Overview of the CHEMDNER patents task

Martin Krallinger[1⋆], Obdulia Rabal[2], Analia Lourenço[3], Martin Perez Perez[3], Gael Perez Rodriguez[3], Miguel Vazquez[1], Florian Leitner[4] Julen Oyarzabal[2], and Alfonso Valencia[1]

[1]Structural Computational Biology Group,
Structural Biology and BioComputing Programme,
Spanish National Cancer Research Centre,
Madrid, Spain
[2]Small Molecule Discovery Platform, Center for Applied Medical Research (CIMA), Spain
[3]Departamento de Informatica, Universidade de Vigo, Spain
[4] Computational Intelligence Group, Universidad Politecnica de Madrid, Spain
{mkrallinger,
avalencia}@cnio.es
http://www.cnio.es

Abstract. A considerable effort has been made to extract biological and chemical entities, as well as their relationships, from the scientific literature, either manually through traditional literature curation or by using information extraction and text mining technologies. Medicinal chemistry patents contain a wealth of information, for instance to uncover potential biomarkers that might play a role in cancer treatment and prognosis. However, current biomedical annotation databases do not cover such information, partly due to limitations of publicly available biomedical patent mining software. As part of the *BioCreative V* CHEMDNER patents track, we present the results of the first named entity recognition (NER) assignment carried out to detect mentions of chemical compounds and genes/proteins in running patent text. More specifically, this task aimed to evaluate the performance of automatic name recognition strategies capable of isolating chemical names and gene and gene product mentions from surrounding text within patent titles and abstracts. A total of 22 unique teams submitted results for at least one of the three CHEMDNER subtasks. The first subtask, called the CEMP (chemical entity mention in patents) task, focused on the detection of chemical named entity mentions in patents, requesting teams to return the start and end indices corresponding to all the chemical entities found in a given record. A total of 21 teams submitted 93 runs, for this subtask. The top performing team reached an f-measure of 0.89 with a precision of 0.87 and a recall of 0.91. The CPD (chemical passage detection) task required the classification of patent titles and abstracts whether they do or do not contain chemical compound mentions. Nine teams returned predictions for this task (40 runs). The top run in terms of Matthew's correlation coefficient (MCC) had a score of 0.88, the highest sensitivity

⋆ Corresponding author

score was of 0.99, the best specificity was 0.94 and the best accuracy 0.95. The four participants (16 runs) of the GPRO (gene and protein related object) task had to detect gene/protein mentions that could be linked to at least one biological database, such as SwissProt or EntrezGene. For this task, the best f-measure was of 0.81, the best recall was of 0.85 and the best precision was of 0.82. For the CHEMDNER task 21,000 medicinal chemistry patent abstracts were manually annotated, resulting in the CEMP (chemical entity mention in patents) and GPRO (gene and protein related object) gold standard corpora. Each of these two corpora comprised subsets of 7,000 records: the training set and development sets (used for training and optimization of participating systems) and the test set (used for comparative evaluation purposes). Manual annotations and automated predictions could be examined through an online text annotation visualization system called *Markyt* (http://markyt.org).

Key words: CHEMDNER patents; ChemNLP; BioCreative; Named Entity Recognition; Chemical compounds; Genes/proteins; Text Mining

1 Introduction

Patents represent a very challenging type of document for automated text mining and information extraction approaches, showing considerable differences in terms of linguistic structure when compared to common English language. Moreover, patents cover a heterogeneous range of subject areas, which are only classified by rather coarse level international patent codes (IPC). There are differences in terms of the constraints and requirements imposed by the various patent authorities, which in turn is reflected in the resulting patent texts. Challenges and hurdles when preprocessing noisy patent texts are well known, being typographical errors, OCR errors and missing spaces only some of the difficulties encountered by automated text-processing pipelines. Patent literature is characterized by a sublanguage of not only technical terms (scientific and technical language) but also has many expressions related to legal aspects of the field of intellectual property. Although currently the main data consumers of patent processing software are legal experts, information professionals and patent analysts, there is an increasing interest of the scientific community to access more efficiently the information contained in patents, especially in the medicinal chemistry domain. Pharma and biotech companies explore patent information intensely for competitive intelligence and drug development purposes [5]. The effort to extract manually, or with the aid of text mining systems, information contained in the scientific literature in order to populate biological annotation databases is considerable, but those databases do almost completely lack the integration of relevant characterizations of biomedical and chemical entities described in patents. Manual curation of patents is especially work intensive because patents are on average up to five times longer when compared to scientific full text articles [5]. Moreover, there are estimations that about 15-20 million patents are related to life sciences and medicinal chemistry, and therefore to detect and characterize

