The chemical corpus of the NLM-Chem BioCreative VII track

Full-text Chemical Identification and Indexing in PubMed articles

Rezarta Islamaj*, Robert Leaman*, David Cissel, Meng Cheng, Cathleen Coss, Joseph Denicola, Carol Fisher, Rob Guzman, Preeti Gokal Kochar, Nicholas Miliaras, Zoe Punske, Keiko Sekiya, Dorothy Trinh, Deborah Whitman, Susan Schmidt and Zhiyong Lu

National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

***Abstract*— The automatic recognition of chemical names and their corresponding database identifiers in biomedical text is an important first step for many downstream text-mining applications. The NLM-Chem track at BioCreative VII aimed to foster the development of algorithms that can predict with high quality the chemical entities in biomedical literature and further identify the chemical substances that are candidates for article indexing. The NLM-Chem track corpus is a manually curated corpus comprehensively annotated with chemical entities and indexed with chemical substances. NLM-Chem BioCreative VII corpus consists of three parts: A high-quality manually annotated corpus of 200 full-text PubMed central articles, the collection of 11,500 PubMed documents previously annotated in the ChemDNER and BC5CDR challenges, which we have enriched with their corresponding chemical substance indexing, and the collection of 1,387 recently published PMC articles, equipped with chemical substance indexing by manual experts at the National Library of Medicine. This document details the characteristics of this novel resource for chemical entity recognition. Using this new resource, we have demonstrated improvements in the chemical entity recognition algorithms.**

Keywords—corpus annotation; inter-annotator agreement; chemical entity recognition, text mining

I. INTRODUCTION

Chemical entities appear throughout the biomedical research literature, in studies from chemistry, to various other disciplines such as medicine, biology, and pharmacology. As such, chemical names are one of the most searched entity types in PubMed (1). Therefore, correctly identifying chemical names has a significant impact on chemical information retrieval: helping scientists retrieve the relevant literature, directly impacting research that relies on a correct understanding of the structure of chemicals, their usage, and interactions with other molecular entities. For example, correct identification of chemicals and their properties directly impacts drug development research (2).

However, chemicals in the biomedical literature often do not appear to conform to the chemical naming rules defined by standardization bodies. Chemicals appear in numerous lexical variations, synonymous names, and abbreviated forms, which are often ambiguous (3). Moreover, these variations and difficulties are often compounded in articles' full-text, compared with the title and abstract, causing a substantial performance reduction in automated chemical named entity recognition (NER) systems trained using only titles and abstracts (4). However, the full-text frequently contains more detailed chemical information, such as the properties of chemical compounds, their biological effects, and interactions with diseases, genes, and other chemicals (5-7).

Developing a chemical entity recognition system that accurately addresses these challenges requires a manually-annotated corpus of chemical entities, with sufficient examples in full-text articles for system training and an accurate evaluation of their performance.

The NLM-Chem track at BioCreative VII consisted of two tasks (8):

- Chemical Identification in full-text: predicting all chemicals mentioned in recently published full-text articles, both span (i.e., named entity recognition) and normalization (i.e., entity linking) using MeSH[1].
- Chemical Indexing prediction task: predicting which chemicals mentioned in recently published full-text articles should be indexed, i.e., appear in the listing of MeSH terms for the document (9).

To support the challenge and address the need of creating high-quality chemical corpora, we developed a rich and comprehensive chemical entity resource that contains manual annotations for chemical entities mentioned in articles' text and manual indexing for the chemical substances that can represent an article's topic and content. This resource consists of three parts:

1. NLM-Chem200

The NLM-Chem200 corpus consists of 200 full-text PMC articles manually annotated for chemical entities by twelve NLM expert annotators. The first 150 articles, provided as the training set, were previously published as the NLM-Chem corpus (4), and the additional 50 full-text articles were specifically annotated for the BioCreative VII challenge and to serve as the Chemical Identification task testing set. Each article was doubly annotated in a three-round annotation process,

*Co-first authors

[1] https://www.nlm.nih.gov/mesh/meshhome.html

where annotator discrepancies were discussed after each round until they reached full consensus. Finally, the articles were enriched with the manually indexed chemical substances.

