2	2	2	2		2	2	1	2		2	19
G						2		2	2					6
Total	**4**	**1**	**9**	**6**	**3**	**6**	**1**	**8**	**8**	**2**	**4**		**4**	**56**

biomedical topics. In the latter case, the language in the documents different from the scientific language, but the documents form an important part of the biomedical domain.

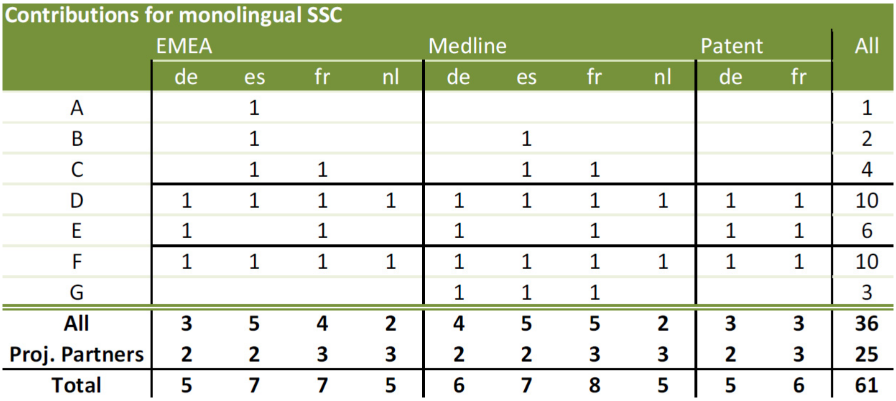
All corpora have been processed and transformed in a representation that linking the non-English text (called "units") to the English part of the same document. For Medline abstracts a single unit is a Medline[4] title, for the EMEA[5] drug labels individual paragraphs from the documents form a unit each, and for the patent texts the claim section forms a unit. The overall statistics are shown in the table above (cf. tbl. 2).

Beware that the parallel corpora for patent texts provide the complete claim section in three languages, i.e. in en, de and fr, whereas for the EMEA drug labels the complete documents are delivered in five languages (en, de, fr, es and nl). For the Medline titles, the parallel units are mostly in two languages, i.e. in English and in one non-English language again covering de, fr, es and nl. The reason for this lack of congruency is the fact that the non-English Medline titles

[4] http://www.nlm.nih.gov/pubs/factsheets/medline.html

[5] http://www.ema.europa.eu/

Table 4. (Generation of the SSC from CLEF-ER submissions): The Table gives an overview on the submissions to the CLEF-ER challenge. For all corpora and for all languages at least one annotated corpus has been contributed. The voting threshold has been set to 3, which is 50 % of the contributions.



Contributions for monolingual SSC											
	EMEA				Medline				Patent		All
	de	es	fr	nl	de	es	fr	nl	de	fr	
A		1									1
B		1				1					2
C		1	1			1	1				4
D	1	1	1	1	1	1	1	1	1	1	10
E	1		1		1		1		1	1	6
F	1	1	1	1	1	1	1	1	1	1	10
G					1	1	1				3
All	**3**	**5**	**4**	**2**	**4**	**5**	**5**	**2**	**3**	**3**	**36**
Proj. Partners	**2**	**2**	**3**	**3**	**2**	**2**	**3**	**3**	**2**	**3**	**25**
Total	**5**	**7**	**7**	**5**	**6**	**7**	**8**	**5**	**5**	**6**	**61**

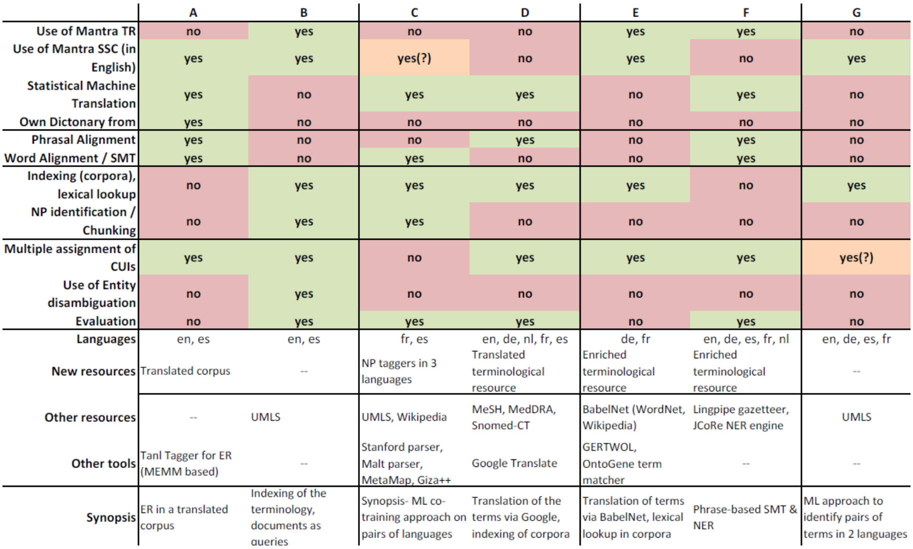
stem from documents that have been delivered from non-English journals and the title has been translated into English and not into any other language.

