ChemProp: A Dataset with Annotations for Instructional Language in Chemical Patents

Sopan Khosla[1,*], Carolyn Rose[2]

[1]AWS AI Labs
[2]Carnegie Mellon University

Abstract

In this paper, we propose a new set of annotations for the ChEMU Chemical Reaction Corpus. Our annotations (ChemProp) non-trivially incorporate the signals from ChEMU 2020 and 2021 schema to extract the instructional structure from chemical patents with details about inputs, outputs, and reaction attributes for each event in the reaction snippet. We propose a semi-automatic algorithm to create ChemProp and benchmark state-of-the-art models proposed for ChEMU 2020 and ChEMU 2021 on it. We hope that ChemProp can play an important part in modeling the instructional language present in chemical patents.

Keywords

Chemical Patents, Information Extraction, Program Synthesis, Coreference Resolution, Relation Extraction

1. Introduction

Chemical research relies heavily on the knowledge of chemical processes and synthesis, which are often described in chemical patents or research literature, with patents also serving as a critical source of information about new compounds [1]. Despite the significant value of the information present in these documents, extraction and organization of this information still heavily relies on costly manual processes [2]. High influx of such documents in chemistry has introduced the need for automatic systems that can extract the structured knowledge present in these texts [3, 4].

CLEF ChEMU shared-task series released the ChEMU Chemical Reaction Corpus that contains reaction snippets extracted from chemical patents. For ChEMU 2020 [3], the authors annotate information about relationships between reaction events (steps) and named-entities involved in that step. ChEMU 2021 [4] on the other hand focuses specifically on extracting chemical relations between a pair of entity-mentions. The framework introduces five domain-specific relations (including bridging and coreference) that link different noun-phrases present in the discourse. Finally, ChEMU 2022 [5] reused the expression-level tasks from 2020 and 2021, and also introduced other document-level information extraction tasks.

None of these shared tasks however fully capture the *entire* instructional structure (e.g. a chronological sequence of *inputs, reaction-steps, conditions,* and *outputs*) of the underlying chemical patent. For example, even though ChEMU 2020 schema tries to relate reaction events with associated compounds or conditions, it only operates on named-entities, and therefore does not cover important lexical items (noun-phrases) that describe relevant reaction conditions and participants using co-referring generic expressions, for example, *the mixture*, *the organic layer*, or *the filtrate*.

In this work, we propose an algorithm that augments CLEF ChEMU 2020 annotations with ChEMU 2021 annotations to create a more complete annotation framework for converting natural language chemical patents into structured recipes. Instructional language is a useful structure that comprises of step-by-step instructions that need to be performed to complete a task. However, most of the prior art in the instructional language paradigm focuses on cooking recipes. We propose a new dataset, **ChemProp**[1], that merges the ChEMU 2020 and 2021 annotations to create labels for the instructional language present in chemical patents. For each reaction snippet, we annotate constituting events (reaction/work-up steps), their relative chronological order, and entities that are associated with each of these events. More specifically, for each reaction step in that snippet, we annotate the trigger event verb, and the noun phrases (entities) that depict the (i) INPUT, (ii) OUTPUT, and (iii) reaction-attributes (RXN_ATTR) of that reaction step. We leverage the raw reaction snippets from the ChEMU Chemical Reaction Corpus as our data and annotate it by (i) automatically combining the annotations of CLEF ChEMU shared tasks 2020 and 2021, and (ii) manually incorporating events/entities that are missed by the two

[1]https://github.com/sopankhosla/chemprop

The Third AAAI Workshop on Scientific Document Understanding, February 14, 2023, Washington, DC

*Work done when the author was a student at CMU.
sopankh@amazon.com (S. Khosla); cprose@cs.cmu.edu (C. Rose)
https://sopankhosla.github.io/ (S. Khosla)

