Practical Work Report: Named Entity Recognition for Chemical Patent Mining

Christian Cadisch*

Supervised by: Dr. Lokesh Mishra†, Dr. Ingmar Meijer†, Lucas Morin*†, Dr. Peter Staar†, Prof. Dr. Fisher Yu*

*ETH Zurich, Switzerland
†IBM Research, Switzerland

***Abstract*—The automatic and accurate identification of chemical names in patent documents is a crucial first step in chemical information extraction pipelines. A precise chemical entity recognition (CER) in patent documents enables an efficient querying of information from this extensive data source. However, this is a formidable challenge due to the diversity of naming conventions and a lack of training data.**

In this practical work, we introduce a novel corpus consisting of 2,700 paragraphs from randomly selected patent documents. We annotate all chemical entities and divide them into three categories, of which we regard chemicals named following the IUPAC nomenclature as the most important category for future downstream processing.

We propose two generative and one model based on the BERT architecture trained on this corpus. Furthermore, we evaluate the few-shot performance of a Falcon 180B model utilized for pre-annotations. Our best model, based on the DistilRoBERTa architecture, reaches a F_1-score of 80% when evaluated on all chemical entities and a F_1-score of 93% when evaluated on the most relevant category of chemical entities.

I. Introduction

In chemical research, patent documents are an important resource for checking freedom-to-operate and tracking recent advances in the field. The molecules found in patent documents may be described within the text or depicted in figures. This practical work focuses on extracting information from the textual descriptions.

In patent document texts, the chemicals are most commonly referred to by names that adhere to the nomenclature standards set by the International Union of Pure and Applied Chemistry (IUPAC) [1]. Alongside IUPAC nomenclature, chemicals may also be named with their trivial names (for example water or vitamins) or by specific brand names. Furthermore, patent texts often also include chemical families, formulas, polymers, substructures and Markush descriptions [2].

However, to make the information more easily queryable and enable a search based on structural similarity, the molecules need to be stored in a unified and simple encoding. A popular encoding for chemical compounds is the simplified molecular-input line-entry system (SMILES) [3]. SMILES encodes molecular structures using short ASCII strings, allowing for a compact representation of chemical compounds.

The objective of this practical work was to tackle the initial task of accurately performing chemical entity recognition (CER). This effort is the foundation for the future development of a translation tool that can translate molecules from their IUPAC names into their SMILES encoding.

In this work, we make the following two contributions. First, we introduce a new high-quality, manually annotated corpus of chemical entities extracted from patent documents. To the best of our knowledge, such a corpus of annotated IUPAC names tailored for patent documents does not exist. The corpus comprises 2,700 paragraphs, each containing at least one chemical entity, resulting in a total of 22,131 annotated entities. The paragraphs were randomly sampled from patent filings in the organic chemistry domain from the US Patent and Trademark Office. We develop an efficient annotation pipeline leveraging the few-shot capabilities of large language models to create pre-annotations on the dataset. Second, we propose and evaluate two types of baseline models, namely a Bidirectional Encoder Representations from Transformers (BERT) model [4] and a generative Text-to-Text Transformer (T5) [5]. Both model types were trained on an augmented version of our dataset.

II. Related Work

This section provides a summary of previous research on CER within text documents. First, we will present frequently used corpora in the CER domain and second, we will give an overview of model architectures that achieve state-of-the-art performances on these corpora.

A. Existing Corpora for CER

Most corpora for CER either originate from scientific literature or patent documents.

One widely used and publicly available corpus from scientific literature is NLM-Chem [6]. It consists of 150 full-text articles from PubMed, that were annotated by ten expert indexers from the National Library of Medicine (NLM). The dataset contains a total of 38,342 manual chemical mention annotations, 4,867 of which are unique chemical

mentions. Their annotation guidelines state that generally, all chemical entities were annotated excluding three types of chemicals. First, very general chemical concepts such as atoms and moieties, second, entities that do not have a one-to-one structural mapping to compounds such as polymers, and third, macromolecular biochemicals such as proteins, nucleic acids or lipids were not annotated.

The CHEMDNER corpus [7] is another popular corpus using scientific literature. It was created using 10,000 abstracts from PubMed and contains a total of 84,355 chemical entity annotations. The annotation guidelines define the following seven labels: abbreviation, family, formula, identifier, multiple, systematic and trivial.

