The overview of the NLM-Chem BioCreative VII track

Full-text Chemical Identification and Indexing in PubMed articles

Robert Leaman*, Rezarta Islamaj* and Zhiyong Lu
National Library of Medicine, National Institutes of Health, Bethesda, MD, USA

***Abstract*— The BioCreative NLM-Chem track calls for a community effort to fine-tune automated recognition of chemical names in biomedical literature. Chemical names are one of the most searched biomedical entities in PubMed and – as highlighted during the COVID-19 pandemic – their identification may significantly advance research in multiple biomedical subfields. While previous community challenges focused on identifying chemical names mentioned in titles and abstracts, the full text contains valuable additional detail. We organized the BioCreative NLM-Chem track to call for a community effort to address automated chemical entity recognition in full-text articles. The track consisted of two tasks: 1) Chemical Identification task, and 2) Chemical Indexing prediction task. For the Chemical Identification task, participants were expected to predict with high accuracy all chemicals mentioned in recently published full-text articles, both span (i.e., named entity recognition) and normalization (i.e., entity linking) using MeSH. For the Chemical Indexing task, participants identified which chemicals should be indexed as topics for the article's topic terms in the NLM article and indexing, i.e., appear in the listing of MeSH terms for the document.**

This manuscript summarizes the BioCreative NLM-Chem track. We received a total of 88 submissions in total from 17 teams worldwide. The highest performance achieved for the Chemical Identification task was 0.8672 f-score (0.8759 precision, 0.8587 recall) for strict NER performance and 0.8136 f-score (0.8621 precision, 0.7702 recall) for strict normalization performance. The highest performance achieved for the Chemical Indexing task was 0.4825 f-score (0.4397 precision, 0.5344 recall). The NLM-Chem track dataset and other challenge materials are publicly available at https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/.

This community challenge demonstrated 1) the current substantial achievements in deep learning technologies can be utilized to further improve automated prediction accuracy, and 2) the Chemical Indexing task is substantially more challenging. We look forward to further development of biomedical text mining methods to respond to the rapid growth of biomedical literature.

Keywords*— *biomedical text mining; natural language processing; artificial intelligence; machine learning; deep learning; text mining; chemical entity recognition; chemical indexing

*Co-first authors

I. Introduction

Identifying named entities is an important building block for many complex knowledge extraction tasks. Errors in identifying relevant biomedical entities are a key impediment to accurate article retrieval, classification, and further understanding of textual semantics, such as relation extraction (1). Chemical entities appear throughout the biomedical research literature and are among the most frequently searched entity types in PubMed (2). Accurate automated identification of the chemicals mentioned in journal publications can translate to improvements in many downstream NLP tasks and biomedical fields; in the near term, specifically in the retrieval of relevant articles, greatly assisting researchers, indexers, and curators (3).

Previous work in biomedical named entity recognition (NER) and normalization (i.e., entity linking) for chemicals includes several community challenges (e.g., CHEMDNER (4) and BC5CDR (5) tasks at previous BioCreative workshops). However, indexing and curation tasks require processing full-text articles, where information retrieval and extraction are different. For example, the full text frequently contains more detailed information, such as chemical compound properties, biological effects, and interactions with diseases, genes and other chemicals.

During the COVID-19 pandemic, the world witnessed the medical research in the search for a cure, treatment, and vaccine that can help ease the suffering worldwide. This process is integrally ingrained with the need to correctly identify chemicals in a timely manner.

To support the efforts of increasing the efficiency and accuracy of the current state of the art algorithms and foster the efforts of researching novel methods and achievements, the NLM-Chem track at BioCreative VII brought together the community to address two tasks:

- 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

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

articles should be indexed, i.e., appear in the listing of MeSH terms for the document.

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 is detailed in the Corpus paper (6) and is available from: https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/.