3.3 Preparation of the Silver Standard Corpus

Commonly systems are trained with and evaluated against gold standard corpora created by human experts. Due to the human involvement those are both expensive to create and limited in size. MANTRA follows the CALBC approach of using silver standard corpora (SSCs) instead [9], which are created by harmonizing multiple automatically annotated contributions. A voting scheme is used to determine which annotations are included in the SSC, e.g. only those annotated by a majority of systems. An SSC can be used to evaluate the contributions it was created from with standard metrics like f-score, yet this evaluation can only judge the averageness of a contribution and not its objective quality. We also created a variant SSC from de-annotated contributions, i.e. contributions from which those annotations trivially derived from the MTR were removed. This SSC was then used to evaluate the de-annotated contributions, allowing a better judgment of the conformity regarding new terms, which are otherwise obscured by the enormous amount of terms already contained in the MTR.

Monolingual Mention Evaluation (Evaluation A). In order to assess the quality of the annotations in all non-English corpora, a mention agreement evaluation against a harmonized Silver Corpus built from the monolingual contributions of the participants and from annotations from project partners was performed. Table 4 shows the number of annotations from the contributors and partners for the centroid-based SSCs [17]. Not all available contributions have been used to

Table 5. (Overview on the CLEF-ER participants systems): The description of the systems that have contributed to the CLEF-ER challenge shows high diversity across the approaches used from the participants. Most participants of the challenge made use of external resources either for their terminology or for word or phrase alignments.



	A	B	C	D	E	F	G
Use of Mantra TR	no	yes	no	no	yes	yes	no
Use of Mantra SSC (in English)	yes	yes	yes(?)	no	yes	no	yes
Statistical Machine Translation	yes	no	yes	yes	no	yes	no
Own Dictonary from	yes	no	no	no	no	no	no
Phrasal Alignment	yes	no	no	yes	no	yes	no
Word Alignment / SMT	yes	no	yes	no	no	yes	no
Indexing (corpora), lexical lookup	no	yes	yes	yes	yes	no	yes
NP identification / Chunking	no	yes	yes	no	no	no	no
Multiple assignment of CUIs	yes	yes	no	yes	yes	yes	yes(?)
Use of Entity disambiguation	no	yes	no	no	no	no	no
Evaluation	no	yes	yes	yes	no	yes	no
Languages	en, es	en, es	fr, es	en, de, nl, fr, es	de, fr	en, de, es, fr, nl	en, de, es, fr
New resources	Translated corpus	--	NP taggers in 3 languages	Translated terminological resource	Enriched terminological resource	Enriched terminological resource	--
Other resources	--	UMLS	UMLS, Wikipedia	MeSH, MedDRA, Snomed-CT	BabelNet (WordNet, Wikipedia)	Lingpipe gazetteer, JCoRe NER engine	UMLS
Other tools	Tanl Tagger for ER (MEMM based)	--	Stanford parser, Malt parser, MetaMap, Giza++	Google Translate	GERTWOL, OntoGene term matcher	--	--
Synopsis	ER in a translated corpus	Indexing of the terminology, documents as queries	Synopsis- ML co-training approach on pairs of languages	Translation of the terms via Google, indexing of corpora	Translation of terms via BabelNet, lexical lookup in corpora	Phrase-based SMT & NER	ML approach to identify pairs of terms in 2 languages

generate the SSC for the evaluation of the participants, because a contributor with several similar contributions would gain too much votes in favor of his system and the SSC would therefore be biased. The decision, which annotated corpus will be included into the SSC production, has been left with the challenge participant. All monolingual SSCs used a voting threshold of 3. Spanish and French are well-resourced in terms of different annotations. For German and especially Dutch, the number of contributions is less optimal.

Cross-lingual Concept Evaluation (Evaluation B). Given the fact that the English terminology covers a lot more concepts and provides more synonyms for them compared to the non-English terminologies, a second evaluation of concept coverage against a harmonized English Silver Standard Corpus built from the Mantra project partners was performed. For each corpus there are 6 different annotations that are harmonized into a centroid-based Silver Standard using a voting threshold of 3. The technical details of the centroid approach for the partner annotations as well as a detailed evaluation of the effect of different voting thresholds can be found in[18]

3.4 Participation and Contributions

Seven groups participated into the CLEF-ER challenge and contributed annotated corpora for the evaluation. Table 5 gives an overview on the approach that has been tested and links the system description to the performance of the tested solutions. As can be seen in tables 3 and 4 the participants contributed different numbers of annotated corpora and in general did not cover all languages. Spanish was the most popular language, i.e. the Spanish corpora have been annotated by the largest number of participants, and the largest number of submissions were linked to Spanish. French was a little bit more popular than German and the least contributions – as expected – were delivered for Dutch. These figures are relevant for the evaluation of the challenge, since a larger number of contributions leads to a larger set of annotated corpora that can be considered for the generation of a SSC in a given language.

Four of seven groups (A, C, D, and F) did apply methods that are linked to statistical machine translation or multi-lingual word alignment. Almost all groups used publicly available resources such as UMLS, Wordnet, Wikipedia and most groups also applied lexical lookup solutions or indexing of the terminological resources. Two groups translated the terms through public resources (i.e. BabelNet, group E) or with the Google translate infrastructure (group D). Altogether, the heterogeneity of the used solutions was high, and it became clear that the CLEF-ER challenge profits from machine translation solutions, although the challenge was announced as an entity recognition task.