The SCAI corpus [8] is a third frequently used corpus based on scientific literature. The dataset was curated from abstracts of MEDLINE publications and annotations were made for IUPAC names, partial IUPAC names, trivial names, abbreviations, chemical families and formulas.

The ChEMU corpus [9] (Cheminformatics Elsevier Melbourne University) is an established dataset created using patent documents. This corpus consists of 1500 chemical reaction snippets sampled from 170 English patent documents originating from both the European Patent Office and the United States Patent and Trademark Office. The dataset is aimed for building models used to extract information from chemical reactions. Accordingly, the annotation guidelines distinguish a total of 10 different entities, describing the roles of chemicals (e.g., reaction product or solvent) as well as quantitative entities relevant for the reactions (e.g., reaction time).

B. CER Model Architectures

Most models achieving state-of-the-art performances for the CER task are either based on Bidirectional Encoder Representations from Transformers (BERT) [4], Long Short-Term Memory architectures (LSTM) [10], Conditional Random Fields (CRF) [11] or a combination thereof. More recently, generative models also reached high performance, but still lag behind the performance of the previously mentioned architectures. In the subsequent paragraphs, there is a concise overview of these architectures and publications that have employed them are presented.

1) BERT: The BERT model architecture is based on the transformer introduced by Vaswani et al [12]. The model is pre-trained on unlabeled text data by jointly conditioning on both left and right context and designed to learn general language representations. There are two model sizes available, BERT-base with 110M parameters and BERT-large with 340M parameters. Copara et al. [13] trained an ensemble model based on five BERT models and used a majority vote rule for tagging the entities. The models were fine-tuned on the ChEMU dataset and evaluated against two baseline models, a CRF and a convolutional neural network (CNN). With a F_1-score of 92.3%, their ensemble approach proved to outperform the two baseline models as well as the individual BERT models.

2) LSTM: LSTMs are a type of recurrent neural network (RNN) which have proven to be effective for learning from sequences of data. Their architecture includes memory cells and gates to regulate the flow of information and thereby overcome issues related to vanishing or exploding gradients usually faced by standard RNNs. Usually, LSTMs are not used in isolation but combined with BERT models or CRFs, as described in section II-B4.

3) CRF: CRFs are a type of discriminative undirected probabilistic graphical model. In CER models, the graph consists of the word which the CRF should predict and its surrounding words, allowing the CRF to model dependencies between words and learning a broader linguistic context. Rocktätschel et al. [14] proposed ChemSpot, a CRF-based model trained for identifying IUPAC entities. ChemSpot was trained on the SCAI corpus and achieved a F_1-score of 68.1%. Leaman et al. [15] proposed tmChem, which is an ensemble model relying on two differently tuned CRF models trained on the CHEMDNER corpus. The authors used two different tokenizers, feature sets, CRF implementations and CRF parameters and thereby achieved a F_1-score of 87.4%.

4) Hybrid models: Liu et al. [16] proposed a hybrid deep learning approach, that consists of four parts: a BERT module, a CRF, a bidirectional LSTM (BiLSTM) and a multi-head attention mechanism. The authors trained and evaluated the model on the CHEMDNER corpus and showed that this hybrid architecture with a F_1-score of 90.8% outperformed existing models. Furthermore, they demonstrated a substantial improvement in recognizing abbreviations, polysemes, and low-frequency entities. Hemati et al. [17] proposed using a two-stage tool based on LSTM networks. In the first stage, the input text is processed by BiLSTM networks and their output is forwarded to a CRF for the final CER prediction. The model was trained and evaluated on the CHEMDNER dataset and achieved a F_1-score of 90.0%.

5) Generative models: Following recent advances in generative models, there have been attempts to formulate the CER task as text-to-text problem and fine-tune generative language models for it. Adams et al. [18] used HTML-like tags (e.g., `<ENTITY>`) to fine-tune a Generative Pre-trained Transformer model (GPT) [19] for a CER task and reported a F_1-score of 62.1% on the NLM-Chem dataset.

III. Methods

This section first introduces the task formulation and describes how the labels for the CER task were defined. Then, the annotation, pre-annotation, and augmentation strategies are presented. Finally, the model architectures and implementation details for the CER models are outlined.