The NLM-Chem track attracted many participating teams worldwide. Ultimately 14 teams submitted official results. We received 53 submission runs for the Chemical Identification task, of which 50 were official runs, and the rest was submitted after the deadline, and 18 total runs for the Chemical Indexing task, of which 5 were official runs, and the rest were submitted after the deadline. For the Chemical Identification task, 77% and 28% of the submissions had higher performance than the benchmark system provided by the organizers for the strict named entity recognition and normalization metrics, respectively. For the Chemical Indexing task, 44% of the submissions had higher strict performance than the baseline system provided by the organizers. Participating teams explored different methodologies, with the majority focusing on deep learning architectures. Both data and evaluation scripts are available from the workshop webpage, same link as above. We encourage further participation from interested teams on the development of biomedical text mining methods to predict chemical mentions (identification), and/or chemical topic terms (indexing) in biomedical full-text articles.

II. Methods

A. The NLM-Chem track corpus

The goal of BioCreative community challenges is to evaluate text mining and information extraction systems as applied to the biological domain. The main emphasis is on the comparison of methods for scientific progress, rather than on the purely competitive aspects. Regarding identification of chemicals in full text articles, both as mentions in text, as well as topic terms reflecting the articles' indexing, the scientific progress is reflected when the developed method is capable to accurately capture the entities present in previously unseen, recently published articles.

To this end, an appropriate training dataset consists of articles that span a variety of journals, are rich in chemical mentions, and cover a plethora of chemical-related topics to be representative of biomedical literature publications that contain chemical mentions (3). Since we need to provide training data for the identification of articles in the full text, as well as titles and abstracts, the dataset needs to contain expert-annotated full-text articles. We describe the NLM-Chem track dataset in detail in (6), but we give a brief overview.

We specifically selected the NLM-Chem track articles to 1) have no restrictions on sharing and distribution, 2) be useful for other downstream biomedical text mining tasks, 3) be suitable for testing real-world tasks, therefore focused on recently published articles.

The organizers provided three collections of articles to the participants:

The NLM-Chem200 corpus consists of the training dataset, as described in (3), 150 full-text articles, doubly annotated by 12 expert NLM indexers for all chemical mentions and their corresponding MeSH identifiers. For this challenge, we augmented this dataset with 50 additional full-text articles, recently published in Spring 2021, to serve as the testing dataset of the Chemical Identification task. These articles were doubly annotated by the same group of NLM indexers, following the same annotation guidelines, as the NLM-Chem corpus. These articles were enriched with the indexing terms corresponding to their chemical substances, assigned by the NLM indexers following the regular indexing process.

We re-purposed the CHEMDNER (4) and the BC5CDR (5) corpora for the NLM-Chem track challenge. The CHEMDNER documents contain title/abstract annotations for chemical NER, and do not include the chemical normalization, however as this could still be useful for training deep learning strategies, we converted all the articles and their annotations in the same format as NLM-Chem corpus documents. The BC5CDR corpus contains title/abstract chemical annotations and their MeSH identifiers, they were also converted to the same format. All articles were enriched with their chemical substance indexing terms assigned by the NLM indexers during the normal indexing process. The full indexing data was filtered to select only the indexing terms representing chemical substances and provided in the same format.

Finally, the last collection consisted of 1,387 recently published articles, which served as the test dataset of the Chemical Indexing task. These articles were published in Spring 2021 and underwent the normal manual indexing process at the NLM during September 2021, after completing the NLM-Chem track challenge. These indexed labels were used as standard gold data for task evaluation.

