ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents

Dat Quoc Nguyen[1,3], Zenan Zhai[1], Hiyori Yoshikawa[1,4], Biaoyan Fang[1], Christian Druckenbrodt[2], Camilo Thorne[2], Ralph Hoessel[2], Saber A. Akhondi[2], Trevor Cohn[1], Timothy Baldwin[1], Karin Verspoor[1⋆]

[1]The University of Melbourne, Australia; [2]Elsevier; [3]VinAI Research, Vietnam; [4]Fujitsu Laboratories Ltd., Japan
v.datnq9@vinai.io; {zenan.zhai,biaoyanf,hiyori.yoshikawa}@unimelb.edu.au
{c.druckenbrodt,c.thorne.1,r.hoessel,s.akhondi}@elsevier.com
{trevor.cohn,tbaldwin,karin.verspoor}@unimelb.edu.au

Abstract. We introduce a new evaluation lab named ChEMU (Cheminformatics Elsevier Melbourne University), part of the 11th Conference and Labs of the Evaluation Forum (CLEF-2020). ChEMU involves two key information extraction tasks over chemical reactions from patents. Task 1—Named entity recognition—involves identifying chemical compounds as well as their types in context, i.e., to assign the label of a chemical compound according to the role which the compound plays within a chemical reaction. Task 2—Event extraction over chemical reactions—involves event trigger detection and argument recognition. We briefly present the motivations and goals of the ChEMU tasks, as well as resources and evaluation methodology.

Keywords: Named entity recognition; Event extraction; Chemical reactions; Patents.

1 Introduction

The chemical industry undoubtedly depends on the discovery of new chemical compounds. However, new chemical compounds are often initially disclosed in patent documents, and only a small fraction of these compounds are published in journals, usually taking an additional 1-3 years after the patent [13]. Therefore, most chemical compounds are only available through patent documents [3]. In addition, chemical patent documents contain unique information, such as reactions, experimental conditions, mode of action, which is essential for the understanding of compound prior art, providing a means for novelty checking and validation as well as pointers for chemical research in both academia and industry [1,2]. As the number of new chemical patent applications has been drastically increasing [11], it is becoming crucial to develop natural language processing (NLP) approaches that enable automatic extraction of key information from the chemical patents [2].

⋆ Correspondence to Karin Verspoor, karin.verspoor@unimelb.edu.au.

In this paper, we propose a new evaluation lab (called ChEMU) focusing on information extraction over chemical reactions from patents. In particular, we will focus on two key information extraction tasks of *chemical named entity recognition* (NER) and *chemical reaction event extraction*. While previous related shared tasks focusing on chemicals or drugs such as CHEMDNER [7] have also included chemical named entity recognition as a task, those have primarily focused on PubMed abstracts. The CHEMDNER patents task [8] was limited to entity mentions and chemical entity passage detection, and only considered titles and abstracts of patents. For our ChEMU lab, we extend the existing corpora in several directions: first, we go beyond chemical NER to require labeling of the role of a chemical with respect to a reaction, and to consider complete chemical reactions in addition to entities. The ChEMU website is available at: `https://chemu-patent-ie.github.io`.

2 Goals and Importance

What are the goals of this evaluation lab? Our goals are: (1) To develop tasks that impact chemical research in both academia and industry, (2) To provide the community with a new dataset of chemical entities, enriched with relational links between chemical event triggers and arguments, and (3) To advance the state-of-the-art in information extraction over chemical patents.

Why is this lab needed? For evaluating information extraction developments in the scientific literature domain, there have been a large number of labs/shared tasks offered within previous i2b2/n2c2, SemEval, BioNLP, BioCreative, TREC and CLEF workshops. However, less attention has been paid to the chemical patent domain. In particular, there has previously been only one shared task on this domain, which is the CHEMDNER patents task at the BioCreative V workshop, involving detection of mentions of chemical compounds and genes/proteins in patent text [8].

