The WEAVE Corpus: Annotating Synthetic Chemical Procedures in Patents with Chemical Named Entities

Ravindra Nittala
Language Technology Research Centre,
IIIT - Hyderabad, India
ravindra.n@research.iiit.ac.in

Manish Shrivastava
Language Technology Research Centre,
IIIT - Hyderabad, India
m.shrivastava@iiit.ac.in

Abstract

The modern pharmaceutical industry depends on the iterative design of novel synthetic routes for drugs while not infringing on existing intellectual property rights. Such a design process calls for analyzing many existing synthetic chemical reactions and planning the synthesis of novel chemicals. These procedures have been historically available in unstructured raw text form in publications and patents. To facilitate automated synthetic chemical reactions analysis and design the novel synthetic reactions using Natural Language Processing (NLP) methods, we introduce a Named Entity Recognition (NER) dataset of the Examples section in 180 full-text patent documents with 5188 synthetic procedures annotated by domain experts. All the chemical entities which are part of the synthetic discourse were annotated with suitable class labels. We present the second-largest chemical NER corpus with 100,129 annotations and the highest IAA value of 98.73% (F-measure) on a 45 document subset. We discuss this new resource in detail and highlight some specific challenges in annotating synthetic chemical procedures with chemical named entities. We make the corpus available to the community to promote further research and development of downstream NLP systems applications. We also provide baseline results for the NER model to the community to improve on.

1 Introduction

There is a renewed interest in academia and industry to access the information regarding chemical and chemical reactions currently available in unstructured raw text form in journal publications and patents (Coley et al., 2017; Segler et al., 2018; Mysore et al., 2019) using machine learning. Also, several chemical NER datasets exist. With increasing demand in automated chemical synthesis design and planning novel chemical reactions, we need to shift away from the annotation of title and abstract of patents or reactions in isolation to the patents' core, the Examples section. The CHEMDNER-patents corpus (Krallinger et al., 2015c) is the only dataset focusing on titles and abstracts. The Chapati corpus (Grego et al., 2009) and BioSemnatics corpus (Akhondi et al., 2014) focus on the full text of patents for annotation. The reason for the insufficiency of these corpora is discussed in detail in Section 3.3 and 3.5. The ChEMU labs introduced a named entity dataset with chemical role labels (Nguyen et al., 2020). As part of the dataset, they have annotated only snippets of reaction text from the patents' experimental section. They also acknowledge the problem of entity often referring to context beyond the current reaction text[1]. This context cannot be accounted for by the snippets of reaction text in isolation. As part of the WEAVE corpus, we would like to annotate the chemical entities in their full reaction discourse. This would enable us to model the context beyond the immediate reaction text. We refer readers to the supporting information containing full-text patents to understand how the discourse varies from section to section.

A patent is the grant of a legal right by a patent office to an inventor. This grant provides the inventor exclusive rights for a designated period of time in exchange for comprehensive invention disclosure. The disclosure should be complete, such that a person well versed in the field should be able to reproduce this patented process, design, or invention. This disclosure is done in the Examples section of a patent. Hence the Examples section is fundamentally different in its linguistic structure from other sections in a patent. It is the most useful part of understanding the synthetic chemical

[1] https://chemu-patent-ie.github.io/resources/Annotation_Guidelines_CLEF2020_ChEMU_task1.pdf

Proceedings of the 17th International Conference on Natural Language Processing, pages 1–9
Patna, India, December 18 - 21, 2020.

reactions given in the patent.