biomedical entities within this large textual data volume is an interesting task for text mining and information extraction technologies. Automatic recognition of entities such as chemical compounds and genes in patents is key to enable better patent retrieval engines as well as to assist database curation of patents. Despite the valuable characterizations of biomedical relevant entities such as chemical compounds, genes and proteins contained in patents, academic research in the area of text mining and information extraction using patent data has been minimal. Pharmaceutical patents covering chemical compounds provide information on their therapeutic applications and, in most cases, on their primary biological targets. Some initial attempts to deal with patent data have been carried out in the context of an information retrieval track to examine search engines applied to chemistry in patents, the so-called TREC-Chem task [7]. On the other hand, efforts have been made to annotate full-text patents with chemical compound and protein target mentions, using automatic pre-annotation together with manual correction of errors and manually adding missing entity mentions [8]. Such approaches constitute a valuable alternative to exhaustive human generated corpora, which are difficult to construct for lengthy full text patents. Nevertheless for competitive assessment purposes it is unclear whether using a systematic pre-annotation step could potentially favor or influence the performance of some of the participating teams/methodologies. Some of the general difficulties for the automatic chemical name recognition in the scientific literature have been already highlighted in the previous BioCreative IV CHEMDNER task [1, 3, 2]. Nevertheless, so far it was less clear how well chemical or gene/protein NER worked when dealing with patent language.

This CHEMDNER patents task thus addressed the automatic extraction of chemical and biological data from medicinal chemistry patents. The aim of this task was to engage in more depth the biomedical text mining community to process noisy text data (patents) and to promote the development of software that helps to derive chemical and biological annotations from patents.

2 Task description

The CHEMDNER task was divided into three tasks, which were carried out on the same patent abstract collections. Teams were provided with a training and development set to construct their predictor and a blinded test set for which they have to submit predictions that were evaluated against manual annotations.

The used patent abstract records were released in the form of plain-text, UTF8-encoded patent abstracts in a tab-separated format with the following three columns: (1) patent identifier, (2) title of the patent, (3) abstract of the patent.

The BioCreative V CHEMDNER patents Track was structured into three subtasks:

* CEMP (chemical entity mention in patents, main task): the detection of chemical named entity mentions in patents.

* CPD (chemical passage detection, text classification task): the detection of patent titles and abstracts that mention chemical compounds.
* GPRO (gene and protein related object task): for the GPRO task teams had to identify mentions of gene and protein related objects (named as GPROs).

The settings of the *CEMP task* were very similar to the CEM task of BioCreative IV [3]. Participating teams had to detect correctly the start and end indices corresponding to all the chemical entities. Chemical entities were manual annotated by domain experts using well-defined annotation guidelines (CEMP annotation guidelines). Those guidelines were similar to the ones used for the CHEMDNER task at BioCreative IV but they also showed several differences updates to make them more suitable for the annotation of patent data. The CEMP annotation guidelines (as well as the GPRO guidelines) were published together with the manually annotated corpora in order for teams to actually understand how the annotations were done and to make it possible to examine how their systems could consider the annotation rules.

For the *CPD task* we asked participating teams to classify patent titles and abstracts whether they do or do not contain mentions of chemical entities. It was thus essentially a text classification task, and represented a sort of preprocessing step to be able to determine in the first place if the patent text did or did not contain chemical mentions. The classification generated by participating teams was compared to the annotations generated manually by chemical domain experts (derived from an exhaustive manual tagging of chemical entities done for the CEMP task). For the CPD task teams returned a binary classification of patent titles and abstracts into: 1 does mention chemicals or 0 does not mention chemicals. For each of the predictions they also provided a rank and a confidence score.

For the *GPRO task* teams had to identify mentions of gene and protein related objects (named as GPROs) mentioned in patent titles and abstracts. The definition of GPRO entity mentions was primarily concerned with capturing those types of mentions that are of practical relevance (both for end users of the extracted data as well as for the named entity recognition systems). Therefore, the covered GPRO entities had to be annotated at a sufficient level of granularity to be able to determine whether the labeled mention can or can not be linked to a specific gene or gene product (represented by an entry of a biological annotation database). The annotation carried out for the CHEMDNER GPRO task was exhaustive for the types of GPRO mentions that were previously specified. We distinguished two types of GPRO entity mentions:

* GPRO entity mention type 1: covering those GPRO mentions that can be normalized to a bio-entity database record.
* GPRO entity mention type 2: covering those GPRO mentions that in principle cannot be normalized to a unique bio-entity database record (e.g. protein families or domains).