Information extraction approaches developed for the scientific literature domain may not apply directly to the chemical patent domain. This is because as legal documents, patents are written very differently as compared to scientific literature. When writing scientific papers, authors strive to make their words as clear and straightforward as possible, whereas patent authors often seek to protect their knowledge from being fully disclosed [15]. In tension with this is the need to claim broad scope for intellectual property reasons, and hence patents typically contain more details and are more exhaustive than scientific papers [9].

There are also a number of characteristics of patent texts that create challenges for NLP in this context. Long sentences listing names of compounds are frequently used in chemical patents. The structure of sentences in patent claims is usually complex, and syntactic parsing in patents can be difficult [4]. A quantitative analysis from [16] showed that the average sentence length in a patent corpus is much longer than in general language use. That work also showed that the lexicon used in patents usually includes domain-specific and novel terms that are difficult to understand.

Table 1. Brief definitions of ChEMU chemical entity types, organised into chemical entity types, a reaction label introduced in the text, and reaction properties.

Entity type	**Definition**
REACTION_PRODUCT	A product is a substance that is formed during a chemical reaction.
STARTING_MATERIAL	A substance that is consumed in the course of a chemical reaction providing atoms to products is considered as starting material.
REAGENT_CATALYST	A reagent is a compound added to a system to cause or help with a chemical reaction. Compounds like catalysts, bases to remove protons or acids to add protons must be also annotated with this tag
SOLVENT	A solvent is a chemical entity that dissolves a solute resulting in a solution.
OTHER_COMPOUND	Other chemical compounds that are not the products, starting materials, reagents, catalysts and solvents.
EXAMPLE_LABEL	A label associated with a reaction specification.
TEMPERATURE	The temperature at which the reaction was carried out must be annotated with this tag.
TIME	The reaction time of the reaction.
YIELD_PERCENT	Yield given in percent values.
YIELD_OTHER	Yields provided in other units than %.

How will the community benefit from the lab? The ChEMU lab will provide a new challenging set of tasks, in an area of significant pharmacological importance. The lab will focus attention on more complex analysis of chemical patents, provide strong baselines as well as providing a useful resource for future research.

What are usage scenarios? Automatically identifying compounds which serve as the starting material or are a product of a chemical reaction would allow more targeted extraction of chemical information from patents and can improve the usefulness of patent resources. Automatic extraction of chemical reaction events supports the construction of cheminformatics databases, capturing key information about chemicals and how they are produced, from the patent resources.

3 Tasks

The ChEMU lab at CLEF-2020[1] offers the two information extraction tasks of Named entity recognition (**Task 1**) and Event extraction (**Task 2**) over chemical reactions from patent documents. Teams may participate in one or both tasks.

[1] `https://clef2020.clef-initiative.eu`

Table 2. An example of a chemical reaction snippet and BRAT annotations in a standoff format [14] w.r.t. Task 1.

(step 1) Synthesis of 3-chloro-6-(trifluoromethyl)pyridazine
To 3-(trifluoromethyl)-1H-pyridazin-6-one (1.1 g, 6.7 mmol) was added phosphorus oxychloride (10 mL) and the mixture was stirred at 100°C for 2.5 hr, and concentrated under reduced pressure. To the obtained residue were added dichloromethane and water, and the mixture was stirred at room temperature for 5 min. After stirring, the mixture was alkalified with potassium carbonate and partitioned. The organic layer was washed with saturated brine, dried over sodium sulfate and the desiccant was filtered off. The solvent was evaporated and the obtained residue was purified by silica gel column chromatography (petroleum ether/ethyl acetate) to give the title compound (0.77 g, 4.2 mmol, 63%).