Not all submissions were considered to be included for the generation of the SSC, which is based on the annotated corpora by the MANTRA project partners and the CLEF-ER participants (cf. tbl. 4). It is important to avoid that one or several participants dominate the outcome of the SSC by contributing a large number of annotated corpora. Therefore, the participants have been asked to point out one corpus that should server as their contribution to the challenge.

4 Resource and Evaluation

4.1 Silver Standards, Multilingual Documents

Table 4 gives an overview on the contributions to the monolingual SSCs. For each corpus and for all covered languages, one SSC has been produced from the MANTRA project partners' contributions to enable task A evaluation, i.e. the mention evaluation, and for the task B evaluation, i.e. CUI assignment. Only for the variant of the task B evaluation, where the trivial annotations have been removed (the "deannotated" corpus) the participants' contributions have been added as well.

In total 36 contributions have been received as part of the challenge, and another 25 annotated corpora have been provided from the MANTRA project partners prior to the challenge termination. Two participants contributed 10 annotated corpora, one for each language and for each corpus, and the other participants provided a smaller number of annotated corpora¿

Evaluation of Challenge Contributions

Two different tasks (and evaluations) have been suggested to the participants. In the evaluation A, the entity annotations are compared against an SSC to

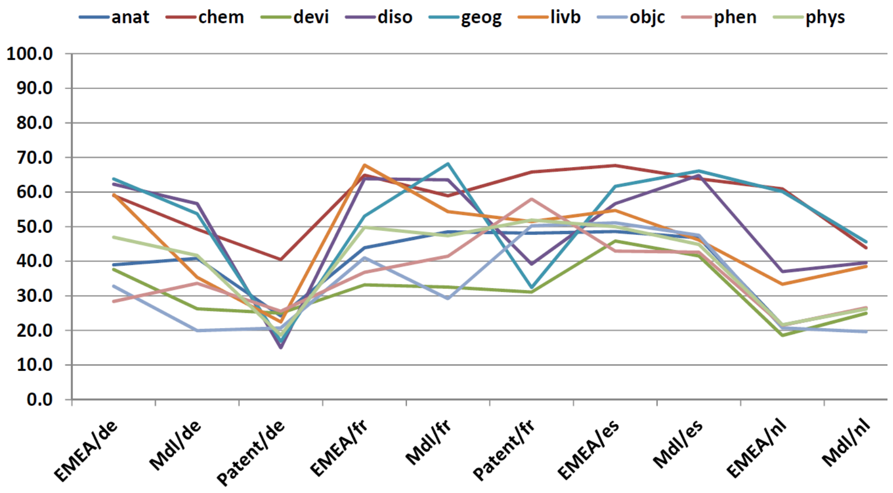


Fig. 1. (Precision, recall and F1-measure for the Evaluation B): All contributions have been evaluated concerning their assignment of the CUI. The evaluation was performed against the English SSC. The figure shows the average precision, recall and F1-measure of all solutions. Note that the both values for precision and recall are above the F1-measure for the EMEA/es corpus, since the diagram shows average figures for all annotation solutions together.

measure the boundary agreement of the participants against the SSC, where the SSC has been produced from annotated contributions from the MANTRA project partners.

In the evaluation B, the CUI assignment in the annotated corpus is evaluated against the prepared English SSC. In this task the participants have to assign the right CUI to a text stretch, which could be the complete unit of the parallel corpus, and the evaluation also does not consider any annotations in the text, but only evaluates against the correct assignment of a CUI to a unit.

Evaluation A and B are complementary in the sense that the boundary annotation (evaluation A) may give the correct mention of an entity, but the entity may still belong to different CUIs, and the correct CUI or mention normalisation may identify the correct concept (or entity), but the assignment to a particular stretch of text is left open.

The first task has been approached in a number of challenges, but not yet in the multi-lingual case covering a large amount of documents. The second task is typical for the biomedical domain and targets the normalisation of entities in non-English documents. This task has not yet been addressed in the multilingual case covering a large amount of parallel documents.

CUI Assignment (Task B). The participants had to produce annotations for their preferred corpus in their preferred languages, which should cover at least