For the GPRO task we only requested the prediction of the GPRO type 1 mention offsets following a similar stetting as done for the CEMP task. Given a

set of patent abstracts, the participants have to return the start and end indices corresponding to all the GPRO type 1 entities mentioned in this document. We did not request that teams had to return the actual database identifiers of normalizable GPRO mentions.

The system predictions were evaluated by comparing them to the manually labeled annotations done by domain experts according to the annotation guidelines. The CHEMDNER patents task was structured into several steps. At the beginning a small sample set with annotations and example predictions was released in order to illustrate the kind of annotations that would be provided for this task. Later, the training and development sets were made public. Finally, during the test phase a collection of patent abstracts without annotations (blinded) was distributed and teams were requested to generate automatic annotations (according to predefined evaluation formats) and return them to the task organisers after a short period of time. Teams could submit for each of the subtasks up to five predictions (runs).

Figure 1 shows example team predictions for the CEMP and GPRO subtasks for an patent abstract.

WO2009026621A1	A:12:24	1	0.99	paliperidone
WO2011115938A1	T:0:17	1	0.99	Spiro-tetracyclic
WO2011115687A2	A:0:12	1	0.99	SP-B
WO2011115687A2	T:0:22	2	0.98989	Alkylated
WO2011115687A2	A:104:117	3	0.98978	SP-B
US20050101595	A:0:13	1	0.99	Aminothiazole
US20050101595	A:60:67	2	0.98989	2-amino
US20050101595	T:0:50	3	0.98978	N-containing
US20050101595	A:29:52	4	0.98967	N-containing

A

CN103371975A	A:271:274	1	0.99	RGD
CN103371975A	A:276:306	2	0.98989	Arginine-Glycine-Aspartic
US20090312385	A:100:112	1	0.99	CB2
US20090312385	T:0:11	2	0.98989	Cannabinoid
WO2014144455A1	A:616:621	1	0.99	CARM1
WO2014144455A1	T:53:58	2	0.98989	carm1
WO2014144455A1	A:676:681	3	0.98978	CARM1
EP1087981B1	T:0:18	1	0.99	Prenyl
EP1087981B1	A:60:79	2	0.98989	prenyl

B

Fig. 1. Example output predictions for the CEMP (A) and GPRO (B) subtasks. The first column corresponded to the patent identifier and the second column to the predictions, i.e. the mention offsets. The offset string consisted in a triplet joined by the ':' character, containing the text type (T: Title, A:Abstract), the start offset and the end offset. The third column corresponded to the actual rank for each prediction given an article and the last column to the corresponding confidence score.

Together with the CHEMDNER patents corpus we provided an evaluation script that calculated the performance of predictions against the Gold Standard data and facilitated checking of the correct submission format. Additionally, teams could use a specially adapted online annotation visualisation system called Markyt. This system not only evaluated in depth the predictions against the manual annotations, but also showed side by side differences between the automatic and manual annotations for a given patent record.

The evaluation metrics used to asses team predictions were micro-averaged recall, precision and F-score (main evaluation metric) for the CEMP and GPRO tasks. Three different result types were scored: False negative (FN) results corre-

sponding to incorrect negative predictions (type II errors); False positives (FP) predictions corresponding to incorrect positive predictions (type I errors) and True positives (TP) results corresponding to correct predictions.

Recall r (also known as coverage, sensitivity, true positive rate, or hit rate) is the percentage of correctly labeled positive results over all positive cases.

$$r := \frac{TP}{TP + FN} \tag{1}$$

Precision p (positive predictive value) is the percentage of correctly labeled positive results over all positive labeled results.

$$p := \frac{TP}{TP + FP} \tag{2}$$

The F-measure F_β is the harmonic mean between precision and recall, where β is a parameter for the relative importance of precision over recall.

$$F_\beta := (1 + \beta^2)\frac{p \cdot r}{\beta^2 p + r} \tag{3}$$

The balanced F-measure ($\beta = 1$, referred to as "F-score" in this work) can be simplified to:

$$F_1 = 2\frac{p \cdot r}{p + r} \tag{4}$$

3 The CHEMDNER patents corpus

A crucial aspect for the preparation of text mining training and evaluation datasets is to carry out a proper selection of the used document, in order to be representative of the kind of data that should be extracted, in this case of medicinal chemistry related patents. In order to make sure that after the competition participating teams would have access to the used dataset and to new as well as older patents of the same kind, we have used the online accessible patent collections provided by Google patents.

