The ChEMU 2022 Evaluation Campaign: Information Extraction in Chemical Patents

Yuan Li[1], Biaoyan Fang[1], Jiayuan He[1,2], Hiyori Yoshikawa[1,3], Saber A. Akhondi[4], Christian Druckenbrodt[5], Camilo Thorne[5], Zenan Zhai[1], Zubair Afzal[4], Trevor Cohn[1], Timothy Baldwin[1], and Karin Verspoor[1,2](✉)

[1] The University of Melbourne, Melbourne, Australia
karin.verspoor@rmit.edu.au
[2] RMIT University, Melbourne, Australia
[3] Fujitsu Limited, Kawasaki, Japan
[4] Elsevier BV, Amsterdam, The Netherlands
[5] Elsevier Information Systems GmbH, Frankfurt, Germany

Abstract. The discovery of new chemical compounds is a key driver of the chemistry and pharmaceutical industries, and many other industrial sectors. Patents serve as a critical source of information about new chemical compounds. The ChEMU (Cheminformatics Elsevier Melbourne Universities) lab addresses information extraction over chemical patents and aims to advance the state of the art on this topic. ChEMU lab 2022, as part of the 13th Conference and Labs of the Evaluation Forum (CLEF-2022), will be the third ChEMU lab. The ChEMU 2020 lab provided two information extraction tasks, named entity recognition and event extraction. The ChEMU 2021 lab introduced two more tasks, chemical reaction reference resolution and anaphora resolution. For ChEMU 2022, we plan to re-run all the four tasks with a new task on semantic classification for tables as the fifth one. In this paper, we introduce ChEMU 2022, including its motivation, goals, tasks, resources, and evaluation framework.

Keywords: Named entity recognition · Event extraction · Anaphora resolution · Reaction reference resolution · Table classification · Chemical patents · Text mining

1 Overview

The ChEMU campaign focuses on information extraction tasks over chemical reactions in patents. The ChEMU2020 lab [5,6,12] provided two information extraction tasks, named entity recognition and event extraction. The ChEMU 2021 lab [4,9,10] introduced two more tasks, chemical reaction reference resolution and anaphora resolution. This year, we plan to re-run all the four tasks with a new task on semantic classification for tables as the fifth one. Together, the tasks support comprehensive automatic chemical patent analysis.

M. Hagen et al. (Eds.): ECIR 2022, LNCS 13186, pp. 400–407, 2022.
https://doi.org/10.1007/978-3-030-99739-7_50

1.1 Why Is This Campaign Needed?

The discovery of new chemical compounds is a key driver of the chemistry and pharmaceutical industries, and many other industrial sectors. Patents serve as a critical source of information about new chemical compounds. Compared with journal publications, patents provide more timely and comprehensive information about new chemical compounds [1,2,13], since they are usually the first venues where new chemical compounds are disclosed. Despite the significant commercial and research value of the information in patents, manual effort is still the primary mechanism for extracting and organising this information. This is costly, considering the large volume of patents available [7,11]. Development of automatic natural language processing (NLP) systems for chemical patents, which aim to convert text corpora into structured knowledge about chemical compounds, has become a focus of recent research [6,8].

1.2 How Would the Community Benefit from the Campaign?

There are three key benefits of this campaign to our community. First, our tasks provide a unique chance for NLP experts to develop information extraction models for chemical patents and gain experience in analysing the linguistic properties of patent documents. Second, several high-quality data sets will be released for a range of complex information extraction tasks that have applicability beyond the chemical domain. Finally, the tasks provided in this campaign focus on the field of information extraction over chemical literature, which is an active research area. The campaign will provide strong baselines as well as a useful resource for future research in this area.

1.3 Usage Scenarios

The details of chemical synthesis are critical for tasks including drug design and analysis of environmental or health impacts of material manufacturing. A key usage scenario for ChEMU is population of databases collecting detailed information about chemicals such as Reaxys®,[1] The tasks within the ChEMU 2022 lab will lead towards detailed understanding of complex descriptions of chemicals, chemical properties, and chemical reactions in chemical patents, addressing a number of natural language processing challenges involving both local and longer-distance relations and table analysis.

2 Tasks

We first briefly introduce the tasks from previous years, then describe the new table classification task. For more details about previous tasks, please refer to the corresponding overview paper ChEMU 2020 [6,12], ChEMU 2021 [4,9], and our website hosting the shared tasks[2].

[1] Reaxys® Copyright ©2021 Elsevier Life Sciences IP Limited. Reaxys is a trademark of Elsevier Life Sciences IP Limited, used under license. https://www.reaxys.com.