1.1 Related work

There is a large body of chemical and biomedical NER literature. We refer readers to Yadav and Bethard (2018) and Huang et al. (2020) for a comprehensive survey. We include a summary of the publicly available datasets as follows: Chapati corpus (Grego et al., 2009) is a manually annotated set of 40 patents with 11,162 annotations. The chemical named entities identified were mapped to the Chemical Entities of Biological Interest (ChEBI) database. BioSemantics corpus (Akhondi et al., 2014) is a manually annotated set of patents. This corpus has two sets: First, a harmonized set of 47 patents with 36,537 annotations, and the second set of 198 patents with 400,125 annotations. Besides chemical entity mentions, they also annotate diseases, targets, modes of actions (MOAs), OCR errors, and spelling errors. It is the largest chemical NER dataset. BC-IV CHEMDNER corpus (Krallinger et al., 2015a) is an annotated set of 10,500 titles and abstracts from the PubMed database with 84,355 annotations. BC-V CHEMDNER-patents corpus (Krallinger et al., 2015c) is an annotated set of 21,000 titles and abstracts from patents with 99,634 annotations. With BC-IV CHEMDNER corpus and BC-V CHEMDNER-patents corpus being the widely cited among these. CHEMDNER-patents corpus exclusively focuses on chemical entity mentions. The entity mention classes are a variant of earlier published CHEMDNER corpus (Krallinger et al., 2015b). Nguyen et al. (2020) have introduced a new evaluation lab named ChEMU. It focuses on two tasks: First, named entity recognition of chemical compounds and assign the compound's role within a chemical reaction. Second, event trigger detection and argument identification of previously detected chemical entities. In the publically available NER dataset[2], there are 20,186 annotations (train + dev) in 1125 reaction snippets extracted from 170 patents.

1.2 Structure of a patent

A typical US patent[3] granted has the following discourse structure: Patent grant number, Title, Bibliography, Abstract, Other Patent Relations, Brief Summary, Detailed Description, and Claims. The intellectual property rights or the innovative part of the patent granted resides in the examples contained in the Detailed Description section. This section will be analyzed thoroughly for any novel synthetic route to be non-infringing on existing intellectual property rights. Therefore in the next section, we present the WEAVE[4] patents corpus, which focuses exclusively on synthetic procedures in the Examples section.

2 The WEAVE patents corpus

An important consideration in preparing a corpus for NER training, development, and evaluation sets is selecting documents representing the distribution of chemical named entities seen in related documents. In the WEAVE corpus, the focus is on synthetic chemical procedures and the chemical entities present. Two considerations influenced document selection in our corpus. First, the documents used in the corpus should be available without copyright protection. Second, they are complementary to existing datasets. We accessed the patents from the United States Patent and Trademark Office (USPTO)[5]. Following criterion were applied to further subset the patents for annotation:

- **IPC code**: The selection of patents for the WEAVE corpus was made based on IPC (International Patent Classification) code. Patents which belonged to at least A61K (Preparations for Medical, Dental, or Toilet purposes)[6] or C07D (Heterocyclic compounds) [7] were selected. This enriched patents with chemical entities in medicinal and organic chemistry. An additional criterion for selection within this subset was the presence of synthetic organic procedures.

- **Date and Publication type**: We decided to select patents that were granted in the years

[2] http://chemu.eng.unimelb.edu.au/

[3] USPTO, https://www.uspto.gov

[4] to form something from several different things or to combine several different things, in a complicated or skilled way https://dictionary.cambridge.org/dictionary/english/weave

[5] USPTO Bulk Data Storage System (BDSS) https://bulkdata.uspto.gov/#pats

[6] https://www.wipo.int/classifications/ipc/en/ITsupport/Version0170101/transformations/ipc/20170101/en/htm/A61K.htm

[7] https://www.wipo.int/classifications/ipc/en/ITsupport/Version0170101/transformations/ipc/20170101/en/htm/C07D.htm

2018 and 2019. This would ensure the availability of patents in XML format and text free from OCR errors.

- **Character encoding and language**: XML character entities were converted to corresponding UTF-8 characters, and the full text was encoded in UTF-8 encoding. As the patents were selected from USPTO, only English language patents were included.

- **Document format**: The patent in XML format was converted to a UTF-8 encoded text file. Only the paragraph elements, headings, subheadings, and tables were written to the text file. All the formatting elements like bold, italics, subscript, and superscript were discarded. Bibliographic details and XML formatting was also discarded. There was no restriction on the number of lines in documents.

- **Documents inclusion and exclusion**: Patents covering Inorganic, Organometallic, Polymers, Natural products, Proteins, DNA/RNA, Polymorphic crystal forms were excluded. The overriding criterion for inclusion was at least one synthetic organic procedure in the Examples section, and this was manually checked in each document.