The following steps were followed to select abstracts for annotation.

Step 1 : IPC code selection criteria. We have selected all the patents that had at least one assigned IPC (International Patent Classification) code corresponding to A61P (or its corresponding child IPCs) and also at least one A61K31 IPC code. This selection criteria ensured that the corresponding patents are enriched in medicinal chemistry patents mentioning chemical entities.

Step 2: Publication date selection criteria. We decided to select only more recent patents, having a publication date from 2005 to 2014.

Step 3: Encoding and HTML tags. HTML tags were converted to their corresponding characters. For the patents text we have specified UTF-8 encoding always.

Step 4: Publication type. We did not distinguish during the patent selection between 'Application' and 'Grant'.

Step 5: OCR error checking. We have checked that neither titles nor abstracts have any OCR error related metadata.

Step 6: Patent agency. We selected patents from the following agencies: the World Intellectual Property Organization (WIPO), the European Patent Office (EPO), the United States Patent and Trademark Office (USPTO), Canadian Intellectual Property Office (CIPO), the German Patent and Trade Mark Office (DPMA) and the State Intellectual Property Office of the People's Republic of China (SIPO).

Step 7: Automatic/machine translation. Some of the patent titles and abstracts that were available in English had been generated through automatic translation from texts in other languages, mainly Chinese and German. We decided therefore only to use those patents where the English patent titles and abstracts were not generated through machine translation.

Step 8: Language detection and filtering of patents containing non-English text. Manual inspection of a sample of patents revealed that there were cases of patents that contained texts in various languages. We applied a simple language detection program based on the fraction of English stop words versus stop words from other languages (French, German) and a used a strict cut-off to filter out abstracts containing text in non-English from the set of EPO and Canadian patents.

Step 9: Length selection criteria. A length cutoff was chosen to avoid patents that are too short. The used cut off was derived from the previous CHEMDNER PubMed abstracts and titles. The minimal length in number of tokens (simple whitespace tokenization) of CHEMDNER PubMed documents was used as a cutoff, corresponding to 12 word tokens.

Step 10: Duplicate abstracts. There was also a redundancy in terms of abstracts, i.e. some abstracts of patents were the same because some patents are published by different agencies. To address those cases only one of each duplicate was selected.

Step 11: Patent-patent relations. In order to avoid redundancy, we tried to retain among all the different publication numbers found under the field 'also published as' only one patent publication number. Specifically, the following order of preference was applied to select a single publication number for related documents: WO (which corresponds to PCT applications) > US (USA applications) > EP (European Applications) > DE (German Applications) > CN (Chinese) > CA (Canadian) .

Step 12: Traditional Chinese Medicine. A considerable number of Chinese patents relate to traditional Chinese medicine. A manual inspection of such patents showed that many of them did not describe any organic compound and provided instructions related to several ingredients, mainly herbs. We constructed a list of stop word frequently occurring in traditional Chinese medicine patents to filter out a number of those patents.

After applying all the previous selection and processing criteria we randomised the remaining patents and divided into the following subsets:

1- Sample set : 200 records
2- Training set : 7,000 records
3- Development set : 7,000 records
4- Test set : 7,000 records
5- Additional background and post-workshop set : 33,000 records

	Training set	Development set	Test set	Entire corpus
Patent abstracts	7,000	7,000	7,000	21,000
Nr. CEMP mentions	33,543	32,142	33,949	99,634
Nr. CEMP names	11,977	11,386	11,433	34,796
Passage with CEMP	9,152	8,937	9,270	27,359
Nr. GPRO mentions	6,876	6,263	7,093	20,232
Nr. GPRO type I mentions	4,396	3,934	4,093	12,423

Table 1. CHEMDNER patents corpus overview. This table provides and overview of the CHEMDNER corpus. Passage with CEMP refers to the number or titles or abstracts with chemical mentions (used for the CPD task).