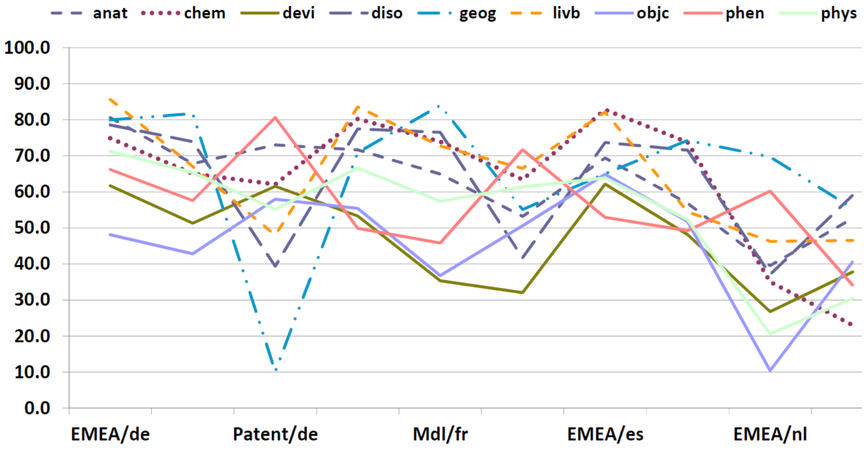


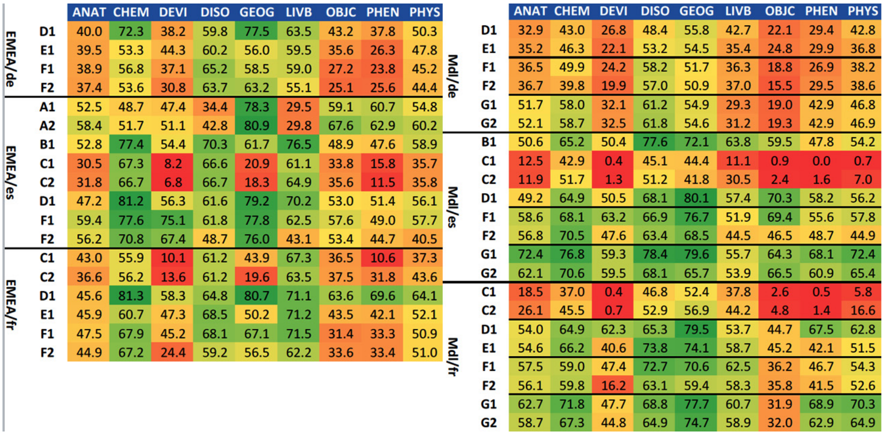
Fig. 2. (Evaluation B for semantic groups): The average F1-measure across all contributing systems has been calculated per semantic group of the annotations

one non-English language. The annotations had to comprise the assignment of a CUI to the entity mention. As can be seen from the system descriptions (cf. tbl. 5), the participants used different kinds of technologies including the translation of the terminology, the alignment and matching of concept mentions, and the translation of the corpus with the identification of corresponding concepts. The comparison of the CUI assignment in the non-English corpus against the English SSC formed the first evaluation and led to the following results (cf. fig. 1). The F1-measure performance over all contributing systems is better for Medline than for EMEA in all languages except for German, and for all languages the precision is higher in Medline than in EMEA. The F1-measure performance for the German patents (19 %) is a lot lower than for the other two corpora in German, and to a certain extend lower for the annotation of the patents in French in comparison to the other two corpora in French. This result indicates that the identification of entities and on concepts in patent documents is more complex than in the scientific biomedical literature, but the F1-measure for the other corpora ranges between 38 % and 48 %.

Table 6 shows the results for individual participants. The performance of the different solutions shows high heterogeneity, i.e. some entity types are identified well from selected solutions, but not in general across the corpus. As explained before, the annotation of French and Spanish text led to better performances than the annotation of German texts.

CUI Assignment per Semantic Group (Task B). The CUIs of the annotations can be categorized according to the semantic group that has been assigned to the CUIs. This grouping can be used to differentiate the performances according to the semantic groups and to give a more detailed analysis on the annotation of

Table 6. (Evaluation B, F1-Measure, challenge's participants): The table to the left shows the individual F1-measure performances of the participants in the evaluation B on the EMEA corpus and on the right for the Medline titles



		ANAT	CHEM	DEVI	DISO	GEOG	LIVB	OBJC	PHEN	PHYS
EMEA/de	D1	40.0	72.3	38.2	59.8	77.5	63.5	43.2	37.8	50.3
	E1	39.5	53.3	44.3	60.2	56.0	59.5	35.6	26.3	47.8
	F1	38.9	56.8	37.1	65.2	58.5	59.0	27.2	23.8	45.2
	F2	37.4	53.6	30.8	63.7	63.2	55.1	25.1	25.6	44.4
EMEA/es	A1	52.5	48.7	47.4	34.4	78.3	29.5	59.1	60.7	54.8
	A2	58.4	51.7	51.1	42.8	80.9	29.8	67.6	62.9	60.2
	B1	52.8	77.4	54.4	70.3	61.7	76.5	48.9	47.6	58.9
	C1	30.5	67.3	8.2	66.6	20.9	61.1	33.8	15.8	35.7
	C2	31.8	66.7	6.8	66.7	18.3	64.9	35.6	11.5	35.8
	D1	47.2	81.2	56.3	61.6	79.2	70.2	53.0	51.4	56.1
	F1	59.4	77.6	75.1	61.8	77.8	62.5	57.6	49.0	57.7
	F2	56.2	70.8	67.4	48.7	76.0	43.1	53.4	44.7	40.5
EMEA/fr	C1	43.0	55.9	10.1	61.2	43.9	67.3	36.5	10.6	37.3
	C2	36.6	56.2	13.6	61.2	19.6	63.5	37.5	31.8	43.6
	D1	45.6	81.3	58.3	64.8	80.7	71.1	63.6	69.6	64.1
	E1	45.9	60.7	47.3	68.5	50.2	71.2	43.5	42.1	52.1
	F1	47.5	67.9	45.2	68.1	67.1	71.5	31.4	33.3	50.9
	F2	44.9	67.2	24.4	59.2	56.5	62.2	33.6	33.4	51.0