ID	Entity type	Offsets		Text span
T1	EXAMPLE_LABEL	6	7	1
T2	REACTION_PRODUCT	22	60	3-chloro-6-(trifluoromethyl)pyridazine
T3	STARTING_MATERIAL	64	102	3-(trifluoromethyl)-1H-pyridazin-6-one
T4	REAGENT_CATALYST	131	153	phosphorus oxychloride
T5	TEMPERATURE	193	203	100°C
T6	TIME	208	214	2.5 hr
T7	SOLVENT	292	307	dichloromethane
T8	SOLVENT	312	317	water
T9	TEMPERATURE	350	366	room temperature
T10	TIME	371	376	5 min
T11	OTHER_COMPOUND	426	445	potassium carbonate
T12	OTHER_COMPOUND	507	512	brine
T13	OTHER_COMPOUND	525	539	sodium sulfate
T14	OTHER_COMPOUND	678	693	petroleum ether
T15	OTHER_COMPOUND	694	707	ethyl acetate
T16	REACTION_PRODUCT	721	735	title compound
T17	YIELD_OTHER	737	743	0.77 g
T18	YIELD_OTHER	745	753	4.2 mmol
T19	YIELD_PERCENT	755	758	63%

3.1 Task 1: Named entity recognition

In general, a chemical reaction is a process leading to the transformation of one set of chemical substances to another [10]. Task 1 involves identifying chemical compounds and their specific types, i.e. to assign the label of a chemical compound according to the role which it plays within a chemical reaction. In addition to chemical compounds, this task also requires identification of the temperatures and reaction times at which the chemical reaction is carried out, as well as yields obtained for the final chemical product and the label of the reaction.

This task involves both entity boundary prediction and entity label classification. We define 10 different entity type labels as shown in Table 1. See examples of those entity types in Table 2.

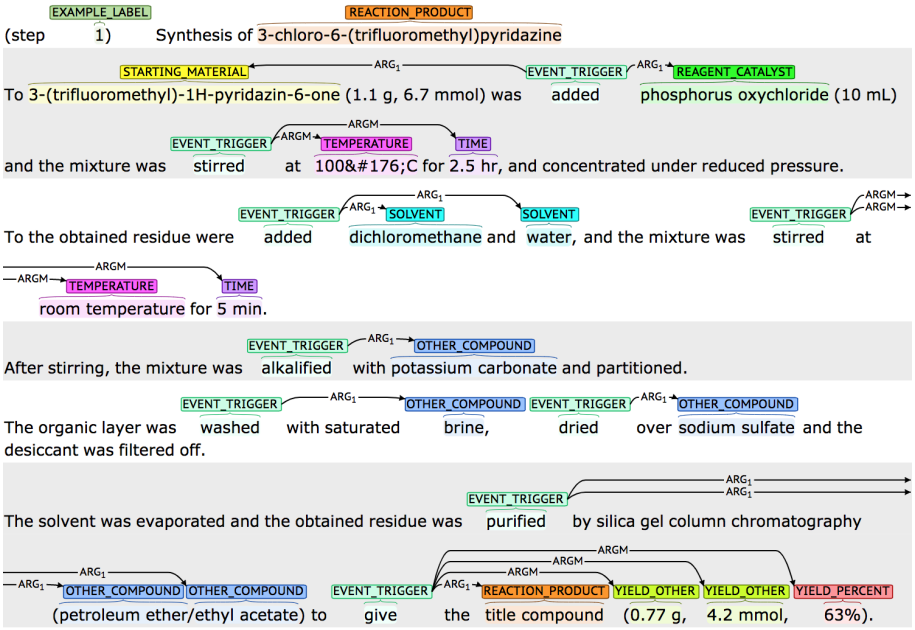

Fig. 1. BRAT visualization of chemical reaction events.