2. The extended chemical entity annotated collection from previous BioCreative challenges

This resource was created by utilizing the other chemical entity corpora built in previous BioCreative challenges (CHEMDNER (3), and BC5CDR (10)). These articles were further enriched with the manually indexed chemical substances.

3. The Chemical Indexing Task testing dataset

This resource consists of 1,387 recently published full-text articles in the PMC Open Access collection, manually indexed with chemical substances. This set of articles was used as the testing set for the Chemical Indexing task.

II. Methods

A. Document Selection Procedure

The chemical corpus of the NLM-Chem BioCreative VII track had these targets:

- be representative of biomedical literature publications that contain chemical mentions.
- target articles for which human annotation was most valuable
- be instrumental in training Chemical NER algorithms to produce high-quality results in full-text publications, as well as article abstracts.

To select candidate articles for human annotation for the NLM-Chem200 corpus, we evaluated each article to:

- be rich in chemical entities that current NER tools have trouble identifying
- have no restrictions on sharing and distribution
- be useful for other downstream biomedical entity text mining related tasks.

To select the articles most suitable for algorithm testing, in addition to the constraints above, we focused on recently published articles. Chemical NER and indexing algorithms are most valuable for the incoming flux of published literature. As we experienced with the Covid-19 pandemic, correctly identifying chemicals and drugs discussed in the articles, and grouping those articles by the relevant substances, is most crucial, especially in the race to find an effective cure and a timely vaccine.

The 50 full-text articles that constituted the Chemical Identification task testing set were selected to be as similar as possible to the NLM-Chem corpus of 150 full text articles (4), to be complementary, balancing and a suitable test set, that can also serve as a stand-alone corpus (NLM-Chem200). The selection criteria included: maximization of journal coverage to assure variety, similar distribution of chemical mentions and identifiers per article, similar distribution of other biomedical entities per article, and similar language models.

We re-purposed the CHEMDNER and the BC5CDR corpora for the NLM-Chem track challenge. The CHEMDNER documents are title/abstract annotations for chemical NER, and do not include the chemical normalization. However, as this could still be useful for deep learning strategies, we converted all the articles and their annotations in the same format as NLM-Chem corpus documents. The BC5CDR corpus, on the other hand, contains title/abstract chemical annotations and their MeSH identifiers; we therefore converted these documents in the same format.

We filtered the manual MeSH indexing terms assigned to each article in the MEDLINE collection at the National Library of Medicine to extract the chemical substances to support the Chemical Indexing task. These indexing terms represent chemical substances that are important topics in their respective articles, and therefore are valuable for chemical information retrieval. We extracted the indexed chemical substances and enriched the dataset for every article in the NLM-Chem corpus, CHEMDNER, and BC5CDR corpus.

The Chemical Indexing Testing set consisted of recently published articles and was selected using the same criteria for the Chemical Identification testing set. These articles were manually indexed after the completion of the NLM-Chem track challenge during September 2021, and these indexed labels were used as gold standard data for task evaluation.

B. Annotation Guidelines

The complete NLM-Chem corpus annotation guidelines are publicly available with the corpus (4). Here we give a quick summary.

Our guidelines specify which text elements should be tagged, those that should not be tagged, and how to assign the tagged mentions to their corresponding MeSH identifiers. The primary considerations of the annotation guidelines are: (a) what should be labeled as a chemical, (b) how to place the mention boundaries for those labels, and (c) how to associate those mentions with an entity within one of the chemical trees of MeSH.

Creating high-quality guidelines that fit the annotation task required a multi-step iterative process, starting from an initial draft that was revised until clear and refined guidelines were obtained. We found that defining the text-bound annotations of chemical mentions found in full-text articles was not trivial. It required a deep knowledge of chemistry, supported by the consultation of external knowledge sources. The guidelines were prepared by 12 professional MeSH indexers with degrees in Chemistry, Biochemistry, Biological Sciences, and Molecular Biology and an average of 20 years of experience in indexing PubMed literature with Medical Subject Heading indexing terms.