		ANAT	CHEM	DEVI	DISO	GEOG	LIVB	OBJC	PHEN	PHYS
Mdl/de	D1	32.9	43.0	26.8	48.4	55.8	42.7	22.1	29.4	42.8
	E1	35.2	46.3	22.1	53.2	54.5	35.4	24.8	29.9	36.8
	F1	36.5	49.9	24.2	58.2	51.7	36.3	18.8	26.9	38.2
	F2	36.7	39.8	19.9	57.0	50.9	37.0	15.5	29.5	38.6
	G1	51.7	58.0	32.1	61.2	54.9	29.3	19.0	42.9	46.8
	G2	52.1	58.7	32.5	61.8	54.6	31.2	19.3	42.9	46.9
Mdl/es	B1	50.6	65.2	50.4	77.6	72.1	63.8	59.5	47.8	54.2
	C1	12.5	42.9	0.4	45.1	44.4	11.1	0.9	0.0	0.7
	C2	11.9	51.7	1.3	51.2	41.8	30.5	2.4	1.6	7.0
	D1	49.2	64.9	50.5	68.1	80.1	57.4	70.3	58.2	56.2
	F1	58.6	68.1	63.2	66.9	76.7	51.9	69.4	55.6	57.8
	F2	56.8	70.5	47.6	63.4	68.5	44.5	46.5	48.7	44.9
	G1	72.4	76.8	59.3	78.4	79.6	55.7	64.3	68.1	72.4
	G2	62.1	70.6	59.5	68.1	65.7	53.9	66.5	60.9	65.4
Mdl/fr	C1	18.5	37.0	0.4	46.8	52.4	37.8	2.6	0.5	5.8
	C2	26.1	45.5	0.7	52.9	56.9	44.2	4.8	1.4	16.6
	D1	54.0	64.9	62.3	65.3	79.5	53.7	44.7	67.5	62.8
	E1	54.6	66.2	40.6	73.8	74.1	58.7	45.2	42.1	51.5
	F1	57.5	59.0	47.4	72.7	70.6	62.5	36.2	46.7	54.3
	F2	56.1	59.8	16.2	63.1	59.4	58.3	35.8	41.5	52.6
	G1	62.7	71.8	47.7	68.8	77.7	60.7	31.9	68.9	70.3
	G2	58.7	67.3	44.8	64.9	74.7	58.9	32.0	62.9	64.9

the different corpora (cf. fig. 2). From this analysis it is possible to derive that chemical entities ('chem') and living beings ('livb') can be identified at a better rate than the entities from the other groups. In the case of the patent corpus, the identification of the chemical entities can be reached at a rate which is high in comparison to the entities from the other semantic groups. Furthermore, it becomes clear that anatomical entities ('anat') and disease & disorder ('diso') can be well recognized in Medline abstracts and EMEA drug guidelines in contrast to patents. Overall, the presented results indicate that the identification of the concepts and entities can be achieved at a higher performance level in French and Spanish in contrast to German and Dutch.

Mention Evaluation (Task A). The evaluation of the mention annotations has been performed against a SSC that has been generated from the annotated corpora contributed by the MANTRA project partners and the participants f the CLEF-ER challenge. The SSC has been generated as described in section 3.3 and a TP is any mention annotation that nests a centroid in the SSC. This can be interpreted as the identification of a portion of the entity representation that has a high agreement between the different annotation solutions. Every annotated corpus has been evaluated against the appropriate SSC, i.e. the same corpus annotated in the same language. (cf. fig. 3)

The performance evaluation indicates that – with a few exceptions – the annotation of the EMEA documents can be achieved with better results than the annotation of the Medline abstracts, or the patent documents. This results is true for all languages except for Dutch. The mention annotation of the patent documents shows a mixed picture, since in general the performance for the annotation in German and French resembles the performance produced on the other