Label	Subclasses	Examples
Complex Chemicals	Chemicals following IUPAC nomenclature with numbers or element symbol	2-[[2-[benzyl(butanoyl)amino]acetyl]amino]-4,5,6,7-tetrahydro-1-benzothiophene-3-carboxamide, 2-(benzhydrylamino)-N-(2-cyanophenyl)acetamide
Simple Chemicals	Chemicals following IUPAC nomenclature without numbers or element symbols	Carbon dioxide, Dinitrogen
	Chemical families and group of compounds	Alkyl, Halogen
	Trivial names	Water, Aspirin
	Formulas	HCl, NaCl
	Polymers	Polypropylene, Polyvinylchloride
	Substructures	Methyl, Ethyl
Markush descriptions	Sentences containing Markush structures with a placeholder letter	R-SH where R is an unsubstituted or substituted alkyl

Table I: **Classification of chemical entities.** The table gives an overview of the labels considered in the CER task. We differentiate between *complex chemicals*, which represent most relevant category of chemicals, *simple chemicals* and *Markush descriptions*. The *simple chemicals* category consists of multiple subclasses of chemical entities, while the *complex chemicals* and *Markush descriptions* only consist of one.

A. Task formulation

We formulate a new task that differentiates three categories of chemical entities, namely *complex chemicals*, *simple chemicals* and *Markush descriptions*. Table I shows an overview of our classification scheme. We aimed at creating precise annotation guidelines that define the CER categories in much detail, such that the corpus can easily be used or extended for future work. In the sections below, we specify what terms are considered a chemical entity, the specific categories they fall into and the guidelines for how entity boundaries are set.

1) Complex Chemicals: *Complex chemicals* are the most relevant category for chemical research. *Complex chemicals* were defined as chemicals named following the IUPAC nomenclature that contain at least one number or an element symbol. These chemicals are fully translatable into a SMILES encoding, as their names encapsulate all the necessary structural information. We have classified these entities as "high-value" chemical names, reflecting their complex structures which are of particular interest to chemical research.

"2-[[2-[benzyl(butanoyl)amino]acetyl]amino]-4,5,6,7-tetrahydro-1-benzothiophene-3-carboxamide" and 2-(benzhydrylamino)-N-(2-cyanophenyl)acetamide are examples of *complex chemicals*.

2) Simple Chemicals: *Simple chemicals* were defined as broad group encompassing various subclasses. They were considered as less valuable than the *complex chemicals*, since they are either simple chemicals, not of high value for chemical research, or because they are not translatable to SMILES due to naming ambiguities. The following subclasses were all labeled as *simple chemicals*:

- Simple IUPAC names: Chemicals named following the IUPAC nomenclature, but not containing a number or element symbol. They can be translated to SMILES. "Carbon dioxide" and "Dinitrogen" are examples of simple IUPAC names
- Groups of compounds: Chemical families and groups of compounds are not directly translatable to SMILES because they do not define a single molecule but a group of molecules sharing certain similarities. "Alkyl" and "Halogen" are examples of groups of compounds.
- Trivial names: Trivial and colloquial names and brand names, that do not follow IUPAC conventions. They can sometimes be translated to SMILES. "Water" and "Aspirin" are examples of trivial names.
- Formulas: Formulas of molecules, giving a count of the atoms contained in a molecule using chemical element symbols. Formulas are not directly translatable to SMILES because they lack structural information. "HCl" and "NaCl" are examples of formulas.
- Polymers: All chemicals that contain the word "poly" are considered polymers. "Poly" stands for an undefined number of repetitions of a chemical structure. Due to this undefined number, polymers are not directly translatable to SMILES. "Polypropylene" and "Polyvinylchloride" are examples of polymers.
- Substructures: Chemical groups that end on -yl. They only describe a part of a molecule that is not in its complete form. Therefore, not the whole molecule is described and substructures are not directly translatable to SMILES. "Methyl" and "Ethyl" are examples of substructures.

3) Markush descriptions: The third category *Markush descriptions* was defined as sentences that contain Markush structures with placeholder letters [2]. The placeholder letters are used to represent variable components or side chains. This method allows for the description of a family of compounds in a single structural formula, where the placeholders can be substituted with various groups or elements. When a *Markush description* appears within a paragraph, the entire paragraph is labeled as such to ensure the full scope of