B. Baseline methods for Chemical Identification and Indexing

In our previous work (3), we described an improved benchmark tool for chemical entity recognition and normalization to illustrate the value of the NLM-Chem full-text corpus. This tool was based on the bluebert variant of BERT (7), which was trained on PubMed abstracts and clinical notes from MIMIC-III (8). The model was then fine-tuned on the combined BC5CDR and the NLM-Chem training sets to provide chemical mention annotations. To assign MeSH identifiers to the chemical mentions found by the bluebert NER tool, we used our sieve-based normalization system Multiple Terminology Candidate Resolution (MTCR), which is optimized for chemical mention normalization. We resolved all abbreviations that appear in the mention text using abbreviation definitions identified by Ab3P in the full article text (9). Then, each mention text is mapped to a set of candidate MeSH concepts using multiple string-matching methods, applied in sequence, with the first method that returns a non-zero number of MeSH concepts used as the overall result. The earlier methods in the sequence provide higher precision while later methods provide higher recall. These methods can be briefly summarized as: exact match to

terminology vocabulary (MeSH), relaxed match: which allows for certain lexical variations in the sequence, relaxed plural match: which processes tokens using a conservative plural stemmer, relaxed match to multiple chemical terminologies.

We adapted this chemical tagger to provide comparison methods for the Chemical Identification and Chemical Indexing tasks. For the Chemical Identification task, we updated the transformer NER model to BioBERT and updated the normalization component to use the 2021 version of MeSH. Thus the comparison method for Chemical Identification sets a very high benchmark. For the Chemical Indexing task, we added a component 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.

C. Evaluation Measures

The evaluation metrics used to assess team predictions were micro-averaged recall, precision and F-score (main evaluation metric) for the Chemical Identification and Chemical Indexing tasks. Three different result types were scored: False negative (FN) results corresponding to incorrect negative predictions; False positives (FP) predictions corresponding to incorrect positive predictions 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 $r = TP/(TP + FN)$. Precision p (positive predictive value) is the percentage of correctly labeled positive results over all positive labeled results $p = TP/(TP+FP)$. 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. $F_\beta = (1+\beta^2)\cdot p \cdot r/(\beta^2 p+r)$. The balanced F-measure ($\beta = 1$), referred to as "F-score" in this work) can be simplified to $F_1 = 2 \cdot p \cdot r / (p + r)$.

We measure the precision, recall, and F_1 measures in a strict and relaxed evaluation setting. Furthermore, the Chemical Identification task consists of both chemical named entity recognition (NER), and normalization using MeSH identifiers. The strict evaluation for both NER and normalization tasks assumes exact match between predicted mention span or MeSH identifier and annotated mention span or MeSH identifier. The relaxed evaluation for NER 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 (LCaF) (10). This measure identifies an approximately minimal set of ancestor identifiers sufficient to ensure that all predicted and annotated identifiers have at least one ancestor in the set. Both the set of predicted identifiers and the set of annotated identifiers are augmented with the ancestor set, allowing partial credit for a predicted identifier related to the predicted identifier.

The evaluation script was made available to all track participants, together with the data and other challenge materials via: https://ftp.ncbi.nlm.nih.gov/pub/lu/BC7-NLM-Chem-track/.

D. Team Invitations and Challenge

We announced the NLM-Chem chemical recognition challenge in full-text articles in Spring 2021.The NLM-Chem corpus as the training dataset, and the BC5CDR, and the CHEMDNER corpora, as additional data were made available in May 2021. A webinar was held in May 2021 to interested teams to introduce them to the challenge motivation and data collections. The testing dataset for the Chemical Identification task, which complements the NLM-Chem200 corpus, was manually annotated during April-June 2021, and the testing dataset for the Chemical Indexing task was manually indexed via the regular NLM indexing pipeline in September 2021.

Seventeen teams submitted a total of 62 runs for the Chemical Identification task, 6 of which failed the evaluation script and were not evaluated further. Of the remaining 56 runs, three were considered unofficial because they were submitted after the deadline. For the Chemical Indexing task, eight teams submitted a total of 26 runs, 8 of which failed the evaluation script and were not evaluated further. Of the remaining 18 runs, 13 were considered unofficial because they were submitted after the deadline.

III. Results

We received 88 submissions from a total of 17 teams. The participating teams represent 9 nations from Europe, Asia, and North America. Two teams were from industry, with the remainder from universities. The teams reported sizes of 2 to 7 (average 4), typically with backgrounds in natural language processing, machine learning, information retrieval, and/or computer science.