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

2.1 Task 1 Expression-Level Information Extraction

Task 1 consists of three sub-tasks, i.e. named entity recognition, event extraction, and anaphora resolution, since they only consider entities or relations between them within a few consecutive sentences.

In our ChEMU corpus, every snippet has been annotated for all three tasks, which opens the opportunity to explore multi-task learning since the input data is the same for all three tasks, as illustrated in Table 1. Fang et al. [3] extended coreference resolution with four other bridging relations as the anaphora resolution task. Results show that the performance of coreference resolution model can be further improved if bridging relation annotations are also available on the same data and the model is jointly trained for 5 relations instead of just coreference. One possible explanation for this is that a large part of the jointly trained model is shared for both coreference resolution and bridging relation tasks so effectively the jointly trained model is making use of more data which reduces the risk of overfitting and improves its ability to generalization. We expect more exploration towards this direction.

Task 1a Named Entity Recognition. This task aims to identify chemical compounds and their specific types. In addition, 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. In total, the participants need to find 10 types of named entities.

Task 1b Event Extraction. A chemical reaction leading to an end product often consists of a sequence of individual event steps. This task is to identify those steps which involve chemical entities recognized from Task 1a. It requires identification of event trigger words (e.g. “added” and “stirred”) and then determination of the chemical entity arguments of these events.

Task 1c Anaphora Resolution. This task requires the resolution of anaphoric dependencies between expressions in chemical patents. The participants are required to find five types of anaphoric relationships in chemical patents, i.e. coreference, transformed, reaction-associated, work-up and contained.

2.2 Task 2 Document-Level Information Extraction

Tasks 2 groups together the two tasks chemical reaction reference resolution and table semantic classification, since both of these tasks take a complete patent document as input rather than the short snippet extracts of Task 1. This increases the complexity of the task from a language processing perspective. The reaction references can relate reaction descriptions that are far apart, and the semantics of a table may depend on linguistic context from the document structure or content (Table 2).

Table 1. Illustration of three tasks performed on the same snippet, namely, Task 1a Named Entity Recognition (NER), Task 1b Event Extraction (EE), and Task 1c Anaphora Resolution (AR).

Text	The title compound was used without purification (1.180 g, 95.2%) as yellow solid
NER	The **title compound** was used without purification (**1.180 g**, **95.2%**) as yellow solid REACTION_PRODUCT: **title compound** YIELD_OTHER: **1.180 g** YIELD_PERCENT: **95.2%**
EE	The **title compound** was *used* without purification (**1.180 g**, **95.2%**) as yellow solid REACTION_STEP: *used* → REACTION_PRODUCT: **title compound** REACTION_STEP: *used* → YIELD_OTHER: **1.180 g** REACTION_STEP: *used* → YIELD_PERCENT: **95.2%**
AR	**The title compound** was used without purification **(1.180 g, 95.2%)** as *yellow solid* COREFERENCE: *yellow solid* → **The title compound (1.180 g, 95.2%)**

Table 2. An example for Task 2a chemical reaction reference resolution, where reaction 2 (RX2) is producing Compound B13 following the procedure that reaction 1 (RX1) produces Compound B11.

	Text
RX1	A mixture of the obtained ester, ... was stirred under argon and heated at 110 °C. for 24 h. ... Column chromatography of the residue (silica gel-hexane/ethyl acetate, 9:1) gave **Compound B11**, ...
	...
RX2	Using 2-ethoxyethanol and following the procedure for **Compound B11** gave Compound B13, bis(2-ethoxyethyl) 3,3'-((2-(bromomethyl)-2-((3-((2-ethoxyethoxy)carbonyl)phenoxy)methyl)propane-1,3-diyl)bis(oxy))dibenzoate,

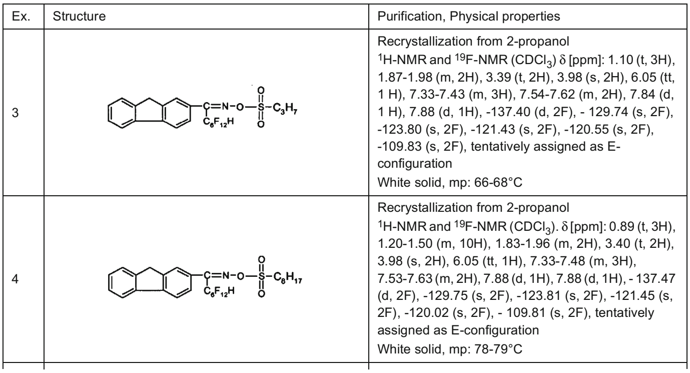
Some preliminary results on these tasks show that traditional machine learning models perform reasonably well and can sometimes do better than neural network models, especially on minority classes. It would be interesting to see if there exists a combined model that has the best of two worlds.