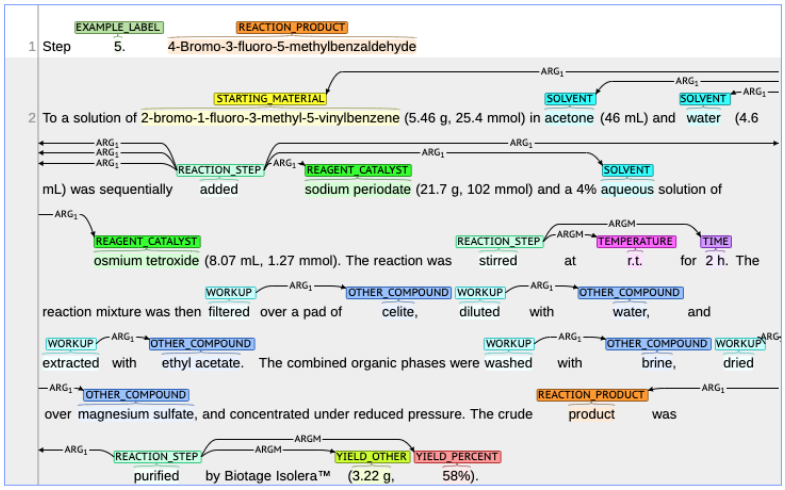

Figure 1: File 0050 with CC20 annotations.

annotation schemes.

Furthermore, we show the significance of these augmentations by evaluating the performance of the best performing models on ChEMU 2020 and 2021 shared-tasks on ChemProp. Our experiments show that models trained on ChemProp training data only achieve 0.69 Micro-F1 points on the test data, thus highlighting the room for improvement. We also show that ChemProp contains novel entities and relationships that are not present in ChEMU shared-tasks thus making it beneficial as a standalone benchmark for instructional language modeling from chemical patents.

2. Prior Art

In this section, we briefly describe the ChEMU Chemical Reaction Corpus and the two state-of-the-art annotation schemes proposed during ChEMU shared-tasks '20 & '21.

CLEF ChEMU 2020 Annotation Schema (CC20). He et al. [3] annotated a corpus of 1,500 patent snippets sampled from 170 patents from the European Patent Office and the United States Patent and Trademark Office. Their annotation schema aims at extraction of chemical reactions (i.e. REACTION_STEP, WORKUP) from patent snippets. It identifies trigger words that describe reaction steps and relates them to named-entities linked to the step (i.e. chemical compounds, time, temperature, and yields; Figure 1). Despite being a comprehensive annotation schema, CC20 suffers from two major drawbacks:

1. CC20 does not annotate reaction steps that do not relate to any named-entity in the discourse snippet. E.g., as shown in Figure 1, CC20 does not annotate the event *concentrated* (in line 7).
2. Furthermore, CC20 does not capture relationships between reaction steps (events) and noun-phrase mentions that denote combinations/ mixtures (e.g., the reaction mixture) or coreferent expressions (e.g., the product).

CLEF ChEMU 2021 Annotation Schema (CC21). Next year, He et al. [4] proposed an additional layer of annotation to the patents corpus, which focuses on the identification of anaphoric references. The new corpus contains annotations for both COREFERENCE and bridging relations (Figure 2). The authors define four domain-specific sub-types for bridging: TRANSFORMED, REACTION_ASSOCIATED, WORK_UP, CONTAINED. As a standalone schema, CC21 suffers from the following issues:

1. CC21 does not contain explicit information about reaction steps. Therefore, it is less useful, in isolation, for information extraction from chemical patents.
2. Furthermore, CC21 differs from CC20 on its definition of *mentions* and therefore makes the combination of two annotations non-trivial. In the next section, we describe the algorithm to handle these ambiguities.