A. Chemical Identification Task Team Submissions

We report the NER performance for all valid submissions in Table 1 and the normalization performance for all valid submissions in Table 2, with the respective performance values for the benchmark and baseline systems. Most teams separated the Chemical Identification task into distinct named entity recognition (NER) and normalization subtasks, combined with a pipeline approach.

NER systems based on BERT transformer models were both popular and performed well. Teams 121 (11), 128 (12), 139 (13), and 143 (14) noted that BERT variants using a vocabulary intended for biomedical text, such as PubMedBERT (15), have noticeably higher performance on NLM-Chem than BERT models using a general vocabulary, such as BioBERT (16). This result may reflect the specialized vocabulary used by chemical names.

TABLE 1 CHEMICAL IDENTIFICATION PERFORMANCE RESULTS: NAMED ENTITY RECOGNITION (* UNOFFICIAL)

Team / Run	Strict			Approximate		
	Precision	Recall	F-score	Precision	Recall	F-score
139 / 3	0.8759	0.8587	**0.8672**	0.9373	0.9161	**0.9266**
139 / 1	0.8747	0.8523	0.8633	0.9361	0.9083	0.9220
139 / 2	0.8775	0.8447	0.8607	0.9441	0.9051	0.9242
128 / 1	0.8544	**0.8658**	0.8600	0.9220	0.9304	0.9262
143 / 1	0.8535	0.8608	0.8571	0.9271	0.9235	0.9253
128 / 4	0.8457	0.8617	0.8536	0.9157	0.9294	0.9225
128 / 2	0.8643	0.8403	0.8521	0.9258	0.8980	0.9117
121 / 2	0.8461	0.8583	0.8521	0.9152	0.9215	0.9183
121 / 1	0.8616	0.8415	0.8515	0.9293	0.9028	0.9158
121 / 3	0.8580	0.8409	0.8494	0.9257	0.9045	0.9149
141 / 1	0.8338	0.8654	0.8493	0.8953	**0.9309**	0.9127
104 / 2	0.8687	0.8249	0.8463	0.9273	0.8791	0.9025
104 / 3	0.8692	0.8239	0.8459	0.9277	0.8761	0.9011
148 / 1	0.8692	0.8239	0.8459	0.9277	0.8761	0.9011
110 / 4	0.8394	0.8515	0.8454	0.9040	0.9229	0.9134
149 / 1	0.8226	0.8614	0.8416	0.8951	0.9204	0.9076
146 / 4	0.8219	0.8622	0.8415	0.8945	0.9235	0.9088