- **Final document sets**: After applying the above selection criteria and prepossessing, we were left with 180 documents. The summary of these sets is given in Table 1. These were randomly assigned to training, development, and test sets. 45 documents from the above settings were used for the Inter-annotator agreement (IAA). For display performance in BRAT, all patents were split into files of 100 lines each before annotation and later concatenated into a single document after annotation.

Set	Documents	Reactions
Evaluation	45	438
Training	60	1311
Development	60	2020
Test	60	1857
Overall	180	5188

Table 1: Document sets. Evaluation set is a subset of overall 180 documents

3 Corpus annotation

3.1 Annotation tools

Neves and Leser (2012) have surveyed the annotation tools available for biomedical literature. They determined that BRAT was easy to use and customizable as per the annotation scheme among the tools reviewed. Hence we used the BRAT Rapid Annotation Tool (BRAT) (Stenetorp et al., 2012) for the entire annotation process and BRAT standoff format for storing the annotations.

3.2 Evaluation metric

We used CoNLL 2003 shared task (Tjong Kim Sang and De Meulder, 2003) evaluation script to compute the macro averaged F-measure on named entity annotations. The annotation output in the BRAT standoff format was converted to CoNLL 2003 shared task format with BIO tagging representation before computing the F-measure. We used F-measure as the evaluation metric for IAA as suggested by Corbett et al. (2007) and Kolarik et al. (2008). CoNLL 2003 shared task evaluation script evaluates an entity to be valid by matching the chemical mention and class label. The use of F-measure provides an advantage in a direct comparison between system performance and inter-annotator agreement (Grouin and Névéol, 2014).

3.3 Annotation scheme

We had to make a choice of designing our own scheme or utilize an existing scheme. Based on publicly available guidelines and corpora, we had a choice between Chapati corpus by Chemical Entities of Biological Interest (ChEBI) and European Patent Office (EPO) (Grego et al., 2009), BioSematics corpus (Akhondi et al., 2014), CHEMDNER corpus (Krallinger et al., 2015a), CHEMDNER-patents corpus (Krallinger et al., 2015c) and ChEMU Labs NER corpus (Nguyen et al., 2020). In Chapati corpus, 40 patents were manually annotated with 11,162 annotations (Grego et al., 2009). The number of annotated patents and the corresponding number of annotations was small in size.

We were left with a choice between BioSemantics, CHEMDNER, and CHEMDNER-patents corpora. On a closer look at BioSemantics corpus, which was based on 15 rules published in their article (Akhondi et al., 2014), we noticed that the IAA (F-score), when considered for only chemical mentions in the corpus, varies from 0.94 to 0.38 depending on entity type and the agreement between the

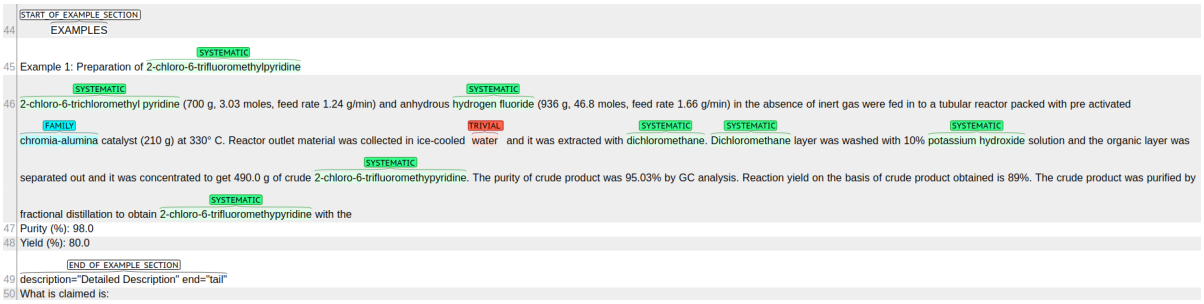

Figure 1: An example an annotated organic reaction, within the Examples section of patent.

four annotator groups on the harmonized patents set (47 patents) (Akhondi et al., 2014). The wide variation in IAA indicates a lack of consistency in guidelines and the need for multiple disambiguation steps. This could be potentially misleading to the annotators.

The near-simultaneous publication of the ChEMU Labs NER dataset[8] (Nguyen et al., 2020) with this publication precluded a full evaluation of the dataset. After reviewing the guidelines[9], it was determined that this dataset is not suitable for the chemical named entity recognition in the full discourse of reaction text in the Examples section.