3. ChemProp: Annotation

CC20 contains relationships between events and named-entities, whereas CC21 connects noun phrases (including named-entities) based on their anaphoric relationships. Together, CC20 and CC21 provide somewhat complementary information about each reaction snippet in the

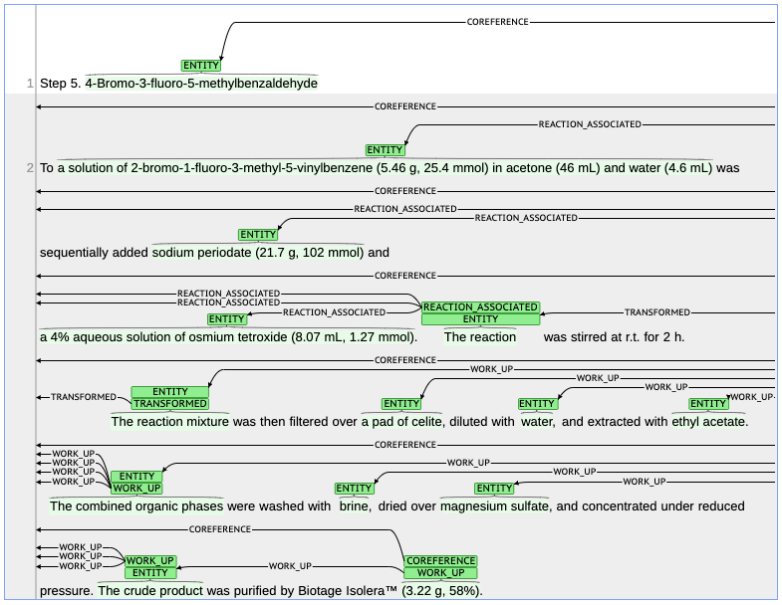

Figure 2: File 0050 with CC21 annotations.

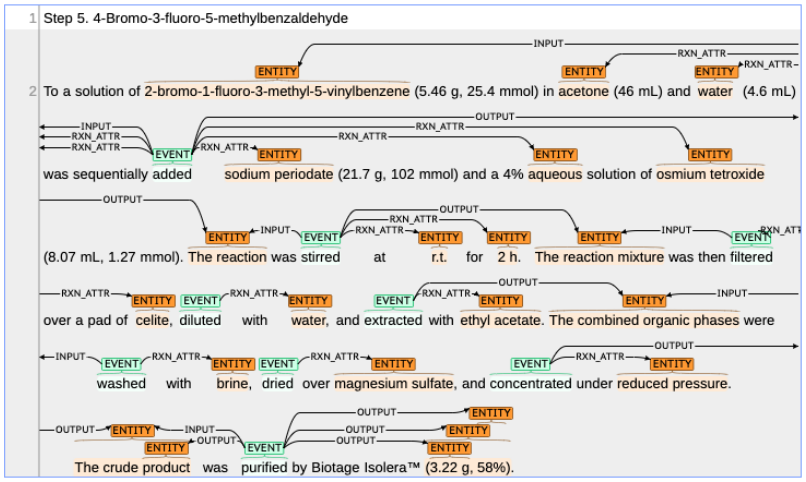

Figure 3: File 0050 with ChemProp annotations.

ChEMU corpus. In this section, we describe the steps we take to merge these somewhat heterogeneous annotation schema to create our new dataset ChemProp.

3.1. Automatic Merging of ChEMU 2020 and 2021

First, we present our algorithm that automatically merges signals from CC20 and CC21 and extracts lexical spans in the discourse, that most closely represent the inputs, outputs, & reaction-attributes for each reaction step.

3.1.1. Pre-processing

As a pre-processing step, we setup data-structures that help with the conversion algorithm. We create

1. A many-to-one map from CC20 named-entities to CC21 named-entity annotations to tackle the small annotation differences between the two schemas. The map is many-to-one because CC21 annotates entire solutions, whereas CC20 keeps the individual compounds/elements, e.g., [a solution of **ethanol**$_{CC20}$ and **water**$_{CC20}$]$_{CC21}$.