146 / 5	0.8222	0.8609	0.8411	0.8951	0.9204	0.9076
121 / 5	0.8618	0.8209	0.8409	0.9303	0.8822	0.9056
110 / 1	0.8354	0.8429	0.8392	0.9027	0.9186	0.9106
110 / 2	0.8421	0.8350	0.8386	0.9066	0.9081	0.9074
149 / 3	0.8639	0.8136	0.8380	0.9238	0.8682	0.8951
139 / 5	0.8706	0.8068	0.8375	0.9286	0.8566	0.8912
149 / 4	0.8644	0.8123	0.8375	0.9242	0.8650	0.8936
149 / 5	0.8641	0.8121	0.8373	0.9242	0.8651	0.8937
148 / 4	0.8835	0.7893	0.8337	0.9367	0.8341	0.8824
148 / 2	0.8824	0.7898	0.8335	0.9363	0.8371	0.8839
148 / 3	0.8828	0.7890	0.8333	0.9367	0.8345	0.8826
146 / 3	0.8280	0.8382	0.8330	0.8996	0.9031	0.9013
148 / 5	0.8270	0.8388	0.8328	0.8993	0.9061	0.9027
146 / 1	0.8273	0.8375	0.8324	0.8996	0.9031	0.9013
157 / 1	0.8476	0.8101	0.8284	0.9128	0.8670	0.8893
157 / 2	0.8476	0.8101	0.8284	0.9128	0.8670	0.8893
157 / 3	0.8476	0.8101	0.8284	0.9128	0.8670	0.8893
157 / 4	0.8476	0.8101	0.8284	0.9128	0.8670	0.8893
157 / 5	0.8476	0.8101	0.8284	0.9128	0.8670	0.8893
139 / 4	0.8720	0.7885	0.8282	0.9345	0.8359	0.8825
146 / 2	**0.9082**	0.7436	0.8177	0.9561	0.7809	0.8597
104 / 1	0.9077	0.7435	0.8174	**0.9562**	0.7812	0.8599
149 / 2	0.9069	0.7437	0.8173	0.9559	0.7835	0.8611
128 / 3	0.8440	0.7896	0.8159	0.9187	0.8541	0.8852
Benchmark	0.8440	0.7877	0.8149	0.9156	0.8492	0.8811
155 / 1	0.8312	0.7967	0.8136	0.9009	0.8596	0.8798
110 / 3	0.8505	0.7662	0.8062	0.9231	0.8295	0.8738
110 / 5	0.8372	0.7416	0.7865	0.9150	0.8081	0.8583
121 / 4	0.8345	0.7374	0.7830	0.9123	0.7993	0.8521
155 / 3*	0.7676	0.6886	0.7259	0.8881	0.8041	0.8440
155 / 2*	0.7541	0.6011	0.6690	0.8682	0.6964	0.7729
143 / 2	0.7817	0.5552	0.6493	0.8544	0.5990	0.7043
114 / 1	0.7219	0.5897	0.6492	0.8348	0.6919	0.7567
130 / 1	0.7208	0.5211	0.6049	0.8933	0.6331	0.7410
116 / 3	0.8234	0.1916	0.3109	0.9196	0.215	0.3485
116 / 1	0.8207	0.1853	0.3023	0.9143	0.2072	0.3378
116 / 2	0.8419	0.1734	0.2876	0.9291	0.1925	0.3189