3.2 Task 2: Event extraction

As illustrated in Figures 1 and 2, a chemical reaction leading to an end product often consists of a sequence of individual event steps. Task 2 is to identify those steps which involve chemical entities recognized from Task 1. Unlike a conventional event extraction problem [6] which involves event trigger word detection, event typing and argument prediction, our Task 2 requires identification of event trigger words (e.g. "added" and "stirred") which all have the same type of "EVENT_TRIGGER", and then determination of the chemical entity arguments of these events.[2]

When predicting event arguments, we adapt semantic argument role labels **Arg1** and **ArgM** from the Proposition Bank [12] to label the relations between the trigger words and the chemical entities: **Arg1** is used to label the relation between an event trigger word and a chemical compound. Here, **Arg1** represents argument roles of being causally affected by another participant in the event [5]. **ArgM** represents adjunct roles with respect to an event, used to label the relation between a trigger word and a temperature, time or yield entity.

An end-to-end process incorporating both Task 1 and Task 2 can be equivalently viewed as a relation extraction task which identifies 11 entity types

[2] Note that those individual event steps are sequentially ordered, thus we do not consider cases where an event is an argument of another event, i.e. we do not label the relationship between two event triggers.

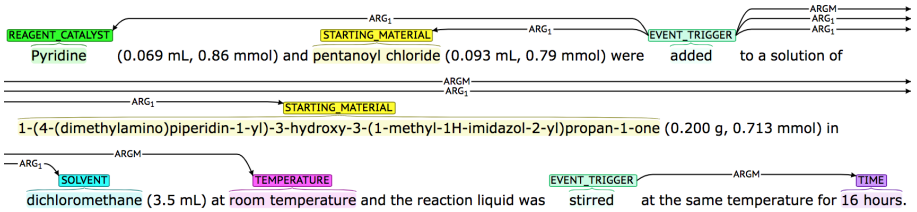

Fig. 2. BRAT visualization of a more complex event with the trigger word "added" involving five arguments.

including 10 types defined in Table 1 plus "EVENT_TRIGGER", and extracts relations between the "EVENT_TRIGGER" entities and the remaining entities.

4 Data and Evaluation

Data: For system development and evaluation, a new corpus of 1500 chemical reaction snippets will be provided for both tasks (an example of a chemical reaction snippet is shown in Table 2). These snippets are sampled from 170 English document patents from the European Patent Office and the United States Patent and Trademark Office. We will mark up every chemical compound or event trigger with both text spans and IDs, and highlight relations and event arguments, as illustrated in Figures 1 and 2. We have begun preparing the corpus and will make available strong baselines for the tasks. Initial publications related to the data and Task 1 appear at the 2019 ALTA and BioNLP workshops, respectively [18,19].

The corpus will be split into 70%/10%/20% training/development/test. Gold annotations for the training and development sets will be provided to task participants in the BRAT standoff format [14] during the development phase. The raw test set will be provided for final test phase.

To support teams who are interested in Task 2 only, a pre-trained chemical NER tagger is provided as a resource [19].

Evaluation: For evaluation, precision, recall and F1 scores will be used, under both strict and relaxed span matching conditions. F1 will be the main metric for ranking the participating teams [17].[3]

5 Conclusion

In this paper, we have presented a brief description of the upcoming ChEMU lab at CLEF-2020. ChEMU will focus on two new tasks of named entity recognition and event extraction over chemical reactions from patents. We expect participants from both academia and industry. We will advertise our ChEMU lab via social media as well as NLP-related mailing lists.

[3] `https://bitbucket.org/nicta_biomed/brateval/src/master/`

Acknowledgments

This work is supported by an Australian Research Council Linkage Project, LP160101469, and Elsevier.

