The WEAVE 2.0 Corpus: Role Labelled Synthetic Chemical Procedures from Patents with Chemical Named Entities

Shubhangi Dutta
shubhangi.dutta@research.iiit.ac.in

Manish Shrivastava
m.shrivastava@iiit.ac.in

Prabhakar Bhimalapuram
prabhakar.b@iiit.ac.in
International Institute of Information Technology,
Hyderabad, Telangana, India

Abstract

Discovering new reaction pathways lies at the heart of drug discovery and chemical experimentation. A huge amount of drug reaction data lies in unannotated patent texts which are not machine readable. Reaction roles play an important role in analysing chemical pathways, and tracing chemicals through them, and while there is a vast body of chemical data available, the unavailability of reaction role annotated data is a blocker to effectively deploy deep learning methods for reaction discovery. This paper introduces a new dataset, WEAVE 2.0, obtained from chemical patents, along with full, manual, annotations of novel chemical reactions with reaction role information. We also provide baseline and state of the art models for chemical entity recognition from our raw dataset. Our dataset and associated models form the foundation of neural understanding of chemical reaction pathways via reaction roles.

1 Introduction

Chemical discovery relies heavily on discovering new synthesis pathways. The search space of all possible syntheses is extremely high dimensional, and cannot be naively enumerated. Thus, to create novel synthesis pathways, we need methods to rapidly search through existing pathways, derive insights, and compare these.

Machine learning methods have proved highly effective in exploring and organising unstructured data in various fields, including vision, language, and physics. There has also been work done on producing similar datasets and models for chemistry. Prior work has focused on extracting data from research paper abstracts, patents, and medical records. These corpora contain a large number of reactions, and are thus key to extracting useful reactions from the literature. A key step required for this task is extracting Named Entities from these corpora, which are needed for most further tasks, such as text summarisation, knowledge-graph building etc. which can in turn be used for chemical pathway generation.

The WEAVE dataset that was introduced in Nittala and Shrivastava (2020) is comprised of chemical reactions from patent data, annotated for chemical named entities. In our paper, we extend prior work by adding the critical information of *reaction role labels* to chemical named entities, which give important information by allowing the same chemical entity to be recognised in the different roles across reactions, as in the step-by-step processes described in patents, the product of one reaction is often the reactant in the next. Furthermore, we experimentally verify that this information is useful by training *baseline* and *improved* models, with various architectures, to recognise the reaction role labels along with the chemical entities themselves. We formulate the problem as two parts, the type labels and the reaction role labels, which can be combined together, as in the baseline models, or trained together with two different classifier heads, as in the joint model, or as a two step process where the classification is done by two separate models. This lets us explore which of the formulations represents the dataset best. Since we use Machine Learning models for this task, and the large number of labels in the dataset may lead to the data for each label being smaller, we also introduce data augmentation methods to increase the data size.

2 Background

Named Entity Recognition (NER) is a text classification task that involves the identification and classification of Named Entities (NEs) which are entities that are unique identifiers of interest, such as names, places, etc. It is often the first step of other information extraction tasks.

Domain-specific NER, as compared to general NER, and in particular, chemical NER, is the task of recognising named entities that are specialised to a particular domain, in our case, chemical re-

actions. This is a more difficult task for machine learning, as there are fewer datasets available, and requires specialised architecture to ensure syntactically different but semantically similar entities (e.g. different chemical formulae) are recognised as different.

Chemical patents contain descriptions of novel chemical inventions or discoveries. They contain a detailed process such that anyone with the relevant skills is able to replicate the results of the patent. Thus, they tend to contain an "*EXAMPLE*" section that details the steps of these processes. This section contains the chemicals (reactants, reagents, catalysts, and products), as well as the relevant identifying chemical tests, such as mass spectroscopy, NMR spectroscopy etc. This information is relevant to replicate and extrapolate chemical information from these reactions.

2.1 Datasets

There are many existing datasets for chemical NER (and the closely related biomedical NER) which extract named entities from various sources. CHEMDNER (Krallinger et al., 2015) is one of the largest resources, with named entities from 10,000 PubMed abstracts from chemistry-related disciplines. It is one of the largest and most commonly used datasets, also used as the corpus for the BioCreative-IV Task 2, that involves detection of chemical compounds and drugs from larger texts. The Cheminformatics Elsevier Melbourne University (ChEMU) evaluation lab 2020, is an evaluation lab having 1500 chemical reaction snippets, including named entities and roles, from 170 patents. It also has Event Extraction as one of the tasks, which extracts "events" that cause the reaction to proceed from one step to the next. The Pistachio dataset from NextMove is a large reaction dataset, with reactions extracted from 9 million patents, along with a querying and searching feature.