TABLE 2 CHEMICAL IDENTIFICATION PERFORMANCE RESULTS: NORMALIZATION (* UNOFFICIAL)

Team / Run	Strict			Approximate		
	Precision	Recall	F-score	Precision	Recall	F-score
110 / 4	**0.8621**	0.7702	**0.8136**	**0.8302**	0.7867	**0.8030**
128 / 2	0.7792	0.8434	0.8101	0.7258	0.8679	0.7864
110 / 1	0.8582	0.7641	0.8084	0.8246	0.7709	0.7910
128 / 1	0.7833	0.8339	0.8078	0.7400	0.8595	0.7909
121 / 1	0.7874	0.8281	0.8072	0.7530	0.8643	0.8015
121 / 3	0.7876	0.8272	0.8069	0.7462	0.8606	0.7959
110 / 2	0.8221	0.7898	0.8056	0.7760	0.8040	0.7849
128 / 4	0.7755	0.8318	0.8027	0.7250	0.8591	0.7822
121 / 2	0.7748	0.8315	0.8021	0.7341	0.8669	0.7914
121 / 5	0.7821	0.8226	0.8019	0.7468	0.8569	0.7936
128 / 3	0.7780	0.8257	0.8011	0.7316	0.8517	0.7827
157 / 3	0.7338	0.8683	0.7954	0.6954	0.8976	0.7760
110 / 3	0.8124	0.7760	0.7938	0.7759	0.8017	0.7828
157 / 5	0.7306	0.8658	0.7925	0.6782	0.8919	0.7625
Benchmark	0.8151	0.7644	0.7889	0.7917	0.7889	0.7857
141 / 1	0.7890	0.7849	0.7870	0.7192	0.8254	0.7628
110 / 5	0.8310	0.7411	0.7835	0.8051	0.7648	0.7781
139 / 2	0.7256	0.8505	0.7831	0.7113	0.8966	0.7883
139 / 4	0.7383	0.8281	0.7806	0.7365	0.8777	0.7957
157 / 2	0.7078	**0.8698**	0.7805	0.6612	**0.9018**	0.7554
139 / 1	0.7212	0.8471	0.7791	0.7107	0.8916	0.7850
157 / 4	0.7038	0.8670	0.7769	0.6424	0.8961	0.7399
157 / 1	0.7038	0.8670	0.7769	0.6421	0.8959	0.7395
155 / 1	0.7886	0.7644	0.7763	0.7309	0.7917	0.7551
139 / 3	0.7120	0.8499	0.7749	0.6924	0.8969	0.7757
121 / 4	0.7571	0.7886	0.7725	0.7311	0.8441	0.7774
139 / 5	0.7300	0.8159	0.7705	0.7257	0.8676	0.7837
155 / 3*	0.7836	0.7243	0.7527	0.7083	0.7499	0.7227
155 / 2*	0.7634	0.7050	0.7330	0.7012	0.7323	0.7094
104 / 1	0.6720	0.7475	0.7078	0.6319	0.8097	0.7039
149 / 2	0.6645	0.7451	0.7025	0.6260	0.8174	0.7033
148 / 4	0.6481	0.7629	0.7008	0.6043	0.8281	0.6923
148 / 3	0.6477	0.7626	0.7004	0.6044	0.8284	0.6925
148 / 2	0.6401	0.7607	0.6952	0.5984	0.8348	0.6909
149 / 4	0.6306	0.7730	0.6946	0.5825	0.8385	0.6811
149 / 5	0.6303	0.7727	0.6943	0.5828	0.8392	0.6815
104 / 2	0.6248	0.7699	0.6898	0.5778	0.8466	0.6813
149 / 3	0.6225	0.7708	0.6887	0.5765	0.8456	0.6793
146 / 1	0.5931	0.7816	0.6744	0.5418	0.8558	0.6581
148 / 5	0.5871	0.7806	0.6702	0.5396	0.8616	0.6580
146 / 2	0.6298	0.7105	0.6677	0.5816	0.7806	0.6617
148 / 1	0.5939	0.7344	0.6567	0.5424	0.8156	0.6473
104 / 3	0.5939	0.7344	0.6567	0.5417	0.8153	0.6468
146 / 3	0.5587	0.7445	0.6384	0.5081	0.8312	0.6262
146 / 5	0.5497	0.7454	0.6328	0.5005	0.8337	0.6206
149 / 1	0.5496	0.7457	0.6328	0.4998	0.8334	0.6201
146 / 4	0.5467	0.7491	0.6321	0.4992	0.8402	0.6216
114 / 1	0.8334	0.4645	0.5965	0.8273	0.5279	0.6368
143 / 1	0.4326	0.6541	0.5208	0.4418	0.8108	0.5664
143 / 2	0.4393	0.5392	0.4842	0.4843	0.7222	0.5736
130 / 1*	0.4575	0.4449	0.4511	0.5461	0.6093	0.5662

In line with previous work on NER in chemicals (17), Teams 121, 139, 141 and 143 reported that ensemble methods are effective (11, 14, 18). In addition, several teams reported that fine-tuning a BERT model directly on the additional datasets (BC5CDR, CHEMDNER) resulted in lower performance than models fine-tuned on only the NLM-Chem data. However, Team 139 reported increased performance by pretraining on the additional datasets, followed by fine-tuning on NLM-Chem (13). Finally, Teams 128 and 139 – the teams with the highest NER performance – both augmented their training sets with synthetic data, either by replacing the annotated chemical mentions with chemical names from a lexicon or with random strings (12, 13).