We also store the inverse one-to-many map from CC21 to CC20. So for the given example, we store

$$\underline{\text{ethanol}}_{CC20} \leftrightarrow \underline{\text{a solution of ethanol and water}}_{CC21}$$
$$\underline{\text{water}}_{CC20} \leftrightarrow \underline{\text{a solution of ethanol and water}}_{CC21}.$$

2. The mapping from ARG1 of CC21 reaction-oriented relations (i.e., REACTION_ASSOCIATED (R_ASSOC), WORK_UP, and TRANSFORMED (TRANS)) to their corresponding ARG2 mentions. For example, for the relation TRANSFORMED between *The reaction mixture* and *The reaction* in Figure 2 (line 4, 5), we store

$$\underline{\text{The reaction mixture}}_{CC21} \xrightarrow{\text{Rxn:TRANS}} \underline{\text{The reaction}}_{CC21}.$$

3. The mapping from CC20 events (ARG1) to their corresponding argument mentions (ARG2). For example, for the event *stirred* in Figure 1 (line 4), we store

$$\underline{\text{stirred}}_{CC20} \xleftrightarrow{\text{Event:TEMP}} \underline{\text{r.t.}}_{CC20}$$
$$\underline{\text{stirred}}_{CC20} \xleftrightarrow{\text{Event:TIME}} \underline{\text{2 h}}_{CC20}.$$

4. Finally, a dictionary to store the COREFERENCE relationships between different CC21 mentions. We create a separate map for coreference as they denote that both mentions in the pair are equivalent i.e., point to the same underlying entity in the discourse:

$$\text{ARG1}_{CC21} \xleftrightarrow{\text{Coref}} \text{ARG2}_{CC21}.$$

The above mappings store the relationships annotated in CC20 and CC21 that are relevant to creating ChemProp.

3.1.2. Algorithm

We allow for three relationships between reaction events and entities in our new ChemProp benchmark – INPUT, OUTPUT, RXN_ATTR. CC20 already annotates RXN_ATTRs (reaction-attributes) and named-entity INPUTS/ OUTPUTS. Therefore, to complete the schema, we devise an algorithm that can find a mapping between each non named-entity noun-phrase in CC21 to a reaction event in CC20 ($\text{ARG}_{CC21} \leftrightarrow \text{Rxn}_{CC20}$).

For each CC21 entity (ARG1_{CC21}) that is related to other CC21 entities (ARG2_{CC21}) (that occur before it in the discourse) via one of the three reaction-oriented CC21 relations (Rxn_{CC21}) i.e., REACTION_ASSOCIATED, WORK_UP, and TRANSFORMED, two possibilities need to be considered. For each ARG2^i_{CC21} in ARG2_{CC21}:

1. If ARG2^i_{CC21} is also present in the CC20 annotation, we use this ARG2^i_{CC21} as a pivot to combine the two annotations. We extract the CC20 relation (Rxn^i_{CC20}) it is linked with.
2. If ARG2^i_{CC21} is not present in CC20, it means that ARG2^i_{CC21} is a non named-entity noun-phrase and therefore needs to be resolved further. To ground such cases, we rely on the fact that the starting compounds in each reaction snippet starts are named-entities. Therefore, we can assume that in order to reach current ARG1_{CC21}, all other ARG2_{CC21}s that also appeared as ARG1_{CC21} (noun-phrases) have been resolved in earlier iterations of this algorithm. Therefore, the latest Rxn^i_{CC20} in the patent snippet between the CC20 reaction event associated with ARG2^i_{CC21} (i.e. $\text{ARG2}^i_{CC21} \rightarrow \text{Rxn}_{CC20}$) and ARG1_{CC21} is returned.

From this list of relations, we consider the Rxn^i_{CC20} that is closest to ARG1_{CC21}, but occurs before it, to be the *lexical event trigger* that outputs ARG1_{CC21}:

$$\text{Cand\_Rxn}_{CC20} = [\text{Rxn}^1_{CC20}, \text{Rxn}^2_{CC20}, ..., \text{Rxn}^j_{CC20}, ...]$$
$$\text{Rxn}^j_{CC20} = \textit{closest\_before}(\text{ARG1}, \text{Cand\_Rxn}_{CC20})$$
$$\text{ARG1}_{CC21} \xleftrightarrow{\text{Output:Rxn}} \text{Rxn}^j_{CC20}.$$

Consider the case where ARG1_{CC21} is *The combined organic phases* (Figure 2; line 6). It is related to four mentions (ARG2_{CC21}s) *{The reaction mixture, a pad of celite, water, ethyl acetate}* by the relation WORK_UP (Rxn_{CC21}). Three of these mentions *{a pad of celite, water, ethyl acetate}* are also present in the mapping created in preprocessing step 1, whereas *{The reaction mixture}* is not. As discussed earlier, for ARG2^i_{CC21}s present in CC20, we first create a list of Rxn_{CC20}s (Cand_Rxn_{CC20}) they correspond to *{filtered, diluted, extracted}* (Figure 1). For *The reaction mixture*, the corresponding Rxn_{CC20}, based on its resolution in the previous step, is *The reaction mixture* $\leftrightarrow$ *stirred*.