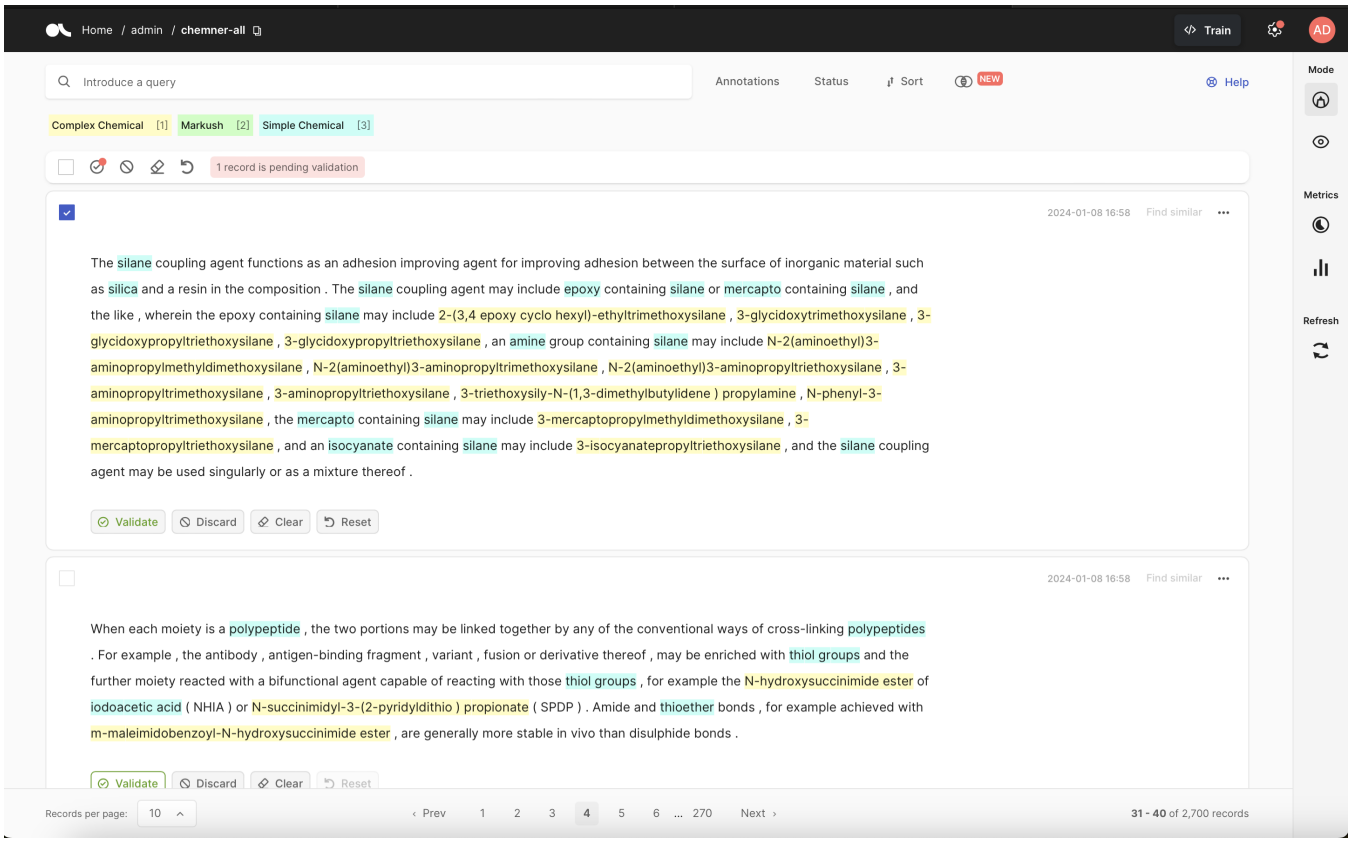

Figure 1: **Screenshot of annotation interface Argilla.** Argilla provides an intuitive user interface for annotating text. In this screenshot, examples of *simple chemicals* are highlighted in turquoise and examples of *complex chemicals* are highlighted in yellow.

potential molecules is accurately captured, including the necessary context. Due to the nested and variable properties of *Markush description*, a direct translation into the SMILES is not possible in most cases.

"R-SH where R is a substituted alkyl" and "$CH_2-CH_2-S(O)_2-R^{15-}CH(-OH)-R^{15-}S(O)_2-CH=CH_2$, where each R^{15} is independently selected from C_{1-3} alkanediyl, and substituted C_{1-3} alkanediyl, wherein the one or more substituent groups is —OH" are examples of *Markush descriptions*.