First, it was decided that very general chemical concepts (such as atom(s), moiety (moieties)) and terms that cannot be associated directly to a chemical structure such as molecule(s), drug(s), and polymer(s) should be excluded from the annotation.

In addition, macromolecular biochemicals, namely, proteins (including enzymes), lipids, nucleic acids (DNA, RNA) were excluded from annotation. In addition, embedded chemical concepts in other biomedical entities such as "sodium channel gene," where the chemical concept "sodium" is embedded in a phrase indicating a different type of biochemical entity "gene," were tagged as OTHER. Each rule defined in the guidelines was also represented by one or more illustrative examples to simplify comprehension and application.

C. Annotation Procedure for the NLM-Chem resource

The NLM-Chem200 full-text articles are doubly annotated by 12 NLM experts in three annotation rounds using the TeamTat annotation tool (11). All articles were pre-annotated using the NLM-Chem improved chemical recognition tool (4). Articles were randomly assigned to pairs of annotators in such a way that the annotation burden is equally distributed. The first round of manual annotations consisted of each annotator working on and completing the annotations of the assigned articles independently. At this stage, the annotators do not know the identities of their partners. After completion, these annotations were reviewed by the technical team to identify differences and discrepancies. Inter-annotator agreement was measured. All pairwise annotations were merged into one document, and the agreements and disagreements were marked and made available in the annotation tool for annotation round two. The second round of annotations consisted of each annotator working independently in their own annotation space, without knowing the identities of their partner-annotators. They reviewed their own decisions and considered their partners' decisions editing the documents until they were satisfied. After completion, the annotations were again reviewed, inter-annotator agreement was computed, and remaining differences and discrepancies were analyzed. All annotations were again merged into one document, agreements and remaining disagreements were marked, and the documents were made available to the respective annotators' accounts. In the third and final round of annotations, the annotation partners for each document were revealed, and every pair of annotators collaboratively reviewed and discussed any remaining differences and finalized the shared document annotation reaching complete consensus.

D. Document Format

While annotations can be represented in various formats, we used the BioC (XML and JSON) format due to several considerations: 1) the format (12) supports full-text articles and annotations representing both mention span (location) and entity identifier, 2) articles in the PMC text mining subset (13) are already available in BioC, 3) our annotation tool of choice TeamTat, and the NLM-Chem NER tool already support the format, 4) the format is simple and easy to modify, allowing additional analysis tools to be applied rapidly as needed.

TABLE 1 DATA CHARACTERISTICS OF THE NLM-CHEM TRACK DATASETS

	NLM-Chem200	BC5CDR	CHEMDNER
Number of Chemical Annotations per article (unique)			
Minimum	2 (1)	1 (1)	0 (0)
Maximum	1,318 (214)	55 (22)	67 (40)
Average	300.4 (66.6)	10.6 (4.1)	8.4 (4.6)
Median	279 (60)	9 (3)	7 (4)
Number of Unique MeSH Identifiers per article			
Minimum	1	1	*NA*
Maximum	127	16	*NA*
Average	41.0	3.0	*NA*
Median	39.5	2	*NA*
Number of Unique Indexed Substances per article			
Minimum	0	0	0
Maximum	14	11	19
Average	1.8	2.3	2.2
Median	1	2	2

III. RESULTS

A. Corpus characteristics

The NLM-Chem track chemical resources are rich in manual chemical annotations and currently the largest corpus, compatible with previously annotated corpora, targeted for developing chemical NER text mining tools. The NLM-Chem training dataset of 150 full-text articles contains 38,339 manual chemical mention annotations; corresponding to 4,862 unique chemical name strings, normalized to 1,810 MeSH identifiers. The Chemical Identification task test set of 50 recently published full-text articles contains 3,740 unique chemical strings and 1,352 unique MeSH IDs. The BC5CDR set contains 15,951 chemical mention annotations, corresponding to 2,693 unique chemical name strings, normalized to 1,269 MeSH identifiers. The CHEMDNER set contains 84,036 chemical mention annotations, corresponding to 19,803 unique chemical name strings. The Chemical Indexing test set of 1,387 recently published full-text articles contains 1,677 unique MeSH IDs. The statistics of annotations per article per dataset are detailed in Table 1.