The latest CC20 relation between *stirred* and ARG2_{CC21}, *extracted*, is then inserted into Cand_Rxn_{CC20}. Combining the four relations we get *{extracted, filtered, diluted, extracted}*. From these, *extracted* occurs closest to the current ARG1 while occurring before it. Therefore, *The combined organic phases* is considered to be the OUTPUT of *extracted* (*The combined organic phases* $\leftrightarrow$ *extracted*; Figure 3).

Exceptional Cases. Although, most of the events can be fully annotated in ChemProp format using the above steps, there are certain exceptions that arise due to the mismatch between the motivation of CC20 and CC21 schemas.

1. (E1) For a small number of cases, we find that the reaction-event (Rxn^{j}_{CC20}) that occurs right before an ARG1_{CC21} might not be the event that outputs it.
   Example. Consider the phrase – "***the reaction mixture*** *is* ***filtered*** *in* **Celite**, and **ethanol** is ***added*** *to* ***the filtrate***". In this case, ***the filtrate*** refers to a state before the addition of ***ethanol*** and is the output of event ***filtered***. We use regex expressions to find such template patterns and resolve them automatically. We tackle the more complex occurrences in the manual quality assurance phase (as described in the next section).
2. (E2) In some instances, we find that ARG2^{i}_{CC21} and ARG1_{CC21} are related to each other by a Rxn_{CC21} but no corresponding Rxn_{CC20} is present. In such cases, we introduce a *pseudo* relation in between them and leave its annotation to the manual step.
   Example. Consider the phrase "***the reaction mixture*** *is* ***filtered***, *and* ***the filtrate*** *is* ***heated*** *for* ***20 min***". For this phrase, ***filtered*** would not be annotated in CC20 (no related named-entity), however, CC21 would annotate the pair (***the filtrate***, ***the reaction mixture***) as TRANSFORMED. In this case, we automatically introduce a *pseudo* relation whose input is ***the reaction mixture*** and output is ***the filtrate***.

The patent snippets are an ordered sequence of event steps that transform a starting product to an end product. Therefore, one can, with sufficient confidence, also consider the *outputs* of a particular event to be the best lexical representation of the *inputs* of the immediate next event. This allows us to annotate both **INPUT**s and **OUTPUT**s of the CC20 reaction-steps using a single algorithm. Finally, we add the **RXN_ATTR** (reaction-attribute) annotations present in CC20 on top to get our final dataset, ChemProp.

We note that although, we only consider three types of relations, where each relation is between an EVENT mention and an ENTITY mention (Figure 3), the fine-grained classification of these two types of mentions provided in CC20 can be easily ported over to ChemProp to make the new annotation schema more informative.

3.2. Manual Quality Assurance

Next, we manually go through the development and test data to fix the *exceptions* (described in the previous section).

As discussed earlier, the CC20 annotation does not annotate reaction steps that do not relate to any named-entity in the snippet. However, in our case such events are equally relevant and need to be extracted to get a complete picture of the reaction snippet. In order to annotate such events, we manually go through all of the development and test files and fix the annotations manually thus arriving at a gold dev and gold test sets. However we do not perform this quality assurance step on the training data due to it's large size, and therefore only obtain a silver training set. We find, however, that cases which require a manual inspection occur rather infrequently, and therefore do not deteriorate the quality of our training data much (sterling silver). Figure 3 shows an example patent snippet from the development set.