The extensive guidelines documentation (30 pages), illustrated with examples, led us to choose the annotation scheme developed for BioCreative IV (BC-IV) CHEMDNER (Krallinger et al., 2015a). As modified in BioCreative V (BC-V), CHEMDNER-patents task (Krallinger et al., 2015c) to be used for the WEAVE corpus annotation process. CHEMDNER-patents task had annotated titles and abstracts from 21,000 patents with 99,625 annotations (Krallinger et al., 2015c). SYSTEMATIC, IDENTIFIER, FORMULA, TRIVIAL, ABBREVIATION (ABBV), FAMILY and MULTIPLE entity mention classes as reported by Krallinger et al. (2015c) were utilized. We chose to annotate the Examples section of the patent with synthetic organic procedures against title and abstract only in the CHEMDNER-patents task (Krallinger et al., 2015c). This is illustrated in Figure 1.

3.4 Annotation process

The entire annotation process was done in two stages. The first stage work was done to establish the inter-annotator agreement on the evaluation set of 45 documents. The documents were annotated by nine chemistry domain experts with no formal linguistics experience and were equally divided between them (5 each).

These 45 documents were independently double annotated by another chemistry domain expert, designated as lead annotator with formal linguistics experience. The lead annotator's annotations were designated as the gold standard for evaluating the quality of annotation by the nine annotators. These 45 documents were compared to the gold standard using F-measure. Once the annotation consistency was established, the second stage work was done on the rest of the 135 documents. With each annotator getting 15 documents. Following the concept of annotator-reviser (or adjudicator) agreement (Campillos et al., 2018; Bada et al., 2012), annotators were free to consult the lead annotator throughout the annotation process regarding guidelines.

3.5 IAA statistics

CLASS	Precision	Recall	F1
ABBV.	98.50%	99.88%	99.19
FAMILY	90.86%	97.28%	93.96
FORMULA	98.84%	95.63%	97.21
IDENTIFIER	80.00%	72.73%	76.19
MULTIPLE	75.00%	100.00%	85.71
SYSTEMATIC	99.02%	99.13%	99.07
TRIVIAL	98.85%	100.00%	99.42
Overall	98.66%	98.81%	98.73

Table 2: IAA statistics.

Table 2 presents IAA statistics for 45 documents set. The average F-measure was 98.73%. Bada et al. (2012) have reported 90+% IAA level following the annotator-reviser (or adjudicator) agreement concept. Hence the F-measure reported by us is

[8] http://chemu.eng.unimelb.edu.au/

[9] https://github.com/chemu-patent-ie/chemu-patent-ie.github.io/tree/master/resources/Annotation_Guidelines_CLEF2020_ChEMU_task1.pdf

consistent with published results. This IAA value is the highest reported to date on the chemical entity mention dataset. The F-measure at the micro-level was the lowest for IDENTIFIER (76.19%) and MULTIPLE (85.71%). This can be attributed to the data sparsity in the corpus for these two classes. Tables 4, 5 and 6 demonstrate that the data sparsity for these two classes can also be seen in BC-IV CHEMDNER (Krallinger et al., 2015a) and BC-V CHEMDNER-patents task (Krallinger et al., 2015c).

Akhondi et al. (2014) have reported an annotated chemical patent corpus, which besides chemical mentions, also annotates diseases, protein targets, and MOAs in the patents. The best-reported IAA value among a set of values was 78% (F-score). Krallinger et al. (2015b) in BC-IV CHEMDNER task has reported the IAA value of 91% (F-score) while matching the chemical mention ignoring the class label. When the class label was also considered, the IAA value was 85.26% (F-score). Krallinger et al. (2015c) in BC-V CHEMDNER-patents task have not reported any IAA value and have proposed an IAA study based on a blind annotation of 200 patent abstracts in case of the chemical entity mentions. To the best of our knowledge, this has not yet been published.

Despite no published IAA study for CHEMDNER-patents corpus, we relied on the extensive guidelines published as part of their corpus.