The normalization methods were significantly more varied; however, most teams used a hybrid of two or more methods, often in a sieve configuration (19). Almost all teams included a dictionary approach using MeSH with string transformations as their basic approach, employing an approximate match approach only if the direct match was unsuccessful. Teams 110, 139, 141 and 157 reported that converting chemical mentions and names to vectors, then identifying the closest match(es) using cosine similarity to be an effective approximate match (13, 18, 20, 21). The results show this approach seems to have been particularly effective

TABLE 3 CHEMICAL INDEXING PERFORMANCE RESULTS (* UNOFFICIAL)

Team / Run	Strict			Approximate		
	Precis-ion	Recall	F-score	Precis-ion	Recall	F-score
128 / 2*	0.4397	0.5344	**0.4825**	0.5519	0.7230	**0.5778**
128 / 1*	0.4424	0.5286	0.4817	0.5538	0.7168	0.5769
110 / 1	0.5351	0.4133	0.4664	0.6097	0.6046	0.5624
110 / 9*	**0.5457**	0.4053	0.4651	0.6188	0.6038	0.5670
110 / 8*	0.4452	0.4736	0.4590	0.5951	0.6705	0.5825
110 / 5	0.5308	0.3812	0.4437	0.6031	0.5794	0.5483
110 / 10*	0.3665	0.5560	0.4418	0.5042	0.7319	0.5494
110 / 7*	0.4514	0.3952	0.4214	0.5929	0.5800	0.5336
Baseline	0.3134	0.6101	0.4141	0.4510	0.7816	0.5329
110 / 4	0.5173	0.3236	0.3981	**0.6322**	0.5395	0.5376
110 / 6*	0.3622	0.4369	0.3961	0.5286	0.6199	0.5090
110 / 2	0.4882	0.3284	0.3927	0.6183	0.5467	0.5348
110 / 3	0.4910	0.3236	0.3901	0.6252	0.5459	0.5402
128 / 4*	0.3805	0.3814	0.3809	0.4737	0.5879	0.4735
157 / 1*	0.2753	**0.6163**	0.3806	0.3889	**0.7924**	0.4865
128 / 3*	0.3776	0.3781	0.3779	0.4695	0.5826	0.4691
141 / 1*	0.4073	0.2822	0.3334	0.4844	0.4612	0.4380
157 / 2*	0.3267	0.3276	0.3271	0.4083	0.5377	0.4189

at achieving high recall; the 5 runs submitted by Team 157 are notable for achieving the 5 highest recall values for the strict normalization evaluation(21). Team 121 found edit distance to be a useful approximate match method, though computationally expensive (11). Relatively few teams reported using additional chemical name resources, though Team 141 reported using an additional 5 vocabularies (18), Team 130 report adding PubChem (22), Team 155 report using UMLS (23), and Team 128 (12) also used the chemical mappings from PubTator (24) as a lexicon.

The teams did not seem to employ a separate step to filter non-chemicals from the results, instead relying on the NER results to determine whether the span was a chemical or not. We note that the three teams who predicted the greatest number of chemical annotations with composite identifiers (i.e., more than one MeSH identifier per mention) were also the three teams with the highest f-scores for the strict normalization evaluation. Finally, some teams reported cascading errors from the NER system, suggesting that an end-to-end approach might be beneficial.

B. Chemical Indexing Task Team Submissions

The task participants reported finding the Chemical Indexing task significantly more challenging than the Chemical Indexing task. This subjective evaluation is supported by the number and types of submissions – a total of 18 submissions, with 13 unofficial – and by the significantly lower performance relative to the Chemical Identification task. The teams reported that the most readily available information for determining whether a specific chemical identifier should be indexed is the document structure, followed by frequency. Team 128 found a binary classifier with engineered features to be effective (12). Teams 110 and 157 both reported using hybrid methods, including a TF-IDF variant to prioritize the chemical identifiers found during the identification task (20, 21).