The annotation process/settings for the annotation of chemical mentions in patents was essentially similar to the one used for the BioCreative IV CHEMDNER corpus. We relied on the same text annotation platform and used the same chemical entity mention classes: SYSTEMATIC, IDENTIFIERS, FORMULA, TRIVIAL, ABBREVIATION, FAMILY and MULTIPLE chemical mention types. We also introduced additional annotation rules and example cases with the aim of reaching a better distinction between the various class types during the manual annotation process. The annotation guidelines were adapted to mine patents with a stronger focus on identifying any wide definition of chemical terms rather than in obtaining highly refined chemical mentions that can be translated into a chemical structure. The common characteristic among all the chemical mention types used for the CHEMDNER-patent task was that i) they could be associated to chemical structure information to at least a certain degree of reliability or ii) they could be associated to general chemical structural information according to the terms commonly found in patents, specially those describing the characteristics of the substituents of the Markush formula (e.g., 'heteroaryl and aromatic bicycles' that describe topological classes). This implied that, compared to the previous CHEMDNER task (on scientific articles), an extended range of general chemical concepts (e.g., describing topologic features) were included.

The definition of GPRO entity mentions that were annotated for the CHEMDNER patents task was primarily concerned with capturing those types of mentions that are of practical relevance (both for end users of the extracted data as well as for the named entity recognition systems). Therefore, the covered GPRO entities had to be annotated at a sufficient level of granularity to be able to determine whether the labeled mention can or can not be linked to a specific gene

or gene product (represented by an entry of a biological annotation database). In case of the GPRO corpus, the human annotators had to mark up manually through a customized web-interface the mentions of GPROs (and related entities) in text. The selected GPRO entity mentions were classified by hand into one of the following GPRO entity mention classes:

NO CLASS : Names of individual protein domains and names of sequence or structural motifs. Here also DNA/RNA structural motifs should be labeled. Identifiers of protein domains (PFAM).

NESTED MENTIONS : Nested mentions of a single entity. This would correspond to nested mentions where the actual decomposed entity mentions could be normalized to one unique entity.

IDENTIFIER : Database identifiers of genes or gene products (proteins, RNA).

SEQUENCE : Mentions of protein (amino acid) sequences, nucleotide sequences, mutation and residue mentions of DNA, RNA and proteins.

FULL NAME : Full name of a GPRO, including names of precursor proteins (in case of cleaved proteins). It also includes multi-word terms referring to specific gene/protein named entities. It covers single word GPRO what do not correspond to abbreviations or symbols.

ABBREVIATION : Abbreviation of full name GPROs, GPRO acronyms, gene/protein symbols or symbolic names.

FAMILY : GPRO families that can be associated to some gene/protein family (or group of GPROs). It includes groups of genes/proteins at the sequence or taxonomic level. It comprises plural mentions of proteins assuming that they potentially refer to various different GPRO entities.

MULTIPLE : Mentions that do correspond to GPROs that are not described in a continuous string of characters.

The annotation carried out for the CHEMDNER GPRO task was exhaustive for the types of GPRO mentions that were previously specified. This implies that mentions of other entities, such as chemicals or substances, should not be labeled as GPROs.

We distinguish two types of GPRO entity mention types:

(1) GPRO entity mention type 1 : covering those GPRO mentions that can be normalized to a bio-entity database record. GPRO type 1 includes the following classes: NESTED MENTIONS, IDENTIFIER, FULL NAME and ABBREVIATION

(2) GPRO entity mention type 2 : covering those GPRO mentions that in principle cannot be normalized to a unique bio-entity database record. GPRO type 2 includes the following classes: NO CLASS, SEQUENCE, FAMILY and MULTIPLE.

For evaluation purposes only GPRO entity mentions of type 1 were considered.

4 Results

A total of 22 unique teams submitted results for at least one of the three CHEMDNER subtasks. Team results were submitted through the Markyt system and evaluated against the manual test set annotations. In order to avoid manual corrections of the test set results, together with the 7,000 test set patent abstracts we also released 33,000 additional patent abstracts for which teams also had to generate predictions.

In the case of the CEMP task a total of 21 teams submitted 93 different runs. The top performing team of the CEMP task reached an f-measure of 0.89 with a precision of 0.87 and a recall of 0.91. The top scoring run in terms of f-score was generated by team 274 (run 1). Table 2 shows the results obtained for all the submitted runs evaluated for the CEMP task. Two of the top performing teams, namely team 274 and team 288 could reach an f-score of over 0.88. The highest precision was obtained by run 2 of team 274 (0.89711) while the highest recall was obtained by the same team for run 5 (0.97617).