3.6 Error Analysis

Table 3 presents the error analysis of the doubly annotated 45 documents. In the table, rows represent the gold standard labels, and columns represent the annotator's labels. Of the 7503 gold labels, 90 labels (1.2%) were assigned outside the reaction discourse. These should have been assigned to the OTHER class. 78 labels (1.0%) where they should have been assigned one of seven class labels, they were assigned, OTHER class. Only 4 (0.05%) were assigned the incorrect label within the seven class labels.

The error analysis demonstrates that annotators were able to assign the class labels to the chemical entities. The majority of the errors occurred at the boundary of reaction discourse. These errors were communicated to the annotators. They were trained to identify the reaction discourse boundaries and the chemical entities present. They were also encouraged to consult the lead annotator in case of any doubt.

3.7 Corpus statistics

Tables 4, 5 and 6 present the counts of chemical entity mention class labels in the WEAVE corpus (180 documents). These were randomly divided into Training, Development, and Test sets and compared with similar counts from BC-IV CHEMDNER (Krallinger et al., 2015a) and BC-V CHEMDNER-patents task (Krallinger et al., 2015c). Table 7 presents the statistics for the counts of annotations in the WEAVE corpus and CHEMDNER-patents corpus. There are a total of 100,129 annotations with an average of 556 annotations per document. As shown in the table, there is a wide variation between average and median counts per document. This skew is due to a small number of documents having a large number of annotations (Bada et al., 2012). This assertion is supported by the minimum and maximum count across 180 documents.

The top three entity mention classes as a percentage of total annotations in WEAVE corpus was: SYSTEMATIC (49.73%), FORMULA (26.58%), and ABBREVIATION (11.25%). The corresponding distribution of the top three classes in BC-IV CHEMDNER task was: SYSTEMATIC (30.36%), TRIVIAL (22.69%) and ABBREVIATION (15.55%), and in BC-V CHEMDNER-patents task was: FAMILY (36.49%), SYSTEMATIC (28.79%) and TRIVIAL (26.11%). The statistical distribution of entities mentions classes between WEAVE corpus and CHEMDNER-patents corpus is different. Hence the need for annotation of the Examples section of patents was felt. This would significantly help develop machine learning models tailored for the Examples section and downstream processing of synthetic organic reactions in patents.

4 Experiments

To establish some baseline performance parameters for the evaluation of the WEAVE corpus, we applied the NER model[10] developed by Yadav et al. (2018), which has been successfully applied in Multilingual, Clinical and Drug NER. Morphological features have been successfully applied in named entity recognition. In submissions to BC-IV CHEMDNER task (Krallinger et al., 2015a)

[10] https://github.com/vikas95/Pref_Suff_Span_NN

	ABBV.	FAMILY	FORMULA	IDENTIFIER	MULTIPLE	SYSTEMATIC	TRIVIAL	OTHER
ABBV.	855	1	0	0	0	0	0	0
FAMILY	0	179	0	0	0	0	0	5
FORMULA	2	0	854	0	0	0	0	37
IDENTIFIER	0	0	0	8	0	0	1	2
MULTIPLE	0	0	0	0	6	0	0	0
SYSTEMATIC	0	0	0	0	0	4658	0	34
TRIVIAL	0	0	0	0	0	0	861	0
OTHER	11	17	10	2	2	39	9	498807

Table 3: Error analysis of annotations.

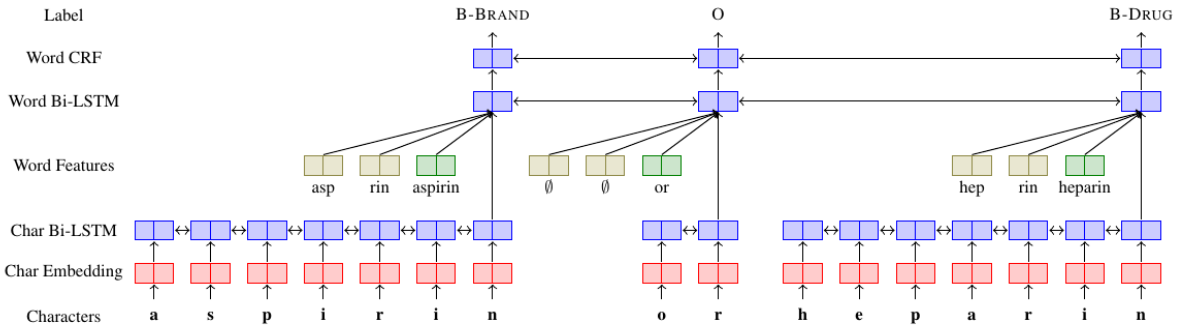

Figure 2: Architecture of NER model proposed by Yadav et al. (2018)