References

1. Akhondi, S.A., Klenner, A.G., Tyrchan, C., Manchala, A.K., Boppana, K., Lowe, D., Zimmermann, M., Jagarlapudi, S.A.R.P., Sayle, R., Kors, J.A., Muresan, S.: Annotated Chemical Patent Corpus: A Gold Standard for Text Mining. PLOS ONE **9**, 1–8 (09 2014)
2. Akhondi, S.A., Rey, H., Schwörer, M., Maier, M., Toomey, J., Nau, H., Ilchmann, G., Sheehan, M., Irmer, M., Bobach, C., Doornenbal, M., Gregory, M., Kors, J.A.: Automatic identification of relevant chemical compounds from patents. Database **2019**, baz001 (2019)
3. Bregonje, M.: Patents: A unique source for scientific technical information in chemistry related industry? World Patent Information **27**(4), 309–315 (2005)
4. Hu, M., Cinciruk, D., Walsh, J.M.: Improving Automated Patent Claim Parsing: Dataset, System, and Experiments. CoRR **abs/1605.01744** (2016)
5. Jurafsky, D., Martin, J.H.: Speech and language processing, 3rd edition, chap. Semantic Role Labeling and Argument Structure (2019)
6. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP'09 shared task on event extraction. In: Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task. pp. 1–9 (2009)
7. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: CHEMDNER: The drugs and chemical names extraction challenge. Journal of Cheminformatics **7**(1), S1 (2015)
8. Krallinger, M., Rabal, O., Lourenço, A., Perez, M.P., Rodriguez, G.P., Vazquez, M., Leitner, F., Oyarzabal, J., Valencia, A.: Overview of the CHEMDNER patents task. In: Proceedings of the Fifth BioCreative challenge evaluation workshop. pp. 63–75 (2015)
9. Lupu, M., Mayer, K., Tait, J., Trippe, A.J.: Current Challenges in Patent Information Retrieval. Springer Publishing Company, Incorporated, 1st edn. (2011)
10. Muller, P.: Glossary of terms used in physical organic chemistry (IUPAC Recommendations 1994). Pure and Applied Chemistry **66**(5), 1077–1184 (2009)
11. Muresan, S., Petrov, P., Southan, C., Kjellberg, M.J., Kogej, T., Tyrchan, C., Varkonyi, P., Xie, P.H.: Making every SAR point count: The development of Chemistry Connect for the large-scale integration of structure and bioactivity data. Drug Discovery Today **16**(23), 1019–1030 (2011)
12. Palmer, M., Gildea, D., Kingsbury, P.: The Proposition Bank: An Annotated Corpus of Semantic Roles. Computational Linguistics **31**(1), 71–106 (2005)
13. Senger, S., Bartek, L., Papadatos, G., Gaulton, A.: Managing expectations: Assessment of chemistry databases generated by automated extraction of chemical structures from patents. Journal of Cheminformatics **7**, 49:1–49:12 (2015)
14. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: brat: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations Session at EACL 2012 (2012)
15. Valentinuzzi, M.E.: Patents and Scientific Papers: Quite Different Concepts: The Reward Is Found in Giving, Not In Keeping [Retrospectroscope]. IEEE Pulse **8**(1), 49–53 (2017)

16. Verberne, S., D'hondt, E., Oostdijk, N., Koster, C.: Quantifying the challenges in parsing patent claims. In: Proceedings of the 1st International Workshop on Advances in Patent Information Retrieval at ECIR 2010. pp. 14–21 (2010)
17. Verspoor, K., Jimeno Yepes, A., Cavedon, L., McIntosh, T., Herten-Crabb, A., Thomas, Z., Plazzer, J.P.: Annotating the biomedical literature for the human variome. Database **2013**, bat019 (2013)
18. Yoshikawa, H., Nguyen, D.Q., Zhai, Z., Druckenbrodt, C., Thorne, C., Akhondi, S.A., Baldwin, T., Verspoor, K.: Detecting Chemical Reactions in Patents. In: Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association. pp. 100–110 (2019)
19. Zhai, Z., Nguyen, D.Q., Akhondi, S., Thorne, C., Druckenbrodt, C., Cohn, T., Gregory, M., Verspoor, K.: Improving Chemical Named Entity Recognition in Patents with Contextualized Word Embeddings. In: Proceedings of the 18th BioNLP Workshop. pp. 328–338 (2019)