The CPD (chemical passage detection) task required the classification of patent titles and abstracts whether they do or do not contain chemical compound mentions. Nine teams returned predictions for this task (40 runs). The top run in terms of Matthew's correlation coefficient (MCC) had a score of 0.88 (team 288, run 1) with also the highest sensitivity score of 0.99, and the best accuracy (0.95). The best specificity was of 0.94, obtained by run 4 of team 313. Table 3 shows the results obtained for all the submitted runs evaluated for the CPD task.

The four participants (16 runs) of the GPRO (gene and protein related object) task had to detect gene/protein mentions that could be linked to at least one biological database, such as SwissProt or EntrezGene. For this task the best f-measure was of 0.81, the best recall was 0.85 and the best precision was 0.82, all of them obtained by team 274. Table 4 shows the results obtained for all the submitted runs evaluated for the GPRO task.

5 Discussion

Overall, the results of the chemical entity mention task were better when compared to the GPRO task. Although the exact reason still remains unclear. A proper inter-annotator agreement (IAA) study based on a blind annotation of 200 patent abstracts in case of the chemical entity mentions and of 500 patent abstracts for GPRO mentions should provide further insights. In addition to the IAA study we plan to determine the baseline predictions using vocabulary transfer. When looking at the results of the top scoring team of the CHEMDNER task of BioCreative IV on PubMed abstracts (F-score of 87.39%) it seems that patent abstracts are not particularly more challenging than scientific literature abstracts. Moreover, the obtained results are also competitive enough to derive in tools that not only could assist manual curation, but also could be used to automatic annotation extraction and patent abstract chemical indexing. Additional aspects that will be analysed by the task organisers include the influence

Team-Id	Run	Precision	Recall	F-score
274	1	0.8752	0.9129	0.8937
274	3	0.8908	0.8918	0.8913
274	2	0.8971	0.8822	0.8896
288	2	0.8718	0.9078	0.8894
288	3	0.8744	0.9047	0.8893
288	1	0.8695	0.9050	0.8869
288	5	0.8756	0.8964	0.8859
288	4	0.8627	0.8938	0.8779
362	1	0.8689	0.8869	0.8778
356	1	0.8553	0.8886	0.8717
293	4	0.8785	0.8623	0.8703
293	3	0.8782	0.8620	0.8700
293	2	0.8779	0.8616	0.8697
356	5	0.8607	0.8785	0.8695
356	2	0.8607	0.8785	0.8695
293	1	0.8778	0.8613	0.8694
276	5	0.8683	0.8681	0.8682
276	4	0.8492	0.8825	0.8655
276	3	0.8588	0.8680	0.8634
356	4	0.8539	0.8718	0.8627
356	3	0.8514	0.8702	0.8607
277	1	0.8788	0.8428	0.8604
274	4	0.7967	0.9314	0.8588
359	1	0.8767	0.8414	0.8587
277	4	0.8807	0.8320	0.8557
277	3	0.8792	0.8327	0.8553
277	2	0.8802	0.8306	0.8547
293	5	0.8615	0.8478	0.8546
359	3	0.8347	0.8743	0.8540
350	1	0.8703	0.8381	0.8539
350	2	0.8644	0.8425	0.8533
350	3	0.8700	0.8362	0.8528
277	5	0.8733	0.8330	0.8527
304	5	0.8293	0.8766	0.8523
304	2	0.8290	0.8768	0.8522
304	4	0.8442	0.8587	0.8514
313	2	0.8561	0.8373	0.8466
286	1	0.8751	0.8192	0.8462
313	5	0.8580	0.8342	0.8460
284	2	0.8859	0.8050	0.8435
304	3	0.8501	0.8359	0.8429
304	1	0.8315	0.8539	0.8425
286	4	0.8523	0.8318	0.8419
296	1	0.8630	0.8215	0.8417
296	5	0.8627	0.8202	0.8409
296	2	0.8632	0.8195	0.8408
296	3	0.8621	0.8199	0.8404
296	4	0.8630	0.8185	0.8402
359	2	0.8547	0.8232	0.8387
276	1	0.7776	0.9084	0.8379
276	2	0.7754	0.9093	0.8370
286	3	0.8297	0.8278	0.8288
315	2	0.8431	0.8130	0.8277
315	3	0.8479	0.8064	0.8266
284	5	0.8663	0.7890	0.8258
284	4	0.8660	0.7886	0.8255
284	1	0.8657	0.7882	0.8251
284	3	0.8655	0.7872	0.8245
359	4	0.7823	0.8701	0.8239
315	1	0.8506	0.7939	0.8213
286	5	0.8175	0.8234	0.8205
278	5	0.8288	0.7971	0.8127
278	2	0.8288	0.7970	0.8126
278	4	0.8200	0.7972	0.8084
278	3	0.8199	0.7971	0.8083
278	1	0.8203	0.7829	0.8011
286	2	0.8052	0.7783	0.7915
308	2	0.8094	0.7576	0.7827
308	1	0.7725	0.7573	0.7648
313	1	0.7940	0.7185	0.7544
348	1	0.8191	0.6971	0.7532
348	2	0.8199	0.6950	0.7523
348	4	0.8182	0.6930	0.7504
348	3	0.8105	0.6908	0.7459
348	5	0.8307	0.6757	0.7452
281	2	0.8312	0.6451	0.7264
281	5	0.8064	0.6143	0.6974
281	4	0.7834	0.6087	0.6851
281	3	0.7716	0.6062	0.6790
274	5	0.5202	0.9762	0.6787
313	3	0.4842	0.9233	0.6352
313	4	0.8698	0.4294	0.5750
281	1	0.6939	0.4739	0.5632
292	5	0.0044	0.0002	0.0004
292	3	0.0037	0.0002	0.0003
292	4	0.0026	0.0001	0.0002
292	1	0.0025	0.0001	0.0002
292	2	0.0025	0.0001	0.0002
337	4	0.0000	0.0000	0.0000
337	2	0.0000	0.0000	0.0000
337	5	0.0000	0.0000	0.0000
337	3	0.0000	0.0000	0.0000
337	1	0.0000	0.0000	0.0000