CLASS	BC-IV	BC-V	WEAVE
ABBV.	4538	588	2520
FAMILY	4090	12209	783
FORMULA	4448	2239	6709
IDENTIFIER	672	99	47
MULTIPLE	202	140	6
NO CLASS	40	-	-
SYSTEMATIC	6656	9570	14547
TRIVIAL	8832	8698	2756
Total	29478	33543	27368

Table 4: Training set.

CLASS	BC-IV	BC-V	WEAVE
ABBV.	4521	454	3857
FAMILY	4223	11710	769
FORMULA	4137	2120	9679
IDENTIFIER	639	125	47
MULTIPLE	188	141	13
NO CLASS	32	-	-
SYSTEMATIC	6816	9194	20106
TRIVIAL	8970	8398	4054
Total	29526	32142	38525

Table 5: Development set.

and BC-V CHEMDNER-patents task (Krallinger et al., 2015c) they feature prominently in the top-performing models.

4.1 Word embeddings

200-dimension GloVe embeddings (Pennington et al., 2014) were trained on text extracted from 100,000 US patents belonging to IPC code A61K[11] and C07D[12]. A window of word co-occurrence of

[11] https://www.wipo.int/classifications/ipc/en/ITsupport/Version0170101/transformations/ipc/20170101/en/htm/A61K.htm

[12] https://www.wipo.int/classifications/ipc/en/ITsupport/Version0170101/transformations/ipc/20170101/en/htm/C07D.htm

CLASS	BC-IV	BC-V	WEAVE
ABBV.	4059	331	4892
FAMILY	3622	12319	597
FORMULA	3443	2459	10231
IDENTIFIER	513	54	57
MULTIPLE	199	137	10
NO CLASS	41	-	-
SYSTEMATIC	5666	9818	15145
TRIVIAL	7808	8831	3304
Total	25351	33949	34236

Table 6: Test set.

Type	WEAVE	BC-V
Total annotations	100,129	99,634
Average per document	522	5
Median per document	366	3
Minimum per document	10	0
Maximum per document	3640	233

Table 7: Statistics for counts of annotations

8 and word frequency of 1 was used to train the uncased text. The resulting embeddings had a dictionary size of 6,828,514 and were used for all experiments.

4.2 Model and Hyper-parameters

Figure 2 presents the architecture of the NER model proposed by Yadav et al. (2018). The model features Character Bi-LSTM, Word features, Word Bi-LSTM, and Word CRF layer for generating BIO tags for the named entities. The above model was used as is with minor modifications in hyper-parameters. The word embeddings size of 200-d was used, train_embeddings was set to false, and batch_size was set to 25. All other parameters were set to the default values given in the model proposed by Yadav et al. (2018).

4.3 NER datasets

The WEAVE corpus of the present study was randomly split into training, development, and test set with 60 documents in each set. The official training, development, and test set of CHEMDNER-patents task (Krallinger et al., 2015c) was used without modification.

4.4 Preprocessing

The WEAVE corpus in the BRAT standoff format was converted into CoNLL 2003 BIO format and truncated to the Examples section. The resulting WEAVE corpus had 73,522 sentences, 3,453,525 tokens, and 15,782 unique tokens. The CHEMDNER-patents corpus in a tab-separated format was converted into CoNLL 2003 BIO format before being used in training and evaluation of the model. The resulting CHEMDNER-patents corpus had 73,383 sentences, 2,511,006 tokens, and 51,570 unique tokens.

5 Analysis

To better understand the WEAVE corpus's baseline performance, we conducted several experiments involving BC-V corpus and its combinations with the WEAVE corpus. In Tables 8 and 9 we present the results of experiments on various combinations of WEAVE and BC-V datasets.