However, we note none of these datasets include role-labelling *along with* the type of chemical entity, that we introduce. This allows the ability of tracing chemical compounds through all the steps of the reaction.

2.2 Models

For NER tasks, specifically, chemical and biomedical NER a combination of BiLSTM and CRF is often used, as shown in Cho and Lee (2019),Luo et al. (2017). In chemical NER, ChemSpot (Rocktäschel et al., 2012) is an early model that uses this architecture, where a CRF along with a dictionary of brand and trivial names to identify NERs. WBI-NER (Rocktäschel et al., 2013) improves upon ChemSpot and makes it purely ML-based, removing the need for a dictionary. Later models include TmChem (Leaman et al., 2015), which has 2 CRFs, both using different features and tokenisations and the model used by Zhang et al. (2016) which uses ChemSpot's output as a feature to generate word embeddings.

The initial ideas of using a CRF and another ML-based model are improved upon in many of these models, and thus this as used as a base for most of our models in this project.

3 WEAVE 2.0

The WEAVE dataset was introduced in Nittala and Shrivastava (2020), and it contains a total of 180 chemical patent documents from the US Patent Office (USPTO). The texts are annotated in the BRAT standoff annotation format (Stenetorp et al., 2012) and classifies the named entities into seven labels: `ABBREVIATION`, `FAMILY`, `FORMULA`, `IDENTIFIER`, `MULTIPLE`, `SYSTEMATIC`, and `TRIVIAL`, based on the naming scheme used to identify the entity.

For example:

> Methanesulphonyl chloride$^{\text{SYSTEMATIC}}$ is added dropwise (1 equiv.) to a solution of the corresponding ethylene glycol$^{\text{SYSTEMATIC}}$ (1 equiv.) and NEt3$^{\text{FORMULA}}$ (0.8 mol equiv.) in THF$^{\text{ABBREVIATION}}$ (100 mL) under an argon$^{\text{SYSTEMATIC}}$ atmosphere and at 0° C.

While a reaction description in a scientific document might have entities belonging to the above mentioned categories, it is important to note that these entities play a specific role within a verbose reaction description.

As can be imagined, there are some very common roles, such as `REACTANT`, `REAGENT`, `PRODUCT`, and `SOLVENT`. But in a reaction description (specifically in a patent document), a number of other interesting categories might be considered important for understanding verbose reaction descriptions. These are reaction participant categories.

Aside from the reaction participants, chemical reactions require specific environments in order to take place. These can comprise of a `CATALYST`, inert gas environments (`ENVINERT`), temperature (`ENVPRES`), pressure (`ENVTEMP`), etc. These are reaction environment categories.

The details and chemical properties of the product of a reaction in a patent may be detailed with relevant entities, including YIELD, and the results of chemical tests such as NMR spectroscopy, MASS spectroscopy, and measurement of the EE or enantiomeric excess. These constitute the yield property categories.

A description of a process in a scientific text may also require references to different parts of the text itself, or to other texts, using discourse connectives. Patent data may therefore contain named STEP, METHOD, and EXAMPLE entities. These, as well as other chemical entities, may be referred to in other parts of the text using either a COREFERENCE for unnamed references (e.g. "above crude product") or a REFERENCETO a named product (e.g. REFERENCETO_STEP "1").

This work introduces the WEAVE 2.0 dataset[1], which contains 33 manually role-labelled documents that comprise randomly chosen subset of the documents from the WEAVE dataset, that contain an expanded tagset. WEAVE 2.0 adds role label tags that introduce reaction roles for each chemical, as well as introduces tags from the environment, yield properties, and discourse connective categories. Therefore, many labels have two parts, separated by an underscore, e.g. SYSTEMATIC_REACTANT. The first part of the labels refers to the "type of nomenclature" of the chemical name in the NER, and is similar to the labels introduced in WEAVE. These labels are referred to as "type labels" through the text to distinguish them from role labels, which are the second parts of the labels, depicting the role of the chemical entity in the reactions.

Using the same example as for the WEAVE dataset above, the WEAVE 2.0 annotations appear as:

Methanesulphonyl chloride$^{\text{SYSTEMATIC_}\textit{REACTANT}}$ is added dropwise (1 equiv.) to a solution of the corresponding ethylene glycol$^{\text{SYSTEMATIC_}\textit{SOLVENT}}$ (1 equiv.) and NEt3$^{\text{FORMULA_}\textit{REACTANT}}$ (0.8 mol equiv.) in THF$^{\text{ABBREVIATION_}\textit{SOLVENT}}$ (100 mL) under an argon$^{\text{SYSTEMATIC_}\textit{ENVINERT}}$ atmosphere and at 0° C.