B. Annotation interface

We used the open-source tool Argilla to annotate our data. It features an intuitive user interface (see Fig. 1) and offers simple ways for importing or exporting raw and annotated data in multiple formats. Furthermore, it supports pre-annotations and is equipped with the string query syntax of Elasticsearch [20], facilitating paragraph search for specific strings or substrings.

C. Pre-annotations with large language model

Brown et al [19] have demonstrated that language models have formidable few-shot performance on NLP tasks without requiring the model to be fine-tuned on the specific task. We leveraged this few-shot learning capability of large language models to pre-annotate our corpus, thereby enhancing the efficiency and accuracy of our annotation process.

Two factors led to the decision to use the model Falcon 180B [21] for pre-annotating our data. First, the model is open-source, and the license allows for both scientific as well as commercial use. Second, according to Almazrouei et al. [21] the Falcon 180B outperforms other open-source models such as PaLM [22] or LLaMA 2 [23], while having a lower inference cost.

Due to the lack of data to identify the most effective prompting strategy, we opted for an approach that utilizes HTML-like tags, similar to the stratety proposed by Adams et al. [18]. Figure 2 shows an example prompt with the tagging scheme we developed for the pre-annotation process. This scheme was designed to effectively differentiate between input text and predictions, while also separating individual chemical entities, as detailed below.

- **Input text:** First, the input text was given to the model. We denoted the beginning of the text with a `<text>` tag and the end of the text with the `</text>` tag.

- **Chemical entities:** Second, we add a <chemical> tag to indicate that this will be the desired output. Each chemical mentioned in the text above was listed and between single entities a <sep> tag was placed. To show the end of the annotations, the </chemical> tag was used.
- **Entity prediction:** We included five correctly annotated example paragraphs before adding the paragraph where Falcon 180B should make the predictions. The patent text was added between the <text> and </text> tag, added the <chemical> to denote the start of the annotations and let the model predict all chemical entities.

```
<text>
The materials used to make compounds 1 -
4 shown in reaction scheme (2) were
obtained from commercial sources or made
using known organic methods. Potassium
acetate, dioxane, 1,1   -Bis (di-tert-
butylphosphino) ferrocene]
dichloropalladium (II) ((PdCl_2-(dppf)),
dppf, bis (pinacolato) diboron, 2-chloro
-4,6-diphenyl-1,3,5-triazine, 2,4-
dichloro-6-phenyl-1,3,5-triazine,
tetrakis (triphenylphosphine) palladium
(0), tetrahydrofuran (THF), potassium
carbonate, ethyl acetate, hexane,
magnesium sulfate (MgSO_4), and
chloroform were obtained from Sigma-
Aldrich  (USA).
</text>
<chemical>
Potassium acetate<sep>dioxane<sep>1,1   -
Bis (di-tert-butylphosphino) ferrocene]
dichloropalladium<sep>bis (pinacolato)
diboron<sep>2-chloro-4,6-diphenyl-1,3,5-
triazine<sep>2,4-dichloro-6-phenyl-1,3,5-
triazine<sep>tetrakis (triphenylphosphine
) palladium<sep>tetrahydrofuran<sep>
potassium carbonate<sep>ethyl acetate<sep
>hexane<sep>magnesium sulfate<sep>MgSO_4<
sep>chloroform
</chemical>
```

Figure 2: **HTML-like prompting scheme for generative models.** The <text> tag marks the beginning of an input paragraph and the </text> marks the end. The <chemical> tag indicates the start of the list containing of chemicals contained in the text and the </chemical> tag indicates the end. The <sep> tag was used to separate single chemical entities.

D. Data augmentation

To increase the volume and diversity of our training corpus and ensure a uniform distribution of number of chemical entities per paragraph, the corpus was augmented. This involved curating two datasets: one for *complex chemical* entities and the other for *simple chemical* entities, both sourced from publicly available chemical names from PubChem [24] adhering to the IUPAC naming conventions.

To prevent our models to be biased towards a certain number of chemical entities per paragraph, we aimed for a uniform distribution. Specifically, the dataset contains paragraphs with the number of entities per paragraph ranging from one to twenty. The augmentation process was developed as follows:

- **Inclusion of Original Paragraphs:** We began by incorporating all original, non-augmented paragraphs that matched the desired number of chemicals per paragraph into the augmented dataset.
- **Augmentation with Curated Entities:** To create augmented data, we selected random paragraphs that contained an equal or fewer number of chemicals than desired. Within these paragraphs, we replaced all chemical entities with randomly chosen entities from our curated PubChem datasets.
- **Ensuring uniform distribution:** If a selected paragraph had a lesser number of chemical entities than needed, we increased the count by selecting an existing chemical entity and substituting it with two randomly selected chemicals from the datasets. This method ensured that the placement of the chemical entities remains linguistically logical. Through this approach, it was guaranteed that each paragraph achieved the desired target number of chemical entities.