Figure 1 illustrates that new resources such as NLM-Chem200 need to be: 1) compatible – to foster re-use, acknowledge and build on previous efforts of experts, and 2) complementary – to expand on previous knowledge and cover new areas of training data. Further, Figure 1 illustrates the impact of annotations in the full text. As seen, the full text contains much more chemical annotations and a larger variety both in the mention as well as the respective identifiers. The NLM-Chem200 annotated data in full-text articles allows the new algorithms to learn from and explore a space of chemical mentions in the biomedical literature that had not been covered in previously annotated corpora, as illustrated with the overlap with the BC5CDR and CHEMDNER corpora. Finally, NLM-Chem200 and the BC5CDR corpora contain chemical annotations normalized to MeSH identifiers which, via UMLS, can be mapped to different chemical terminologies, as needed. The three resources have been enriched with the MeSH indexed chemical substances, representing chemical topic terms,

opening up new research avenues in chemical information retrieval.

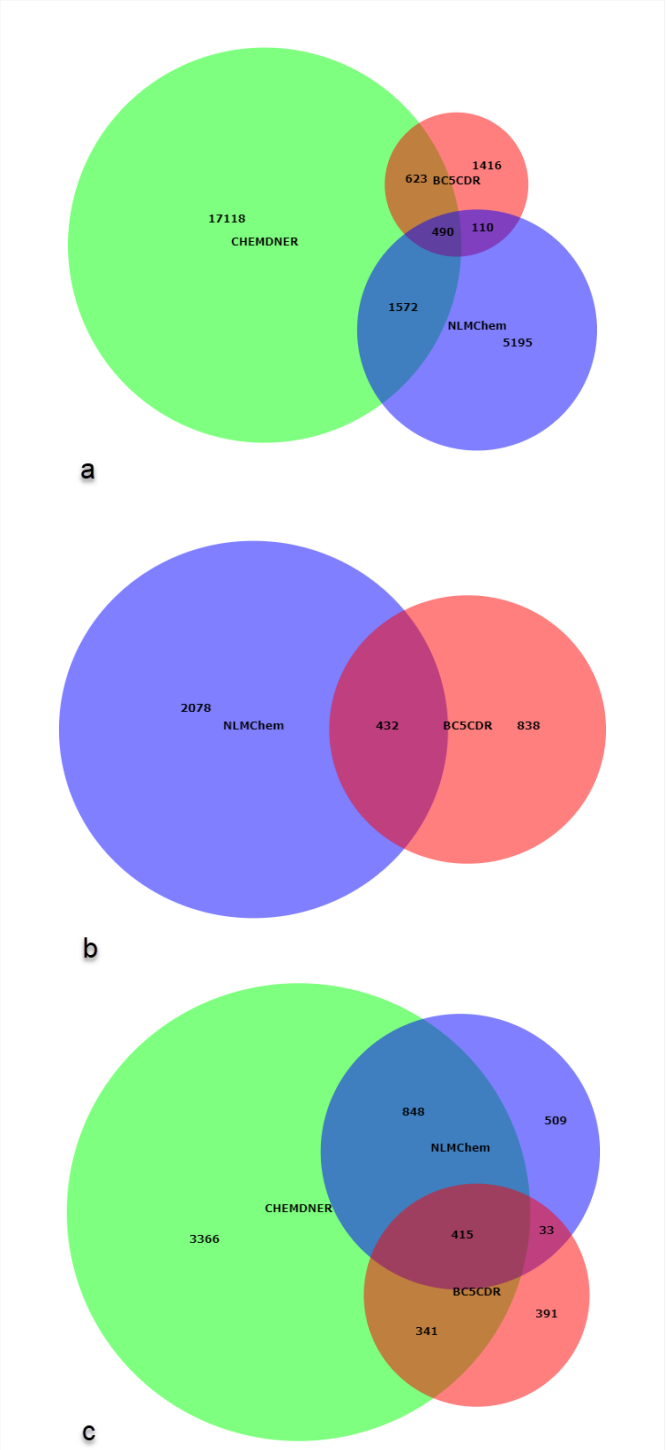

Fig. 1. Illustration of common and different chemical annotations in the chemical annotated data (The overlap of chemical mention annotations is shown in a), the overlap of MeSH ID annotations is shown in b) and the overlap of the indexed chemical substances in the three corpora is shown in c). The NLM-Chem dataset of full-text articles brings in additional previously unseen chemical mentions and chemical identifiers.

B. *Corpus technical validation*

Table 2 shows the results of our benchmark method on the Chemical Identification task. Our benchmark is based on our previously published method and is currently our best performing chemical NER tool. This tool is used in the daily processing of the PubMed and PMC articles as they are queries in our PubTator Central portal (14). This implementation was trained only on the NLM-Chem full-text articles as the training dataset and tested on the NLM-Chem Chemical Identification task (50 full-text articles) dataset. Given the enrichment in chemicals that we observe when we consider the biomedical articles' space covered with the addition of BC5CDR and CHEMDNER corpora, it is reasonable to expect a further improvement in the chemical entity recognition in biomedical articles.