Based on the simple NER model (Yadav et al., 2018), the best result in terms of macro-averaged F-measure was the model on standalone WEAVE corpus and tested on WEAVE test set with 91.37%. Followed by a model trained on BC-V + WEAVE corpus and tested on the WEAVE test set with 91.34%. In comparison, the top-performing team in the BC-V CHEMDNER-patents task had an F-score of 89.37% (Krallinger et al., 2015c). Whereas the model trained on standalone BC-V corpus and tested on BC-V test corpus had an F-measure of 80.89%. The model's worst performance was when trained on WEAVE corpus and tested on the BC-V test set; the F-measure was 29.93%.

The results validate the linguistic structure of the title and abstract of a patent is very different from that of the Examples section. Hence, when combined with the CHEMDNER-patents corpus, the WEAVE corpus are complementary; without losing precision, we have an increase in the recall of the NER model. This also supports our assertion of the need for a focused dataset covering the Examples section of patents. The combined corpus can perform very close to the state-of-the-art results in chemical NER. This combination also gives us many high-quality annotations 199,763 (100,129 WEAVE + 99,634 BC-V) to develop better chemical NER models. The IAA value of 98.73% on 45 documents subset and the best NER model with F-measure of 91.37% is instructive of the NER model's simple nature. There is good scope for researching better NER models, which can reduce this difference.

Training	Development	Test	Precision	Recall	F1
BC-V	BC-V	BC-V	78.62	83.30	80.89
BC-V	WEAVE	BC-V	78.21	80.21	79.19
WEAVE	BC-V	BC-V	35.68	25.78	29.93
WEAVE	WEAVE	BC-V	32.40	24.65	27.99
BC-V + WEAVE	BC-V	BC-V	74.50	78.77	76.58
BC-V + WEAVE	WEAVE	BC-V	73.33	76.34	74.80
BC-V + WEAVE	BC-V + WEAVE	BC-V	73.84	77.93	75.83

Table 8: Experimental results with BC-V Test corpus

Training	Development	Test	Precision	Recall	F1
BC-V	BC-V	WEAVE	67.08	50.80	57.82
BC-V	WEAVE	WEAVE	73.32	48.38	58.29
WEAVE	BC-V	WEAVE	93.24	89.11	91.13
WEAVE	**WEAVE**	**WEAVE**	**93.55**	**89.29**	**91.37**
BC-V + WEAVE	BC-V	WEAVE	92.91	88.76	90.79
BC-V + WEAVE	WEAVE	WEAVE	92.54	88.74	90.60
BC-V + WEAVE	**BC-V + WEAVE**	**WEAVE**	**93.43**	**89.34**	**91.34**

Table 9: Experimental results with WEAVE Test corpus.

6 Discussion

Our results show that a focused annotated NER dataset with a simple NER model can achieve near state-of-the-art results. Complementary datasets can achieve high recall without sacrificing the precision of the chemical NER model. This is illustrated by the rows highlighted as bold in Table 9. The reuse of the existing manually annotated dataset results in substantial savings in manual annotation effort.

Chemical NER models with high precision and recall can be used for downstream processing and analysis of chemical reactions in patents. The present annotated dataset would help better temporal modeling of the synthetic procedures given in the Examples section of patents.

We propose to explore more complex NER models. These models can better account for the high IAA values reported by us. In the future, we would explore the possibility of extending this dataset to chemical reaction role labeling for the identified chemical entities.

7 Supporting Information

The WEAVE corpus described in this paper is available at Github repository: https://github.com/nv-ravindra/the-weave-corpus

Acknowledgments

We thank Vincatis Technologies Private Limited, Hyderabad and anonymous reviewers for their help with this publication.

References

Saber A Akhondi, Alexander G Klenner, Christian Tyrchan, Anil K Manchala, Kiran Boppana, Daniel Lowe, Marc Zimmermann, Sarma ARP Jagarlapudi, Roger Sayle, Jan A Kors, et al. 2014. Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. *PloS One*, 9(9):e107477.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, et al. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):161.

Leonardo Campillos, Louise Deléger, Cyril Grouin, Thierry Hamon, Anne-Laure Ligozat, and Aurélie Névéol. 2018. A French clinical corpus with comprehensive semantic annotations: development of the Medical Entity and Relation LIMSI Annotated Text corpus (MERLOT). *Language Resources and Evaluation*, 52(2):571–601.