E. Model architectures and training

We trained and evaluated two types of models: a model based on the BERT architecture [4] and two models based on the T5 architecture [5]. Since the downstream processing of *Markush descriptions* has yet to be determined and a subsequent translation into a SMILES encoding would not be possible, we decided to exclude paragraphs containing *Markush descriptions* from the scope of this practical work.

For the BERT model, we opted for the DistilRoBERTa-base model, a distilled version of BERT, trained using the optimized RoBERTa pre-training approach [25]. The RoBERTa pre-training procedure mainly differentiates from the BERT training by allowing for a longer pre-training on more training data and larger training batches. DistilRoBERTa-base consists of 83M parameters.

For the T5 based models, we used the T5-small model with 61M parameters and the T5-base model with 223M parameters. The T5 model is an encoder-decoder transformer that closely follows the implementation introduced

Model	#Parameters	Simple Chemicals			Complex Chemicals			Overall		
		Recall	Precision	F_1-score	Recall	Precision	F_1-score	Recall	Precision	F_1-score
Falcon 180B	180B	-	-	-	-	-	-	27.5%	59.7%	37.8%
T5-small	61M	-	-	-	67.4%	43.0%	51.0%	-	-	-
T5-base	223M	-	-	-	70.1%	69.3%	67.6%	-	-	-
DistilRoBERTa	83M	65.9%	79.1%	71.9%	89.7%	96.3%	92.9%	74.3%	85.6%	79.5%

Table II: **Comparison of model performance.** We report the evaluation of state-of-the-art model architectures for CER. We calculate the recall, precision and F_1-score on the *simple chemical* label, the *complex chemical* label and on both labels. The number of parameters for each model is also provided.

by Vaswani et al. [12]. Input sequences get mapped to an embedding which is then passed into the encoder, consisting of self-attention layers and feed-forward networks. The decoder resembles the encoder with the addition of a standard attention mechanism that attends to the output of the encoder.

F. Implementation details

The implementation was carried out using PyTorch 2.1.0 and using the Adam optimizer [26]. For the DistilRoBERTa model, we applied a learning rate scheduler to linearly increase the learning rate from 10^{-10} to 10^{-5} during the first 1,000 training steps, then stay at 10^{-5} for 1,000 steps before exponentially decaying with an exponent of 10^{-3}. For the T5-small and T5-base model, we used 1,000 warmup steps to increase the learning rate from 10^{-10} to 10^{-4}, then stay at 10^{-4} for 1,000 steps before exponentially decaying with an exponent of 10^{-3}. The T5 models were both trained during an early stage of this project where the dataset only consisted of 1,200 annotated paragraphs and we did not yet include the *simple chemicals* in the annotation. Due to computational constraints, these models were not retrained on the whole dataset.

IV. Results

This section first inspects and evaluates the curated dataset for the CER task. Then, the evaluation metrics for the models get introduced. Finally, the performances of the pre-annotation, T5 models and DistilRoBERTa model are presented.

A. Dataset overview

We have created the first corpus specifically designed for the CER task of recognizing chemical entities following IUPAC nomenclature in patent documents. The corpus contains a total of 2,700 annotated paragraphs, resulting in 22,131 annotations in the dataset. The largest group of chemicals in the corpus are the *simple chemicals* with 66.5% (14,712 annotations) of all annotations, followed by the *complex chemicals* with 32.6% (7,219 annotations) and the third group consisting of *Markush descriptions* corresponds only to 0.9% (200 annotations) of the dataset.

When examining the sets of unique chemical names, the distribution becomes more even because the *simple chemicals* are more often repeated. In total, the corpus contains 15,044 unique chemicals, of which 50.6% are *simple chemicals* (7,625 entities), 48.0% are *complex chemicals* (7,219 entities) and 1.3% are *Markush descriptions* (200 entities). In the *simple chemical* group, the most frequently mentioned entity is water with 265 mentions, which is significantly more than the most frequently mentioned entity of the *complex chemical* group, N, N-dimethylformamide with 35 occurrences.