Table 3 shows the results of our baseline method on the Chemical Indexing task. For this task, we added a component to our Chemical Identification benchmark to return the set of MeSH identifiers from annotations found in the title and abstract as the set of indexed chemicals. The indexing component thus represents a straightforward baseline approach with relatively low precision but higher recall.

The strict evaluation for both Chemical Entity Recognition and Normalization tasks assumes an exact match between predicted mention span or MeSH identifier and annotated mention span or MeSH identifier. The relaxed evaluation for Chemical Entity Recognition considers a predicted mention span to match an annotated mention span if they overlap. For chemical entity normalization, which is evaluated both in the Chemical Identification task and the Chemical Indexing task, the relaxed evaluation is the least common ancestor f-score (15).

TABLE 2 BENCHMARK RESULTS CHEMICAL IDENTIFICATION TASK

Chemical Entity Recognition					
Strict			*Approximate*		
Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*
0.8440	0.7877	0.8149	0.9156	0.8492	0.8811
Chemical Entity Normalization					
Strict			*Approximate*		
Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*
0.8151	0.7644	0.7889	0.7917	0.7889	0.7857

TABLE 3 BENCHMARK RESULTS CHEMICAL INDEXING TASK

Chemical Indexing Terms Prediction					
Strict			*Approximate*		
Precision	*Recall*	*F-score*	*Precision*	*Recall*	*F-score*
0.3134	0.6101	0.4141	0.45098	0.78156	0.3134

IV. CONCLUSIONS

The chemical corpus for the NLM-Chem BioCreative track is a high-quality corpus and consists of these parts:

1) The NLM-Chem200 corpus consists of 200 full-text articles doubly annotated by 12 NLM indexers in three rounds of annotation, reaching full consensus and resolving any annotator disagreements. This corpus is currently the largest corpus of full-text articles annotated with chemical entities at a high degree of granularity and their NLM indexed chemical

substances. The NLM-chem training dataset (150 articles) contains a total of 38,339 manual chemical mention annotations, corresponding to 4,862 unique chemical name strings, normalized to 1,810 MeSH identifiers. The NLM-Chem Chemical Identification testing dataset (50 articles) contains 3,740 unique chemical strings and 1,352 unique MeSH IDs. The articles were carefully selected from the PMC Open Access dataset and cover 71 journals.