Table 2. Chemical entity mention in patents (CEMP) result overview.

of the CEMP and GPRO mention classes as well as the underlying patent agency to which the abstracts belong.

Acknowledgments. eTOX Grant Agreement n115002

References

1. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A. (2013, October). Overview of the chemical compound and drug name recognition (CHEMDNER) task. In BioCreative Challenge Evaluation Workshop (Vol. 2, p. 2).

Team-Id	Run	TP	FP	FN	TN	Sens.	Spec.	Accur.	MCC	P_full_R	AUC_PR
288	1	9140	605	130	4125	0.9860	0.8721	0.9475	0.8824	0.6657	0.9347
288	3	9114	602	156	4128	0.9832	0.8727	0.9459	0.8785	0.6627	0.9351
276	1	9087	625	183	4105	0.9803	0.8679	0.9423	0.8703	0.6622	0.9338
276	3	8952	512	318	4218	0.9657	0.8918	0.9407	0.8666	0.6622	0.9412
276	2	9100	668	170	4062	0.9817	0.8588	0.9401	0.8656	0.6622	0.9304
276	4	8992	572	278	4158	0.9700	0.8791	0.9393	0.8632	0.6621	0.9352
288	2	9101	689	169	4041	0.9818	0.8543	0.9387	0.8624	0.6627	0.9275
304	4	8923	529	347	4201	0.9626	0.8882	0.9374	0.8592	0.6622	0.9527
276	5	8911	536	359	4194	0.9613	0.8867	0.9361	0.8561	0.6621	0.9363
356	5	9117	751	153	3979	0.9835	0.8412	0.9354	0.8552	0.6621	0.9233
356	4	9117	751	153	3979	0.9835	0.8412	0.9354	0.8552	0.6621	0.9233
356	2	9117	751	153	3979	0.9835	0.8412	0.9354	0.8552	0.6621	0.9233
356	3	9136	774	134	3956	0.9855	0.8364	0.9351	0.8549	0.6621	0.9220
356	1	9138	784	132	3946	0.9858	0.8343	0.9346	0.8536	0.6621	0.9212
304	5	9066	711	204	4019	0.9780	0.8497	0.9346	0.8529	0.6640	0.9466
304	3	8796	450	474	4280	0.9489	0.9049	0.9340	0.8527	0.6622	0.9552
304	2	9066	713	204	4017	0.9780	0.8493	0.9345	0.8526	0.6640	0.9464
304	1	8967	636	303	4094	0.9673	0.8655	0.9329	0.8487	0.6641	0.9491
286	1	8642	500	628	4230	0.9323	0.8943	0.9194	0.8213	0.6621	0.9643
286	4	8804	672	466	4058	0.9497	0.8579	0.9187	0.8168	0.6621	0.9549
286	3	8908	790	362	3940	0.9610	0.8330	0.9177	0.8139	0.6622	0.9479
308	2	8582	549	688	4181	0.9258	0.8839	0.9116	0.8041	0.6625	0.9527
313	2	8727	686	543	4044	0.9414	0.8550	0.9122	0.8025	0.6629	0.9500
286	5	8935	894	335	3836	0.9639	0.8110	0.9122	0.8013	0.6622	0.9448
313	5	8694	681	576	4049	0.9379	0.8560	0.9102	0.7983	0.6629	0.9497
278	2	8552	589	718	4141	0.9226	0.8755	0.9066	0.7929	0.6621	0.9434
278	5	8552	589	718	4141	0.9226	0.8755	0.9066	0.7929	0.6621	0.9434
286	2	8794	832	476	3898	0.9487	0.8241	0.9066	0.7886	0.6621	0.9441
278	3	8572	668	698	4062	0.9247	0.8588	0.9024	0.7823	0.6621	0.9394
278	4	8572	668	698	4062	0.9247	0.8588	0.9024	0.7823	0.6621	0.9394
308	1	8691	774	579	3956	0.9375	0.8364	0.9034	0.7822	0.6625	0.9428
278	1	8482	649	788	4081	0.9150	0.8628	0.8974	0.7724	0.6621	0.9377
313	1	7629	564	1641	4166	0.8230	0.8808	0.8425	0.6756	0.6623	0.9329
313	3	9073	1919	197	2811	0.9788	0.5943	0.8489	0.6599	0.6631	0.8921
292	5	7925	939	1345	3791	0.8549	0.8015	0.8369	0.6442	0.6624	0.9489
292	1	7805	894	1465	3836	0.8420	0.8110	0.8315	0.6367	0.6622	0.9371
292	4	7768	874	1502	3856	0.8380	0.8152	0.8303	0.6356	0.6622	0.9358
292	3	7421	1180	1849	3550	0.8005	0.7505	0.7836	0.5355	0.6623	0.9135
292	2	7777	1433	1493	3297	0.8389	0.6970	0.7910	0.5344	0.6622	0.8624
313	4	3963	271	5307	4459	0.4275	0.9427	0.6016	0.3812	0.6621	0.8450