The augmented corpus contains a total of 110,000 paragraphs, where we split the corpus into a training set of 100,000 paragraphs and test set of 10,000 paragraphs. The data was split before the augmentation process to avoid data leakage. In the augmented corpus, we have 1,155,000 total annotations, of which 349,137 entities are *simple chemicals* and 805,863 entities are *complex chemicals*. The number of unique chemicals is 1,001,265, of which 245,205 entities are *simple chemicals* and 756,060 are *complex chemicals*.

B. Model evaluation metrics

We evaluate our models performance based on a span-based metric as described below.

We use a strict match strategy, where a prediction is only regarded as correct if all tokens inside an entity were correctly classified. If one or more tokens of an entity were misclassified, we regard the entire prediction as a false negative. Furthermore, if the model falsely predicted a non-chemical entities as chemical, we count it as false positive.

With these definitions for true positive, false negative and false positive we evaluate our models based on recall, precision and F_1-score.

C. Model performances

An overview of the performances of the different models can be seen in table II.

1) Pre-annotation results: The pre-annotations created with the Falcon 180B model resulted in a total of 9,330 predicted entities, of which 5,573 were correct. The prompting strategy we used aimed at detecting both *simple chemicals* and *complex chemicals* without differentiating them. Therefore, we only evaluate this model on its overall performance which yields a recall of 27.5%, precision of 59.7% and F_1-score of 32.8%.

2) T5 results: The T5-small model, which was only trained for predicting the group of *complex chemicals* achieved a recall of 67.4%, a precision of 43.0% and a F_1-score of 51.0%. The larger T5-base model, also only trained for predicting the group of *complex chemicals*, achieves improved results with a recall of 70.1%, a significantly increased precision of 69.3% and a F_1-score of 67.6%.

3) DistilRoBERTa results: The DistilRoBERTa model achieves a recall of 89.7%, a precision of 96.3% and a F_1-score of 92.9% for predicting the *complex chemicals*, and a recall of 65.9%, a precision of 79.1% and a F_1-score of 71.9% for the *simple chemicals*. On the overall CER task, DistilRoBERTa achieves a recall of 74.3%, a precision of 85.6% and a F_1-score of 79.5%.

V. Discussion

We have created a corpus consisting of 2,700 paragraphs originating from patent documents, annotated all chemical entities and categorized them in three groups. While the pre-annotations created with Falcon 180B facilitated the annotation proces, the relatively low recall of 27.5% underscored the need for a dedicated, fine-tuned model for the task.

Our models for CER in patent documents demonstrate a significant difference between generative and BERT models. Consistent with previous research, our best model based on the DistilRoBERTa architecture achieves a high F_1-score of 93% for the most important category of *complex chemicals*, thereby outperforming the T5-base model with a F_1-score of 68% and the T5-small with a F_1-score of 51%.

The T5-base model's improved performance compared to the T5-small model highlight the benefits of an increased number of parameters and their ability extract chemical information from patent documents.

The DistilRoBERTa model with recall, precision and F_1-score around 90% for recognizing *complex chemicals* aligns with state-of-the-art models on similar CER tasks. Additional to the superior performance of DistilRoBERTa compared to the T5-base model, the smaller model size with 83M parameters offers also a run-time advantage.

VI. Conclusion and future work

In this practical work, we created a corpus for chemical entity recognition in the domain of patent documents. A strategy for utilizing the generative model Falcon 180B for creating pre-annotations was developed, resulting in a facilitated annotation process. The corpus contains 2,700 annotated paragraphs, resulting in a comprehensive dataset with over 22,000 chemical entities.

Furthermore, we evaluated generative and models based on the BERT architecture and reached a F_1-score of 68% for recognizing *complex chemicals* with a T5-base model and a F_1-score of 93% with a DistilRoBERTa-base model. The finding that BERT models outperform generative models aligns with related work on similar chemical datasets.

The models developed in this practical work will complement the IBM Deep Search [27] document ingestion pipeline which currently only relies on images [28] to extract molecular structures. This opens up new opportunities for the joint analysis of visual and textual descriptions of chemical structures in patents. Future work could focus on the development of a translation model for translating chemical names into their SMILES encoding, making the extracted data more easily queryable.

References

[1] L. A. Currie, "Nomenclature in evaluation of analytical methods including detection and quantification capabilities (iupac recommendations 1995)," *Pure and Applied Chemistry*, 1995.