The corpus therefore contains a much larger number of labels (71 labels), and a total of 17177 Named Entities, as compared to the WEAVE dataset which has 8 labels and 498807 Entities.

All of our experiments are conducted in the CONLL format, converted from the BRAT dataset using the BRAT toolkit provided [2]. The dataset is tokenised using the CONLL tokenisation method for all experiments. The dataset contains names of chemicals which follow the International Union of Pure and Applied Chemistry (IUPAC) nomenclature. These names are often tokenised as multiple words by many standard tokenisers, including the CONLL tokeniser. The dataset is split into training and test corpora with a 70-30 split, and only the EXAMPLE section of the patent is used in training the models, as they contain the highest density of the NERs, and also contains the description of the chemical reaction pathway, which is of interest to us. This reduces the number of labels to 68, and the total number of named entities to 15740.

4 Baselines

We train three baseline models to perform chemical NER on the WEAVE 2.0 dataset. The architecture for all three models is a BiLSTM-CRF classifier, where GloVe embeddings, fine-tuned BERT embeddings, or a combination of the two is provided as input to the classifier.

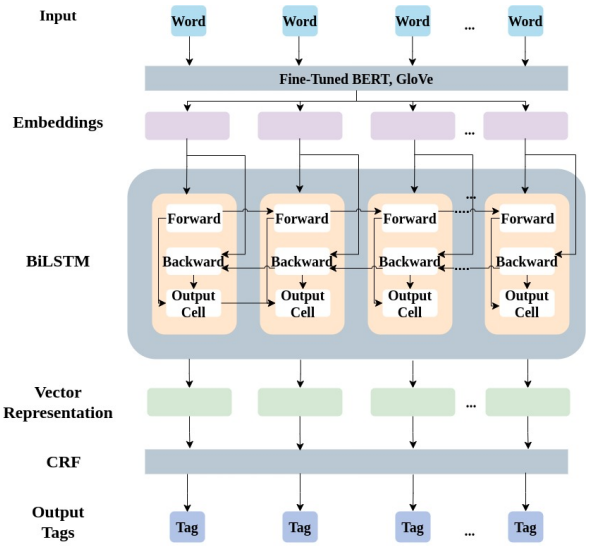

Figure 1: Architecture of baseline model

4.1 Embeddings

Different embeddings were used with the models to check which performed the best. For the baseline models, the embeddings used were the GloVe embeddings and the fine-tuned BERT embeddings.

[1] https://doi.org/10.5281/zenodo.8386296

[2] https://github.com/nlplab/brat

4.1.1 GloVe embeddings

GloVe is an unsupervised, global word embedding algorithm that given a corpus, creates a mapping of the corpus vocabulary to vectors. These vector embeddings encode semantic relationships in the vector space structure. For our task, GloVe word embeddings were trained on a chemical patent corpus taken from the US Patent Trademark Office (USPTO), with patents from 2016 to 2019. About 230,000,000 lines of patent data were used with a window size of 15. Each GloVe vector was of 100 dimensions.

4.1.2 Fine-Tuned BERT Embeddings

BERT is a transformer based language model that performs as a strong baseline in a wide variety of natural language understanding tasks. For our baseline, a pre-trained BERT-uncased model (Devlin et al., 2018) was fine-tuned on patent data from USPTO, with patents from 2016 to 2019, using a total of 300,000 lines of patent data. This was done using Masked Language Modelling (MLM), which masks about 15% of the words in the input, with the model having to predict the masked words. This fine-tuning step ensures domain-specific learning for the embeddings themselves, leading to better representations as well as better recognition of chemistry-related words, as opposed to using the general BERT model. As the size of our dataset is much smaller, compared to the available USPTO dataset, and the NER tagging is not required for this task, we used the latter dataset for fine-tuning the BERT embeddings.

In a separate experiment, the GloVe embeddings and the BERT embeddings were concatenated and used as a sentence embedding as input to the classifier, in order to collate the information both embeddings provide.

4.2 Classifier Architecture

Based on the good performance of BiLSTM-CRF architectures (e.g. Cho and Lee (2019), Dang et al. (2018), Luo et al. (2017)) for NER tasks in general, as well as the usage of a BiLSTM-CRF model for the WEAVE corpus baseline in Nittala and Shrivastava (2020), a BiLSTM-CRF model was used for the baseline models for the WEAVE 2.0 corpus. The BiLSTM was used as the encoder, with 50 hidden states. The output from the BiLSTM was then sent to a CRF layer which classifies the labels. Hyperparameter tuning was also done on the learning rate, number of epochs and the number of hidden layers, with the best performance being at 25 epochs. While only the best results are reported here, the model was trained multiple times to ensure any results were not due to the initial randomness of the weights.