Table 3. Chemical passage detection (CPD) result overview.

Team-Id	Run	Precision	Recall	F-score
274	5	0.8143	0.8131	0.8137
274	4	0.7677	0.8502	0.8069
274	1	0.7835	0.8302	0.8062
274	2	0.8224	0.7852	0.8034
274	3	0.8059	0.7982	0.8020
304	5	0.7853	0.7220	0.7523
304	4	0.7868	0.7203	0.7520
304	2	0.7291	0.7642	0.7463
304	3	0.7880	0.6357	0.7037
304	1	0.7199	0.6819	0.7004
368	1	0.6526	0.6186	0.6351
286	1	0.4113	0.4097	0.4105
286	2	0.4067	0.3882	0.3973
286	5	0.0948	0.0191	0.0317
286	3	0.0850	0.0154	0.0261
286	4	0.0757	0.0120	0.0207

Table 4. Gene and protein related object (GPRO) result overview.

2. Krallinger, M., Rabal, et al. Segura-Bedmar, I. (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics, 7(Suppl 1), S2.

3. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A. (2015). CHEMDNER: The drugs and chemical names extraction challenge. J Cheminform, 7(Suppl 1), S1.
4. Krallinger, M., et al., (2015). The CHEMDNER corpus of chemicals and drugs and its annotation principles. Journal of cheminformatics, 7(Suppl 1), S2.
5. Aras, H., Hackl-Sommer, R., Schwantner, M., Sofean, M. (2014). Applications and Challenges of Text Mining with Patents.
6. Pasche, E., Gobeill, J., Kreim, O., Oezdemir-Zaech, F., Vachon, T., Lovis, C., Ruch, P. (2014). Development and tuning of an original search engine for patent libraries in medicinal chemistry. BMC bioinformatics, 15(Suppl 1), S15.
7. Lupu, M., Huang, J., Zhu, J., Tait, J. (2009, December). TREC-CHEM: large scale chemical information retrieval evaluation at TREC. In ACM SIGIR Forum (Vol. 43, No. 2, pp. 63-70). ACM.
8. Akhondi, S. A., Klenner, A. G., et al. (2014). Annotated chemical patent corpus: A gold standard for text mining.