C. Limitations and Future Work

While we believe that results that fail to receive credit under strict evaluation measures remain useful in many applications, our analysis showed that the approximate measures are highly correlated with the corresponding strict measures. The information value of the approximate measures was therefore relatively limited.

IV. CONCLUSIONS

We presented the NLM-Chem track at BioCreative VII, consisting of two tasks: Chemical Identification and Chemical Indexing. Of the submissions to the Chemical Identification task, 77% achieved higher strict performance for named entity recognition while 28% achieved the same for normalization. Deep learning transformer models reliably improved on the benchmark for named entity recognition, while combining at least one high precision method and one high recall method appeared effective for normalization. Participants found the Chemical Indexing task significantly more challenging – with the maximum f-score less than 0.5 – suggesting a useful direction for additional research.

ACKNOWLEDGMENT

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

REFERENCES

1. Leaman, R., et al., Ten tips for a text-mining-ready article: How to improve automated discoverability and interpretability. *PLoS Biol*, 2020. **18**(6): p. e3000716.

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

3. 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.

4. 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.

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

6. Islamaj, R., et al., The chemical corpus of the NLM-Chem BioCreative VII track: Full-text Chemical Identification and Indexing in PubMed articles in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

7. Peng, Y., S. Yan, and Z. Lu. Transfer Learning in Biomedical Natural Language Processing: An Evaluation of BERT and ELMo on Ten Benchmarking Datasets. in *18th BioNLP Workshop and Shared Task*. 2019. Florence, Italy: Association for Computational Linguistics.

8. Johnson, A.E., et al., MIMIC-III, a freely accessible critical care database. *Sci Data*, 2016. **3**: p. 160035.

9. Sohn, S., et al., Abbreviation definition identification based on automatic precision estimates. *BMC Bioinformatics*, 2008. **9**: p. 402.

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

11. Chiu, Y.-W., et al., Recognizing Chemical Entity in Biomedical Literature using a BERT-based Ensemble Learning Methods for the BioCreative 2021

NLM-Chem Track, in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

12. Erdengasileng, A., et al., A BERT-Based Hybrid System for Chemical Identification and Indexing in Full-Text Articles, in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

13. Kim, H., et al., Improving Tagging Consistency and Entity Coverage for Chemical Identification in Full-text Articles in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

14. Adams, V., et al., Chemical Identification and Indexing in PubMed Articles via BERT and Text-to-Text Approaches, in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

15. Gu, Y., et al., Domain-specific language model pretraining for biomedical natural language processing. *ACM Transactions on Computing for Healthcare (HEALTH)*, 2021. **3**(1): p. 1-23.

16. Lee, J., et al., BioBERT: a pre-trained biomedical language representation model for biomedical text mining. *Bioinformatics*, 2020. **36**(4): p. 1234-1240.

17. Leaman, R., C.H. Wei, and Z. Lu, tmChem: a high performance approach for chemical named entity recognition and normalization. *J Cheminform*, 2015. **7**(Suppl 1 Text mining for chemistry and the CHEMDNER track): p. S3.

18. Bevan, R. and M. Hodgskiss, Fine-tuning transformers for automatic chemical entity identification in PubMed articles, in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

19. D'Souza, J. and V. Ng. Sieve-based entity linking for the biomedical domain. in *Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 2: Short Papers)*. 2015.

20. Almeida, T., et al., Chemical detection and indexing in PubMed full text articles using deep learning and rule-based methods, in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

21. Tsujimura, T., et al., TTI-COIN at BioCreative VII Track 2: Fully neural NER, linking, and indexing models, in *Proceedings of the seventh BioCreative challenge evaluation workshop*. 2021.

22. Kim, S., et al., PubChem in 2021: new data content and improved web interfaces. *Nucleic Acids Res*, 2021. **49**(D1): p. D1388-D1395.

23. Bodenreider, O., The Unified Medical Language System (UMLS): integrating biomedical terminology. *Nucleic Acids Res*, 2004. **32**(Database issue): p. D267-70.

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