4. ChemProp: Baseline

In order to setup a baseline for ChemProp, we train the pipeline-based system from Dutt et al. [6] (also referred to as CC21_BEST going forward) on ChemProp training set. We refer the reader to the paper for more details about the system. We show the performance of two setups as described in Dutt et al. [6], (i) relation-classification on gold-entities and (ii) end-to-end classification.

We also evaluate the CC20 ground-truth [3] and the best-performing model at ChEMU shared-task 2020 [7] on ChemProp. As discussed earlier, the motivation of ChemProp is very similar to CLEF ChEMU 2020. However, in the absence of the support from CC21, the CC20 annotation does not capture all the relationships that make up the instructional language present in patent text. Hence, comparing CC20 against ChemProp would allow us to quantify the additional information present in ChemProp.

4.1. Results

We use the BRAT evaluation script distributed by CLEF ChEMU 2021 shared-task organizers to evaluate different setups. We find that CC20 ground-truth test data achieves an F1 score of 0.74 (Table 1). While the precision is near perfect, CC20 suffers from low recall on INPUT and OUTPUT relations as the annotation excludes some reaction events and does not annotate generic noun-phrases. Furthermore, CC20_BEST [7], a model designed for CC20 shared-task, gets an F1 score of 0.62, 12% below CC20 (ground-truth). These low numbers suggest that ChemProp provides considerably more information about the chemical patent snippets that will be systematically missed by the systems trained on CC20.

We observe that CC21_BEST (gold entities) achieves an F1 score of 0.86 on ChemProp test set. This suggests that the model is able to somewhat reliably figure out which named-entities/ noun phrases are related to which reaction event in the snippet. CC21_BEST (end-to-end), in addition to relation classification, also extracts mentions from raw patent snippets, and therefore expectedly per-

System	Micro F1
CC20 (ground-truth)	0.74
CC20_BEST	0.62
Trained on ChemProp	
CC21_BEST (gold-entities)	0.86
CC21_BEST (end-to-end)	0.69

Table 1
Performance on ChemProp test set

forms much worse than CC21_BEST (gold entities) with an overall F1 score of 0.69.

5. Conclusion

In this work, we propose a new corpus ChemProp that non-trivially combines properties from ChEMU 2020 and 2021 annotation schema to extract the instructional structure from chemical patents. We provide a semi-automatic algorithm to create ChemProp. Evaluating state-of-the-art models on the our new dataset suggests that there is still room for improvement in extracting relevant instructional triggers from patent text. We believe that ChemProp can act as an important benchmark for instructional language modeling.

References

[1] S. Senger, L. Bartek, G. Papadatos, A. Gaulton, Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents, Journal of cheminformatics 7 (2015) 1–12.

[2] S. Muresan, P. Petrov, C. Southan, M. J. Kjellberg, T. Kogej, C. Tyrchan, P. Varkonyi, P. H. Xie, Making every sar point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data, Drug Discovery Today 16 (2011) 1019–1030.

[3] J. He, D. Q. Nguyen, S. A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, Z. Afzal, Z. Zhai, B. Fang, H. Yoshikawa, et al., Overview of chemu 2020: named entity recognition and event extraction of chemical reactions from patents, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2020, pp. 237–254.

[4] J. He, B. Fang, H. Yoshikawa, Y. Li, S. A. Akhondi, C. Druckenbrodt, C. Thorne, Z. Afzal, Z. Zhai, L. Cavedon, et al., Chemu 2021: reaction reference resolution and anaphora resolution in chemical patents, in: ECIR (2), 2021.

[5] Y. Li, B. Fang, J. He, H. Yoshikawa, S. A. Akhondi, C. Druckenbrodt, C. Thorne, Z. Afzal, Z. Zhai, T. Baldwin, et al., Overview of chemu 2022 evaluation campaign: Information extraction in chemical patents, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2022, pp. 521–540.

[6] R. Dutt, S. Khosla, C. P. Rosé, A pipelined approach to anaphora resolution in chemical patents., in: CLEF (Working Notes), 2021, pp. 710–719.

[7] J. W. Y. R. Z. Zhang, Y. Zhang, Melaxtech: A report for clef 2020–chemu task of chemical reaction extraction from patent (2020).