4.3 BERT Embeddings with Fully Connected Layer

This model consists of BERT embeddings with a fully connected layer to act as the classifier. The fine-tuned BERT embeddings are used as an input to a single fully connected layer to create a classifier, in place of the BiLSTM-CRF model, for all the architectures detailed above.

These models generally performed poorly. While transformer architecture is state-of-the-art in many domains, the BERT model in this case is unable to learn a representation that is able to differentiate the different named entities, without a decoder that is able to capture more information.

5 Improved Models

We improve upon our baselines by performing joint improvements to the word embeddings and the classifier, and we also perform data augmentation to correct label imbalances and increase the total number of labels, as certain labels have very few occurrences.

5.1 Improvements to Embeddings

We use pre-trained BERT embeddings [3], trained from the SciBERT checkpoint, with chemical texts, including chemical Wikipedia articles. They perform significantly better than the BERT embeddings which were fine-tuned locally by us on chemical patent data. These embeddings are referred to as ChemicalBERT in this text.

5.2 Classifier Architecture

All the models are based on the baseline model. Each one consists of an embedding that is the input to the model, one or multiple encoder layers, and one or multiple decoder layers. Similar to the baseline models, each improved model was trained multiple times to ensure any results were not due to the initial randomness of the weights.

5.2.1 Joint Model

This model consists of a BiLSTM-CRF based model, with a joint hidden layer, with different clas-

[3] https://huggingface.co/recobo/chemical-bert-uncased

sifier heads for the type labels and role labels. The model is given one input, and it predicts two outputs, classifying each word into the type of chemical NER name it has (e.g. SYSTEMATIC), and the role label (e.g. REACTANT). During training, the loss for both the outputs is back-propagated into the same hidden layer, as well as to the ChemicalBERT model. A joint hidden layer allows for better feature representations in the latent space that contains information from both the type and role labels.

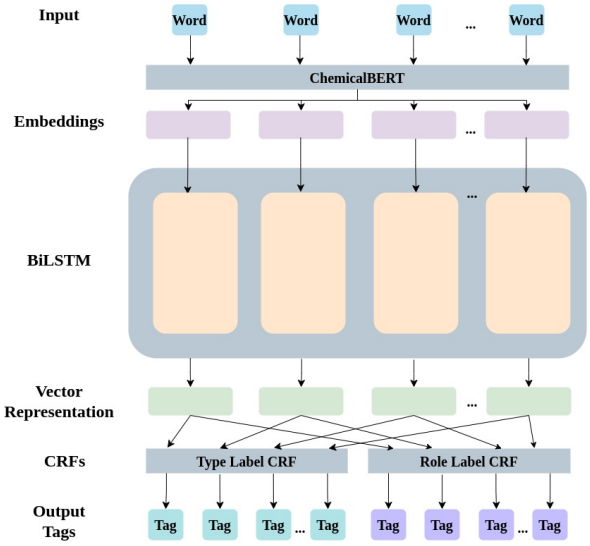

Figure 2: Architecture of ChemicalBERT + Joint Model

5.2.2 Two Step Model

This model consists of a two step process. Each of the individual models used is a BiLSTM-CRF, similar to the baseline model. Each of the the training data labels are split into two labels for this model, creating separate lists for the type and role labels. In the case of the labels that did not have 2 parts, e.g. YIELD, the same tag was repeated in both the sections.

For the first step, the model was trained on the WEAVE 2.0 corpus for the type label, and the model is not sent the role labels. ChemicalBERT generates embeddings for the BiLSTM, and the model predicts the type label associated with each word. This model is trained for a given number of epochs.

The predictions from the first step are then fed into the second model, along with the sentence embeddings from ChemicalBERT. This model then predicts the role labels.

During training, while the second model is trained, the loss is back-propagated into both models, as well as the ChemicalBERT model.

This two-step formulation of the task allows the large number of labels to be reduced, without losing the amount of information. Further, since there are more and larger datasets available for the type labels, (e.g. WEAVE, CHEMDNER), this allows the first step to potentially be trained on a larger dataset and leverage this information to better predict the role labels.