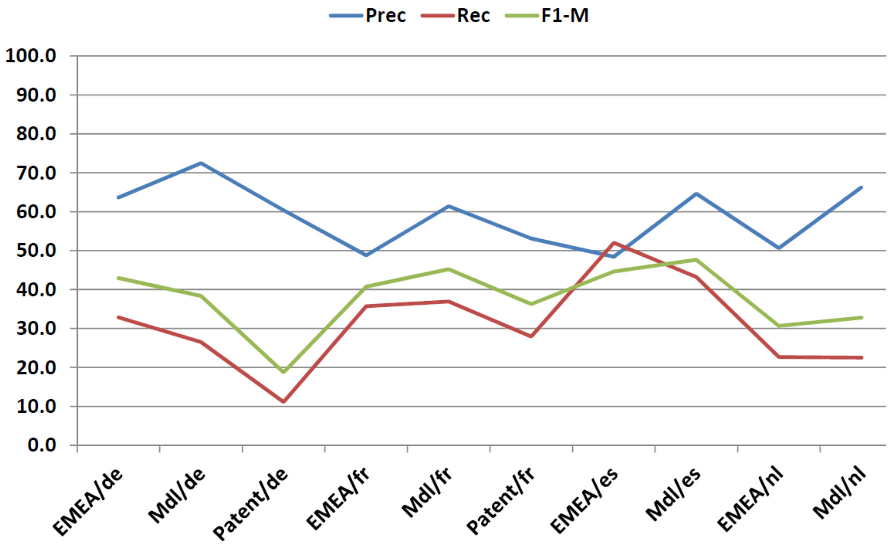


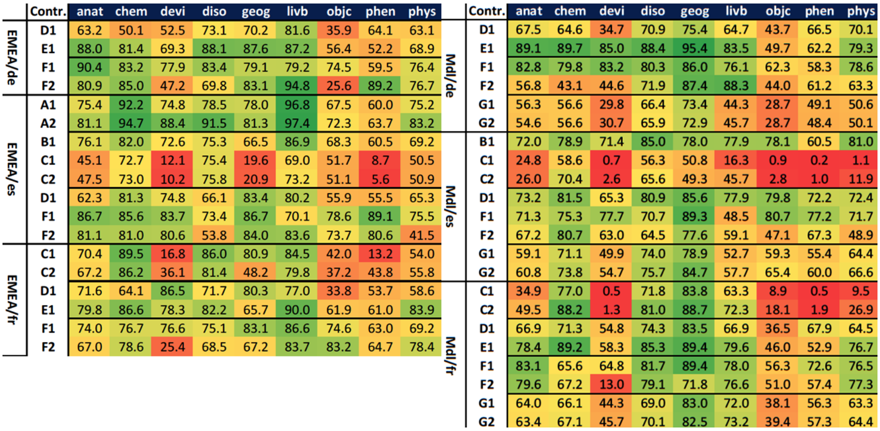
Fig. 3. (Evaluation of mentions): The figure shows the average of the F1-measure across all contributing systems for the mention annotation

two corpora, and comparing the different semantic groups it becomes clear that for selected groups the performance is good (e.g., phenotype – 'phen', 'anat', 'livb' and 'chem').

Again, table 7 shows the results for individual participants, but now for the mention annotation. The measured performances are similar to the results from the task B evaluation (cf. tbl. 6). On the other side, the performances on the German corpora has improved for the mention annotation in comparison to the CUI annotations.

CUI Assignment, Non-trival Cases (Task A). Finally we ignored all the trivial assignments of a CUI to the non-English documents, where a 'trivial' assignment is determined by the fact that the non-English term is already known in the terminological resources as a synonym to a given English term. This evaluation uses a smaller number of term candidates in the English SSC and focuses the evaluation towards those terms where new term candidates – in comparison to the original terminological resources – can be expected. The performances of the annotation solutions against this set of candidate terms (cf. fig. 4) shows a different picture than the previous analysis (cf. fig. 2). Now the performances of the annotation solutions in French and Spanish are now lower than previously and do not differ much from the annotation solutions in German. It is remarkable that the annotations for the different semantic groups are in a similar range, e.g. for nl, de and es on Medline and EMEA, and it becomes again visible that the

Table 7. (Evaluation A, F1-Measure, challenge's participants): Similar to the the previous table 6, this table shows the F1-measure performances of the individual solutions in the task A evaluation, i.e. annotation of entity mentions in the text



	Contr.	anat	chem	devi	diso	geog	livb	objc	phen	phys
EMEA/de	D1	63.2	50.1	52.5	73.1	70.2	81.6	35.9	64.1	63.1
	E1	88.0	81.4	69.3	88.1	87.6	87.2	56.4	52.2	68.9
	F1	90.4	83.2	77.9	83.4	79.1	79.2	74.5	59.5	76.4
	F2	80.9	85.0	47.2	69.8	83.1	94.8	25.6	89.2	76.7
EMEA/es	A1	75.4	92.2	74.8	78.5	78.0	96.8	67.5	60.0	75.2
	A2	81.1	94.7	88.4	91.5	81.3	97.4	72.3	63.7	83.2
	B1	76.1	82.0	72.6	75.3	66.5	86.9	68.3	60.5	69.2
	C1	45.1	72.7	12.1	75.4	19.6	69.0	51.7	8.7	50.5
	C2	47.5	73.0	10.2	75.8	20.9	73.2	51.1	5.6	50.9
	D1	62.3	81.3	74.8	66.1	83.4	80.2	55.9	55.5	65.3
	F1	86.7	85.6	83.7	73.4	86.7	70.1	78.6	89.1	75.5
	F2	81.1	81.0	80.6	53.8	84.0	83.6	73.7	80.6	41.5
EMEA/fr	C1	70.4	89.5	16.8	86.0	80.9	84.5	42.0	13.2	54.0
	C2	67.2	86.2	36.1	81.4	48.2	79.8	37.2	43.8	55.8
	D1	71.6	64.1	86.5	71.7	80.3	77.0	33.8	53.7	58.6
	E1	79.8	86.6	78.3	82.2	65.7	90.0	61.9	61.0	83.9
	F1	74.0	76.7	76.6	75.1	83.1	86.6	74.6	63.0	69.2
	F2	67.0	78.6	25.4	68.5	67.2	83.7	83.2	64.7	78.4