Connor W Coley, Regina Barzilay, Tommi S Jaakkola, William H Green, and Klavs F Jensen. 2017. Prediction of Organic Reaction Outcomes using Machine Learning. *ACS Central Science*, 3(5):434–443.

Peter Corbett, Colin Batchelor, and Simone Teufel. 2007. Annotation of Chemical Named Entities. In *Biological, translational, and clinical language processing*, pages 57–64, Prague, Czech Republic. Association for Computational Linguistics.

Tiago Grego, Piotr Pezik, Francisco M Couto, and Dietrich Rebholz-Schuhmann. 2009. Identification of Chemical Entities in Patent Documents. In *International Work-Conference on Artificial Neural Networks*, pages 942–949. Springer.

Cyril Grouin and Aurélie Névéol. 2014. De-identification of clinical notes in French: towards a protocol for reference corpus development. *Journal of Biomedical Informatics*, 50:151 – 161. Special Issue on Informatics Methods in Medical Privacy.

Ming-Siang Huang, Po-Ting Lai, Pei-Yen Lin, Yu-Ting You, Richard Tzong-Han Tsai, and Wen-Lian Hsu. 2020. Biomedical named entity recognition and linking datasets: survey and our recent development. *Briefings in Bioinformatics*. Bbaa054.

Corinna Kolarik, Roman Klinger, Christoph M. Friedrich, Martin Hofmann-Apitius, and Juliane Fluck. 2008. Chemical Names: Terminological Resources and Corpora Annotation. In *Workshop on Building and evaluating resources for biomedical text mining (6th edition of the Language Resources and Evaluation Conference)*, pages 51–58.

Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015a. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics*, 7(S1):S1.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M Lowe, et al. 2015b. The CHEMDNER corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7(1):1–17.

Martin Krallinger, Obdulia Rabal, Analia Lourenço, Martin Perez Perez, Gael Perez Rodriguez, Miguel Vazquez, Florian Leitner, Julen Oyarzabal, and Alfonso Valencia. 2015c. Overview of the CHEMDNER patents task. In *Proceedings of the fifth BioCreative challenge evaluation workshop*, pages 63–75.

Sheshera Mysore, Zach Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The Materials Science Procedural Text Corpus: Annotating Materials Synthesis Procedures with Shallow Semantic Structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, Florence, Italy. Association for Computational Linguistics.

Mariana Neves and Ulf Leser. 2012. A survey on annotation tools for the biomedical literature. *Briefings in Bioinformatics*, 15(2):327–340.

Dat Quoc Nguyen, Zenan Zhai, Hiyori Yoshikawa, Biaoyan Fang, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Saber A. Akhondi, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2020. ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In *Advances in Information Retrieval*, pages 572–579, Cham. Springer International Publishing.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global Vectors for Word Representation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar. Association for Computational Linguistics.

Marwin HS Segler, Mike Preuss, and Mark P Waller. 2018. Planning chemical syntheses with deep neural networks and symbolic AI. *Nature*, 555(7698):604–610.

Pontus Stenetorp, Sampo Pyysalo, Goran Topić, Tomoko Ohta, Sophia Ananiadou, and Jun'ichi Tsujii. 2012. BRAT: a Web-based Tool for NLP-Assisted Text Annotation. In *Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics*, pages 102–107, Avignon, France. Association for Computational Linguistics.

Erik F. Tjong Kim Sang and Fien De Meulder. 2003. Introduction to the CoNLL-2003 Shared Task: Language-Independent Named Entity Recognition. In *Proceedings of the Seventh Conference on Natural Language Learning at HLT-NAACL 2003*, pages 142–147.

Vikas Yadav and Steven Bethard. 2018. A Survey on Recent Advances in Named Entity Recognition from Deep Learning models. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 2145–2158, Santa Fe, New Mexico, USA. Association for Computational Linguistics.

Vikas Yadav, Rebecca Sharp, and Steven Bethard. 2018. Deep Affix Features Improve Neural Named Entity Recognizers. In *Proceedings of the Seventh Joint Conference on Lexical and Computational Semantics*, pages 167–172, New Orleans, Louisiana. Association for Computational Linguistics.