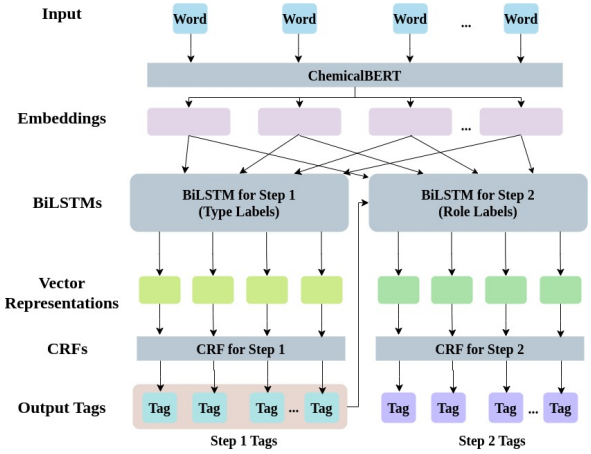

Figure 3: Architecture of ChemicalBERT + 2Step Model

5.2.3 Attention-Based Model

In the attention-based model, in place of the CRF layer, the BiLSTM layer is followed by an attention layer, a fully connected layer, an attention layer, and a final fully connected layer, which acts as the classifier. The same ChemicalBERT embeddings are used for the input. This model performs well generally, which shows that the attention and fully connected layers are a good decoder for the BiLSTM encodings.

5.3 Data Augmentation

The dataset is imbalanced, as it contains a large number of classes but a smaller number of labels. Further, due to the nature of the chemical patents, the number of labels in each class has a large disparity. Since all the models used here are based on neural networks, increasing the quantity and quality of the data leads to better learning by the models. To improve the results in the previous sections, we augment the dataset to correct the large class imbalance. This is done in the following ways:

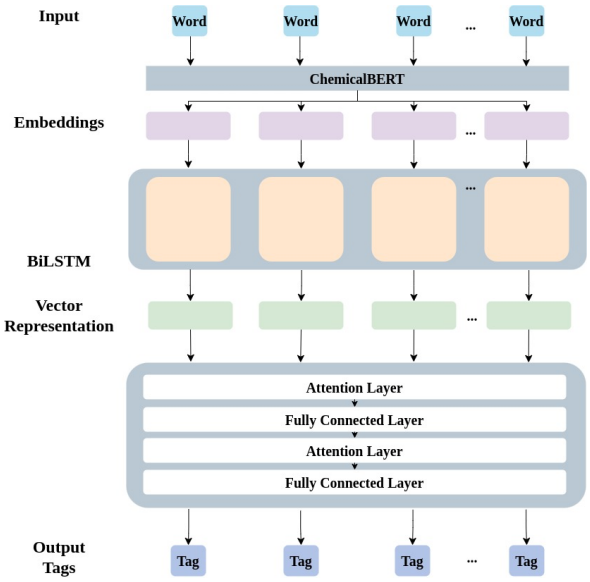

Figure 4: Architecture of ChemicalBERT + Attention Model

5.3.1 Shuffling Sentences

The inputs to all the improved models are sentence embeddings from the BERT models. Therefore, the sentences of the whole training text can be randomly shuffled, while keeping the order of words in each sentence the same, which keeps each sentence embedding the same, which is a method commonly used for data augmentation (Li et al., 2022) (Hu et al., 2020). The resulting corpus is appended to the existing corpus. This does cause the issue of the result text not making sense as a whole document, however, the corpus size and therefore the number of each of the labels is increased to be precisely doubled. This technique can be repeated many times if required, however training the model by repeatedly shuffling may lead to overfitting.

5.3.2 Replacing with Random Strings

Following the process in Task 2 of Erdengasileng et al. (2022), a list of randomly generated strings of length 3-10 characters is created. Some of the Named Entities in the corpus are then probabilistically replaced with one of these strings. The goal is to ensure that the model is able to classify entities it has not seen before in any training data. This method is used in conjunction with the shuffling method described previously. The text generated in this way is appended to the existing corpus to generate the augmented corpus.

5.3.3 Replacing Named Entities

In this method, some of the Named Entities are randomly replaced by other similar Named Entities (e.g. `SYSTEMATIC_REACTANT` may be replaced with `SYSTEMATIC_REAGENT` entities, but not with `TRIVIAL_REAGENT` entities). This leads to sentences that make sense in English, but do not make chemical sense. However, since the goal is to make sure that the model is able to correctly recognise the type and role of the NER based on its position in the sentences, as well as the general structure of the token(s), we are able to augment the dataset using this method. This method is used in conjunction with the shuffling method described previously. This method is the most useful in reducing the class imbalance issue, as the classes with lower number of labels can have a higher number of instances when they are added into the new text. The text generated in this way is appended to the existing corpus to generate the augmented corpus.

All the corpora generated by augmentation methods are tried against the improved models detailed in 5.