	Contr.	anat	chem	devi	diso	geog	livb	objc	phen	phys
Mdl/de	D1	67.5	64.6	34.7	70.9	75.4	64.7	43.7	66.5	70.1
	E1	89.1	89.7	85.0	88.4	95.4	83.5	49.7	62.2	79.3
	F1	82.8	79.8	83.2	80.3	86.0	76.1	62.3	58.3	78.6
	F2	56.8	43.1	44.6	71.9	87.4	88.3	44.0	61.2	63.3
	G1	56.3	56.6	29.8	66.4	73.4	44.3	28.7	49.1	50.6
	G2	54.6	56.6	30.7	65.9	72.9	45.7	28.7	48.4	50.1
Mdl/es	B1	72.0	78.9	71.4	85.0	78.0	77.9	78.1	60.5	81.0
	C1	24.8	58.6	0.7	56.3	50.8	16.3	0.9	0.2	1.1
	C2	26.0	70.4	2.6	65.6	49.3	45.7	2.8	1.0	11.9
	D1	73.2	81.5	65.3	80.9	85.6	77.9	79.8	72.2	72.4
	F1	71.3	75.3	77.7	70.7	89.3	48.5	80.7	77.2	71.7
	F2	67.2	80.7	63.0	64.5	77.6	59.1	47.1	67.3	48.9
	G1	59.1	71.1	49.9	74.0	78.9	52.7	59.3	55.4	64.4
	G2	60.8	73.8	54.7	75.7	84.7	57.7	65.4	60.0	66.6
Mdl/fr	C1	34.9	77.0	0.5	71.8	83.8	63.3	8.9	0.5	9.5
	C2	49.5	88.2	1.3	81.0	88.7	72.3	18.1	1.9	26.9
	D1	66.9	71.3	54.8	74.3	83.5	66.9	36.5	67.9	64.5
	E1	78.4	89.2	58.3	85.3	89.4	79.6	46.0	52.9	76.7
	F1	83.1	65.6	64.8	81.7	89.4	78.0	56.3	72.6	76.5
	F2	79.6	67.2	13.0	79.1	71.8	76.6	51.0	57.4	77.3
	G1	64.0	66.1	44.3	69.0	83.0	72.0	38.1	56.3	63.3
	G2	63.4	67.1	45.7	70.1	82.5	73.2	39.4	57.3	64.4

annotation of EMEA can be achieved at higher performance levels than the annotation of Medline.

5 Conclusions

The CLEF-ER challenge has targeted the task of entity recognition in multilingual and parallel documents. The approach is based on the development of an SSC, which would be made available in the English version for the participants of the challenge, and – later on – for the non-English corpora for any further evaluation of the participants' contributions. At the current state, only preliminary results are available indicating that the task requires the integration of different technologies to achieve ER in multilingual documents. Different approaches have been tested, but further investigation is required to state, which solutions perform best on the given task.

Nonetheless, it becomes clear that evaluation A ("monolingual mention evaluation") as well as evaluation B ("cross-lingual concept evaluation") gives us an indication of how well an individual contribution complies with the harmonized contribution where the harmonized contribution ("SSC") is composed of at least 3 contributions and their agreement induced by the e-centroid method.

On the other side, the analysis shows that the French corpora allow a higher agreement with the SSC than the German and the Spanish corpora. For the Dutch corpora, a high agreement has been achieved through the annotation solutions, but this is biased, since only a very small number of annotated corpora was available.

In the next phase, the contributions from the participants will be analysed for their individual performances on the challenge tasks. Furthermore, the

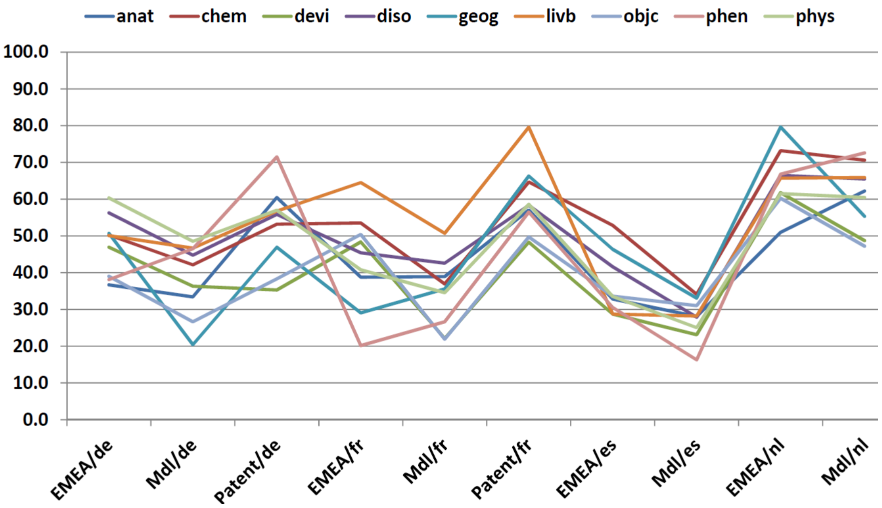


Fig. 4. (Evaluation B for semantic groups after term reduction): Similarly to the previous figure (cf. fig. 2), the average F1-measure of all contributing systems for each semantic group has been calculated, but in contrast to the previous figure the evaluation only considers a subset of all annotations. This subset is specific to novel findings of mentions that are linked to the mention in the parallel English document, but is not confirmed by a synonym in the terminological resource.

MANTRA project partners will mine the contributions for novel terms and will generated a gold standard corpus to evaluate the contributions of the participants on a smaller scale and against the opinion of an expert.

Acknowledgement. This work was funded by the European Commission STREP grant number 296410 (”Mantra”, FP7-ICT-2011-4.1).