[2] T. Ebe, K. A. Sanderson, and P. S. Wilson, "The chemical abstracts service generic chemical (markush) structure storage and retrieval capability. 2. the marpat file," *Journal of Chemical Information and Computer Sciences*, 1991. [Online]. Available: https://pubs.acs.org/doi/abs/10.1021/ci00001a004

[3] D. Weininger, "Smiles, a chemical language and information system. 1. introduction to methodology and encoding rules," *J. Chem. Inf. Comput. Sci.*, vol. 28, 1988.

[4] J. D. et al., "Bert: Pre-training of deep bidirectional transformers for language understanding." Association for Computational Linguistics.

[5] C. R. et al., "Exploring the limits of transfer learning with a unified text-to-text transformer," *Journal of Machine Learning Research*.

[6] R. I. et al., "Nlm-chem, a new resource for chemical entity recognition in pubmed full text literature," *Nature Journal*.

[7] M. K. et al., "The chemdner corpus of chemicals and drugs and its annotation principles," *Journal of Cheminformatics*.

[8] C. K. et al., "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)*.

[9] D. Q. N. et al., "Chemu: Named entity recognition and event extraction of chemical reactions from patents", booktitle="advances in information retrieval." Springer International Publishing.

[10] S. H. et al., "Long short-term memory," *Neural computation*.

[11] J. L. et al., "Conditional random fields: Probabilistic models for segmenting and labeling sequence data."

[12] A. V. et al., "Attention is all you need." Curran Associates, Inc.

[13] J. C. et al., "Named entity recognition in chemical patents using ensemble of contextual language models." Proceedings of CLEF (Conference and Labs of the Evaluation Forum) 2020 Working Notes.

[14] T. R. et al., "Chemspot: a hybrid system for chemical named entity recognition," *Bioinformatics*.

[15] R. L. et al., "tmchem: a high performance approach for chemical named entity recognition and normalization," *Journal of Cheminformatics*.

[16] J. L. et al., "A hybrid deep-learning approach for complex biochemical named entity recognitionon," *Knowledge-Based Systems*.

[17] W. H. et al., "Lstmvoter: chemical named entity recognition using a conglomerate of sequence labeling tools," *Journal of Cheminformatics*.

[18] V. A. et al., "Chemical identification and indexing in pubmed articles via bert and text-to-text approaches," *Database*.

[19] T. B. B. et al., "Language models are few-shot learners," vol. 33. Curran Associates, Inc.

[20] Elastic, "elasticsearch/elasticsearch," 2015. [Online]. Available: https://github.com/elasticsearch/elasticsearch

[21] E. A. et al., "The falcon series of open language models," https://synthical.com/article/64bd055b-e67b-4f1a-bbaf-3205930c3eb4, 10 2023.

[22] A. C. et al., "Palm: Scaling language modeling with pathways," vol. 24, pp. 240:1–240:113, 2022. [Online]. Available: https://api.semanticscholar.org/CorpusID:247951931

[23] H. T. et al., "Llama 2: Open foundation and fine-tuned chat models."

[24] S. Kim, J. Chen, T. Cheng, A. Gindulyte, J. He, S. He, Q. Li, B. A. Shoemaker, P. A. Thiessen, B. Yu, L. Zaslavsky, J. Zhang, and E. E. Bolton, "PubChem 2023 update," *Nucleic Acids Research*, pp. D1373–D1380, 2022. [Online]. Available: https://doi.org/10.1093/nar/gkac956

[25] Y. L. et al., "Roberta: A robustly optimized bert pretraining approach," 2019, cite arxiv:1907.11692. [Online]. Available: http://arxiv.org/abs/1907.11692

[26] D. P. Kingma and J. Ba, "Adam: A method for stochastic optimization," 2015. [Online]. Available: http://arxiv.org/abs/1412.6980

[27] P. W. J. Staar *et al.*, "Corpus conversion service," in *Proceedings of the 24th ACM SIGKDD International Conference on Knowledge Discovery & Data Mining*. ACM, jul 2018. [Online]. Available: https://doi.org/10.1145\%2F3219819.3219834

[28] L. Morin, M. Danelljan, M. I. Agea, A. Nassar, V. Weber, I. Meijer, P. Staar, and F. Yu, "Molgrapher: Graph-based visual recognition of chemical structures," in *Proceedings of the IEEE/CVF International Conference on Computer Vision (ICCV)*.