6 Evaluation

We study the performance of our dataset by first using the dataset as a baseline, and then augmenting the dataset by (a) shuffling sentences, (b) replacing NERs with random strings, and (c) replacing NERs with semantically similar NERs. We demonstrate that this augmentation is beneficial, with the final dataset providing the best performance. Further, we also test the accuracy of our best model using the CHEMDNER dataset, and achieve a high F1 score.

6.1 Without Data Augmentation

The first experiments were conducted without using any data augmentation, using only the EXAMPLE section of the dataset.

These include the baseline models tabulated in Table 1. The baseline BiLSTM-CRF had an 0.80 F1 score, and 0.81 precision score, when used with the fine-tuned BERT embeddings. The combination of GloVe and BERT embeddings with the same model architecture has the highest recall score of 0.85. The ChemicalBERT + Fully-connected Layer performed worse than all the baseline models.

Most of the improved models, as tabulated in Table 2, performed better, and the best recall and F1 scores are by the Attention-Based Model, with

Model Name	Precision	Recall	F1
GloVe+BiLSTM-CRF	0.80	0.80	0.79
BERT+BiLSTM-CRF	**0.81**	0.83	**0.80**
GloVe+BERT+BiLSTM-CRF	0.72	**0.85**	0.78
BERT+Fully-connected Layer	0.74	0.69	0.71

Table 1: Baseline Models using WEAVE 2.0 corpus

Model Name	Precision	Recall	F1
ChemicalBERT+Joint BiLSTM-CRF	0.79	0.83	0.79
ChemicalBERT+2Step BiLSTM-CRF	**0.83**	0.79	0.80
ChemicalBERT+Attention Model	0.82	**0.85**	**0.82**

Table 2: Improved Models using WEAVE 2.0 corpus

the 2-Step Model having the best precision score.

6.2 With Data Augmentation: Adding shuffling of sentences

The augmented data generated by sentence shuffling (as explained in 5.3.1) is tested on the improved models, as well as the BERT+Fully-connected Layer model. The performance of the BERT+FC model is lower than the other models. The best results are achieved by the Attention Model, however, we note that in general the results of all models improve with this data augmentation technique.

6.3 With Data Augmentation: Adding shuffling of sentences and replacing words with random strings

The improved models and the BERT+FC model are all tested against the augmented data that is generated by the process detailed in 5.3.2 The best results are again achieved using the Attention-based Model, however, this form of augmentation appears to decrease the performance, compared to simply shuffling the sentences.

6.4 With Data Augmentation: Adding shuffling of sentences and replacing NERs with other NERs of similar types

An augmented dataset is generated by the process described in 5.3.3, and then tested against all the improved models and the BERT+FC Model. The

Model Name	Precision	Recall	F1
BERT+Fully-connected Layer	0.71	0.67	0.69
ChemicalBERT+Joint BiLSTM-CRF	0.84	0.87	0.84
ChemicalBERT+2Step BiLSTM-CRF	0.82	0.85	0.80
ChemicalBERT+Attention Model	**0.85**	**0.88**	**0.86**

Table 3: Models using shuffled sentences

Model Name	Precision	Recall	F1
BERT+Fully connected Layer	0.71	0.67	0.69
ChemicalBERT+Joint BiLSTM-CRF	0.84	0.87	0.84
ChemicalBERT+2Step BiLSTM-CRF	0.80	0.84	0.80
ChemicalBERT+Attention Model	**0.84**	**0.87**	**0.86**

Table 4: Models using shuffled sentences and replacing words with random strings

Model Name	Precision	Recall	F1
BERT+Fully connected Layer	0.71	0.67	0.68
ChemicalBERT+Joint BiLSTM-CRF	0.84	0.87	0.86
ChemicalBERT+2Step BiLSTM-CRF	0.82	0.85	0.80
ChemicalBERT+Attention Model	**0.85**	**0.88**	**0.87**

Table 5: Models using shuffled sentences and replacing NERs with similar NERs

results are tabulated in Table 4. The Attention-based Model performs the best, however, all the model results show that this augmentation method produces the best results overall, as the F1 score increases for all the models.

6.5 Performance of Best Model on CHEMDNER

The CHEMDNER corpus is a widely used corpus in chemical NER. It was also used by the WEAVE (Nittala and Shrivastava, 2020) dataset as a comparison. however, due to a lack of role labels, and therefore also having a smaller number of labels, it does not have the same task description as WEAVE 2.0.

We show that our Attention-Based Architecture performs well across all the datasets. When trained and tested on the CHEMDNER dataset it achieves 95% precision, 96%recall and a 95% F1 score. We show it can therefore be used in similar tasks.