References

1. Roberts, A., Gaizauskas, R., Hepple, M., Davis, N., Demetriou, G., Guo, Y., Kola, J.S., Roberts, I., Setzer, A., Tapuria, A., et al.: The CLEF corpus: semantic annotation of clinical text. In: AMIA Annual Symposium Proceedings, vol. 2007, p. 625. American Medical Informatics Association (2007)
2. Lussier, Y.A., Shagina, L., Friedman, C.: Automating icd-9-cm encoding using medical language processing: A feasibility study. In: Proceedings of the AMIA Symposium, p. 1072. American Medical Informatics Association (2000)
3. Catarci, T., Ferro, N., Forner, P., Hiemstra, D., Karlgren, J., Penas, A., Santucci, G., Womser-Hacker, C.: CLEF 2012: information access evaluation meets multilinguality, multimodality, and visual analytics. ACM SIGIR Forum 46, 29–33 (2012)
4. Roda, G., Tait, J., Piroi, F., Zenz, V.: CLEF-IP 2009: retrieval experiments in the Intellectual Property domain. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mandl, T., Mostefa, D., Peñas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 385–409. Springer, Heidelberg (2010)

5. Krallinger, M., Leitner, F., Rodriguez-Penagos, C., Valencia, A.: Overview of the protein-protein interaction annotation extraction task of BioCreative II. Genome Biology 9(suppl. 2), S4 (2008), http://genomebiology.com/2008/9/S2/S4
6. Morgan, A., Lu, Z., Wang, X., Cohen, A., Fluck, J., Ruch, P., Divoli, A., Fundel, K., Leaman, R., Hakenberg, J., Sun, C., Liu, H.H., Torres, R., Krauthammer, M., Lau, W., Liu, H., Hsu, C.N., Schuemie, M., Cohen, K.B., Hirschman, L.: Overview of BioCreative II gene normalization. Genome Biology 9(suppl. 2), S3 (2008), http://genomebiology.com/2008/9/S2/S3
7. Cohen, K.B., Demner-Fushman, D., Ananiadou, S., Pestian, J., Tsujii, J., Webber, B. (eds.): Proceedings of the BioNLP 2009 Workshop. Association for Computational Linguistics, Boulder (2009), http://www.aclweb.org/anthology/W09-13
8. Rebholz-Schuhmann, D., Yepes, A.J., Mulligen, E.M.V., 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, 163–179 (2010)
9. Rebholz-Schuhmann, D., Jimeno-Yepes, A., Li, C., Kafkas, S., Lewin, I., Kang, N., Corbett, P., Milward, D., Buyko, E., Beisswanger, E., Hornbostel, K., Kouznetsov, A., Witte, R., Laurila, J., Baker, C., Kuo, C.J., Clematide, S., Rinaldi, F., Farkas, R., Móra, G., Hara, K., Furlong, L., Rautschka, M., Lara Neves, M., Pascual-Montano, A., Wei, Q., Collier, N., Mahbub Chowdhury, M.F., Lavelli, A., Berlanga, R., Morante, R., Van Asch, V., Daelemans, W., Marina, J., van Mulligen, E., Kors, J., Hahn, U.: Assessment of NER solutions against the first and second CALBC Silver Standard Corpus. J. Biomedical Semantics 2(suppl. 5), S11 (2011)
10. Hersh, W., Voorhees, E.: TREC genomics special issue overview. Inf. Retr. Boston 12, 1–15 (2009)
11. Lu, Z.: PubMed and beyond: a survey of web tools for searching biomedical literature. Database (Oxford), 2011:baq036 (2011)
12. Rebholz-Schuhmann, D., Clematide, S., Rinaldi, F., Kafkas, S., van Mulligen, E.M., Bui, C., Hellrich, J., Lewin, I., Milward, D., Poprat, M., Jimeno-Yepes, A., Hahn, U., Kors, J.A.: Multilingual semantic resources and parallel corpora in the biomedical domain: the CLEF-ER challenge. In: Proceedings CLEF Conference, vol. 2013 (2013)
13. Bodenreider, O.: The Unified Medical Language System (UMLS): integrating biomedical terminology. Nucleic Acids Res. 32, D267–D270 (2004)
14. Brown, E.G., Wood, L., Wood, S.: The medical dictionary for regulatory activities (MedDRA). Drug Safety 20(2), 109–117 (1999)
15. Stearns, M.Q., Price, C., Spackman, K.A., Wang, A.Y.: SNOMED clinical terms: overview of the development process and project status. In: Proceedings of the AMIA Symposium, vol. 662, American Medical Informatics Association (2001)
16. Smith, B., Ashburner, M., Rosse, C., Bard, J., Bug, W., Ceusters, W., Goldberg, L.J., Eilbeck, K., Ireland, A., Mungall, C.J., Leontis, N., Rocca-Serra, P., Ruttenberg, A., Sansone, S.A., Scheuermann, R.H., Shah, N., Whetzel, P.L., Lewis, S.: The OBO Foundry: coordinated evolution of ontologies to support biomedical data integration. Nat. Biotechnol. 25, 1251–1255 (2007)
17. Lewin, I., Kafkas, S., Rebholz-Schuhmann, D.: Centroids: Gold standards with distributional variations. In: Proceedings of the Language Resources Evaluation Conference, Istanbul, Turkey (2012)
18. Lewin, I., Clematide, S.: Deriving the Mantra Silver Standard. In: Proceedings CLEF Conference, vol. 2013 (2013)