2) The extended chemical entity annotated collection from previous BioCreative challenges (CHEMDNER and BC5CDR). These articles were enriched with the manually indexed chemical substances.

3) The Chemical Indexing testing dataset. This resource consists of 1,387 recently published full-text articles in the PMC Open Access collection, manually indexed with chemical substances. This set of articles was used as the testing set for the Chemical Indexing task.

To provide a robust test of the corpus utility in chemical entity recognition and normalization that could translate to real life applications, we tested the new corpus with our best performing chemical NER and normalization tool, based on a deep learning architecture for the name entity recognition component and a multi-terminology candidate resolution (MTCR) architecture for the normalization component.

The NLM-Chem track chemical resource provides these contributions: 1) High-quality manual annotation of chemical entities in the full text, 2) Chemical entity normalization to MeSH identifiers, which via UMLS, can be easily mapped to other chemical terminologies, if needed, and 3) Chemical terms indexing of all articles, representing the chemical topic terms for these articles as indexed by the expert literature indexers at the National Library of Medicine. The annotation guidelines are compatible with previously annotated corpora; therefore these (abstract-only) corpora can be used as additional data. The enriched chemical resource of the NLM-Chem track challenge will be invaluable for advancing text-mining techniques for chemical extraction tasks in biomedical text.

ACKNOWLEDGMENT

This research was supported by the NIH Intramural Research Program, National Library of Medicine.

REFERENCES

1. Islamaj Dogan, R., et al., Understanding PubMed user search behavior through log analysis. *Database (Oxford)*, 2009. **2009**: p. bap018.

2. Krallinger, M., et al., Information Retrieval and Text Mining Technologies for Chemistry. *Chem Rev*, 2017. **117**(12): p. 7673-7761.

3. Krallinger, M., et al., The CHEMDNER corpus of chemicals and drugs and its annotation principles. *J Cheminform*, 2015. **7**(Suppl 1 Text mining for chemistry and the CHEMDNER track): p. S2.

4. Islamaj, R., et al., NLM-Chem, a new resource for chemical entity recognition in PubMed full text literature. *Sci Data*, 2021. **8**(1): p. 91.

5. Islamaj Dogan, R., et al., The BioC-BioGRID corpus: full text articles annotated for curation of protein-protein and genetic interactions. *Database (Oxford)*, 2017. **2017**.

6. Bada, M., et al., Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 2012. **13**: p. 161.

7. Kilicoglu, H., Biomedical text mining for research rigor and integrity: tasks, challenges, directions. *Brief Bioinform*, 2018. **19**(6): p. 1400-1414.

8. Leaman, R., R. Islamaj, and Z. Lu, Overview of the NLM-Chem BioCreative VII track: Full-text Chemical Identification and Indexing in PubMed articles. *Proceedings of the seventh BioCreative challenge evaluation workshop.* , 2021.

9. Aronson, A.R., et al., The NLM Indexing Initiative's Medical Text Indexer. *Stud Health Technol Inform*, 2004. **107**(Pt 1): p. 268-72.

10. Li, J., et al., BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database (Oxford)*, 2016. **2016**.

11. Islamaj, R., et al., TeamTat: a collaborative text annotation tool. *Nucleic Acids Res*, 2020. **48**(W1): p. W5-W11.

12. Comeau, D.C., et al., BioC: a minimalist approach to interoperability for biomedical text processing. *Database (Oxford)*, 2013. **2013**: p. bat064.

13. Comeau, D.C., et al., PMC text mining subset in BioC: about three million full-text articles and growing. *Bioinformatics*, 2019. **35**(18): p. 3533-3535.

14. Wei, C.H., et al., PubTator central: automated concept annotation for biomedical full text articles. *Nucleic Acids Res*, 2019. **47**(W1): p. W587-W593.

15. Tsatsaronis, G., et al., An overview of the BIOASQ large-scale biomedical semantic indexing and question answering competition. *BMC Bioinformatics*, 2015. **16**: p. 138.