7 Conclusion

We introduce a new dataset, WEAVE 2.0, using actual manually annotated patent data, that adds **role labels** alongside the existing kinds of labels, usually denoting type of nomenclature of the chemical entity, to chemical NER tasks, that would enable downstream tasks to have more information, and allow easier tracking and searching of chemical entities through patents. Training models on this dataset also enables other, unannotated patent data, as well as data annotated without role labels to be classified using role labels. The dataset also presents a challenging task due to the high number of labels, each of which has two parts, and can thus be formulated in different ways in different

architectures.

We also introduce baseline models for the dataset, as well as improved models, that are structured for a two-part label, domain-specific task, including domain specific embeddings. We show that these improved models not only perform better than the baseline, but on comparing the best model on a different but similar task (CHEMDNER), it is able to achieve good results.

References

Hyejin Cho and Hyunju Lee. 2019. Biomedical named entity recognition using deep neural networks with contextual information. *BMC Bioinformatics*, 20.

Thanh Hai Dang, Hoang-Quynh Le, Trang M Nguyen, and Sinh T Vu. 2018. D3NER: biomedical named entity recognition using CRF-biLSTM improved with fine-tuned embeddings of various linguistic information. *Bioinformatics*, 34(20):3539–3546.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. Cite arxiv:1810.04805Comment: 13 pages.

Arslan Erdengasileng, Qing Han, Tingting Zhao, Shubo Tian, Xin Sui, Keqiao Li, Wanjing Wang, Jian Wang, Ting Hu, Feng Pan, Yuan Zhang, and Jinfeng Zhang. 2022. Pre-trained models, data augmentation, and ensemble learning for biomedical information extraction and document classification. *Database*, 2022. Baac066.

Anwen Hu, Zhicheng Dou, Jian-Yun Nie, and Ji-Rong Wen. 2020. Leveraging multi-token entities in document-level named entity recognition. *Proceedings of the AAAI Conference on Artificial Intelligence*, 34(05):7961–7968.

Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel Lowe, Roger Sayle, Riza Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Wang Qi, and Alfonso Valencia. 2015. The chemdner corpus of chemicals and drugs and its annotation principles. *Journal of Cheminformatics*, 7:S2.

Robert Leaman, Chih-Hsuan Wei, and Zhiyong lu. 2015. Tmchem: A high performance approach for chemical named entity recognition and normalization. *Journal of Cheminformatics*, 7:S3.

Bohan Li, Yutai Hou, and Wanxiang Che. 2022. Data augmentation approaches in natural language processing: A survey. *AI Open*, 3:71–90.

Ling Luo, Zhihao Yang, Pei Yang, Yin Zhang, Lei Wang, Hongfei Lin, and Jian Wang. 2017. An attention-based BiLSTM-CRF approach to document-level chemical named entity recognition. *Bioinformatics*, 34(8):1381–1388.

Ravindra Nittala and Manish Shrivastava. 2020. The WEAVE corpus: Annotating synthetic chemical procedures in patents with chemical named entities. In *Proceedings of the 17th International Conference on Natural Language Processing (ICON)*, pages 1–9, Indian Institute of Technology Patna, Patna, India. NLP Association of India (NLPAI).

Tim Rocktäschel, Torsten Huber, Michael Weidlich, and Ulf Leser. 2013. Wbi-ner: The impact of domain-specific features on the performance of identifying and classifying mentions of drugs.

Tim Rocktäschel, Michael Weidlich, and Ulf Leser. 2012. ChemSpot: a hybrid system for chemical named entity recognition. *Bioinformatics*, 28(12):1633–1640.

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.

Yaoyun Zhang, Jun Xu, Hui Chen, Jingqi Wang, Yonghui Wu, Manu Prakasam, and Wang Qi. 2016. Chemical named entity recognition in patents by domain knowledge and unsupervised feature learning. *Database*, 2016:baw049.

A Appendix: Distribution of Tags in the WEAVE 2.0 dataset

The following images show the distribution of the type and role labels in the WEAVE 2.0 corpus. The frequency of each tag is also depicted in tabular form.

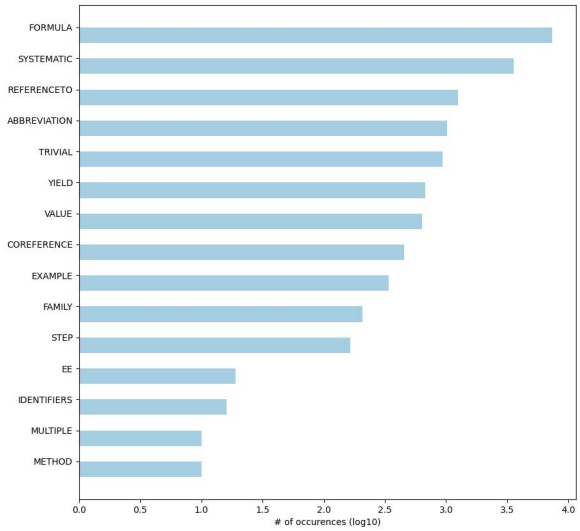

Figure 5: Distribution of the type labels in WEAVE 2.0

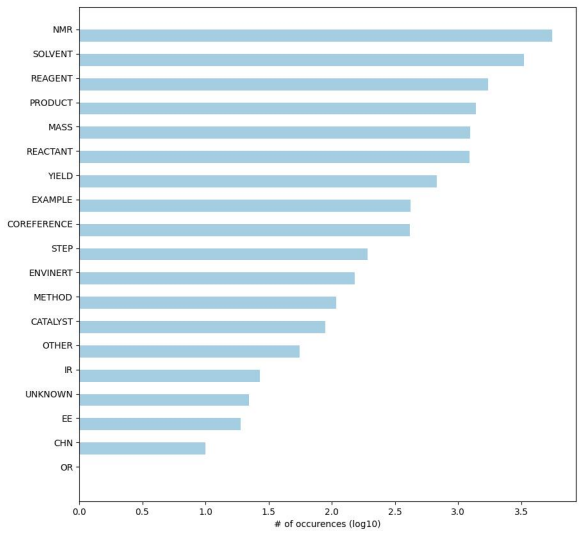

Figure 6: Distribution of the role labels in WEAVE 2.0

Label	Occurences
TRIVIAL_CHN	1
TRIVIAL_MASS	1
MULTIPLE_UNKNOWN	1
REFERENCETO_REAGENT	1
REFERENCETO_NMR	1
SYSTEMATIC_MASS	1
FAMILY_NMR	1
TRIVIAL_UNKNOWN	1
SYSTEMATIC_OR	1
MULTIPLE_REACTANT	2
TRIVIAL_CATALYST	2
ABBREVIATION_PRODUCT	2
ABBREVIATION_UNKNOWN	2
IDENTIFIERS_REAGENT	3
TRIVIAL_NMR	3
REFERENCETO_OTHER	3
MULTIPLE_SOLVENT	3
MULTIPLE_PRODUCT	4
SYSTEMATIC_UNKNOWN	4
TRIVIAL_REACTANT	4
IDENTIFIERS_CATALYST	4
FAMILY_REACTANT	5
FAMILY_OTHER	5
IDENTIFIERS_UNKNOWN	8
FAMILY_REAGENT	8
FORMULA_CHN	9
SYSTEMATIC_NMR	9
FAMILY_UNKNOWN	9
ABBREVIATION_CATALYST	9
ABBREVIATION_OTHER	9
METHOD	10
IDENTIFIERS_OTHER	12
TRIVIAL_PRODUCT	13
COREFERENCE_PRODUCT	15
TRIVIAL_OTHER	15
FORMULA_ENVINERT	17
EE	19
COREFERENCE_REACTANT	20
FAMILY_PRODUCT	21
FORMULA_CATALYST	27
SYSTEMATIC_IR	27
FORMULA_REACTANT	35
REFERENCETO_STEP	40
SYSTEMATIC_CATALYST	51
FORMULA_OTHER	52
ABBREVIATION_REACTANT	66
REFERENCETO_METHOD	71
SYSTEMATIC_OTHER	88
REFERENCETO_EXAMPLE	94
ABBREVIATION_REAGENT	100
SYSTEMATIC_ENVINERT	150
STEP	152
FAMILY_SOLVENT	167
FORMULA_PRODUCT	168
REFERENCETO_SOLVENT	250
ABBREVIATION_NMR	262
REFERENCETO_REACTANT	321
TRIVIAL_REAGENT	330
EXAMPLE	359
REFERENCETO_PRODUCT	407
COREFERENCE_COREFERENCE	420
FORMULA_REAGENT	569
ABBREVIATION_SOLVENT	604
FORMULA_MASS	608
TRIVIAL_SOLVENT	629
VALUE_MASS	636
FORMULA_SOLVENT	668
YIELD	686
SYSTEMATIC_REAGENT	787
SYSTEMATIC_REACTANT	816
SYSTEMATIC_PRODUCT	819
SYSTEMATIC_SOLVENT	1130
FORMULA_NMR	5330

Table 6: Tabular distribution of the labels in the WEAVE2.0 corpus