Task 2a Chemical Reaction Reference Resolution. Given a reaction description, this task requires identifying references to other reactions that the reaction relates to, and to the general conditions that it depends on. The participants are required to find pairs of reactions where one of them is the general condition for or is analogous to the other reaction.

Task 2b Table Semantic Classification. This task is about categorising tables in chemical patents based on their contents, which supports identification of tables containing key information. We define 8 types of tables as shown in Table 3. Figure 1 shows an example SPECT table. Please refer to Zhai et al. [15] for the dataset and Zhai et al. [16] for more details on the settings of this task.

Table 3. 8 labels defined for Task 2b semantic classification on tables, and examples of expected content.

Label	Description	Examples
SPECT	Spectroscopic data	Mass spectrometry, IR/NMR spectroscopy
PHYS	Physical data	Melting point, quantum chemical calculations
IDE	Identification of compounds	Chemical name, structure, formula, label
RX	All properties of reactions	Starting materials, products, yields
PHARM	Pharmacological data	Pharmacological usage of chemicals
COMPOSITION	Compositions of mixtures	Compositions made up by multiple ingredients
PROPERTY	Properties of chemicals	The time of resistance of a photoresist
OTHER	Other tables	–



Ex.	Structure	Purification, Physical properties
3		Recrystallization from 2-propanol ^{1}H-NMR and ^{19}F-NMR ($CDCl_3$) δ [ppm]: 1.10 (t, 3H), 1.87-1.98 (m, 2H), 3.39 (t, 2H), 3.98 (s, 2H), 6.05 (tt, 1 H), 7.33-7.43 (m, 3H), 7.54-7.62 (m, 2H), 7.84 (d, 1 H), 7.88 (d, 1H), -137.40 (d, 2F), - 129.74 (s, 2F), -123.80 (s, 2F), -121.43 (s, 2F), -120.55 (s, 2F), -109.83 (s, 2F), tentatively assigned as E-configuration White solid, mp: 66-68°C
4		Recrystallization from 2-propanol ^{1}H-NMR and ^{19}F-NMR ($CDCl_3$). δ [ppm]: 0.89 (t, 3H), 1.20-1.50 (m, 10H), 1.83-1.96 (m, 2H), 3.40 (t, 2H), 3.98 (s, 2H), 6.05 (tt, 1H), 7.33-7.48 (m, 3H), 7.53-7.63 (m, 2H), 7.88 (d, 1H), 7.88 (d, 1H), - 137.47 (d, 2F), -129.75 (s, 2F), -123.81 (s, 2F), -121.45 (s, 2F), -120.02 (s, 2F), - 109.81 (s, 2F), tentatively assigned as E-configuration White solid, mp: 78-79°C

Fig. 1. An example table in SPECT category.

2.3 Changes Proposed for Rerunning Previous Tasks

The number of participating teams in ChEMU 2021 was much lower than that in ChEMU 2020 (2 vs. 11 teams). We believe the primary reason for this was

Table 4. A summary of the information about participation, data, and baseline models for all tasks. NER is short for Named Entity Recognition, EE for Event Extraction, AR for Anaphora Resolution, CR3 for Chemical Reaction Reference Resolution, TSC for Table Semantic Classification.

Task	Continued?	Data	Baseline models
1a NER	2020 task 1	Existing 1500 snippets as train and	He et al. [6]
1b EE	2020 task 2	dev sets. 500 new snippets will be	
1c AR	2021 task 2	annotated and used as the test set	Fang et al. [3]
2a CR3	2021 task 1	Data for ChEMU 2021 will be reused	Yoshikawa et al. [14]
2b TSC	New task	All the data is ready for release	Zhai et al. [16]

that the time given to participants was too short. The data for both tasks of ChEMU 2021 was released in early April, while the deadline for submitting the final predictions on test set was in mid-May, which left only 6 weeks to the participants to build and test their models. Additionally, the pandemic is not over yet, and one team mentioned that they faced several related challenges. Both teams that participated in ChEMU 2021 Task 2 asked for extensions to the various deadlines. This year, we will release the data for ChEMU 2022 by the end of this year, so that the participants will have a few months instead of a few weeks to work on them.

Furthermore, we will simplify the two tasks from ChEMU 2021 (2022 Tasks 1c and 2a), by providing the gold spans of mentions and chemical reactions, respectively. Since both teams have proposed a few potential directions for improving their relation extraction component, we hope to support exploration of more ideas on the this part. The simplification will also make it easier for participants to build models, and could potentially attract more people.

2.4 Data and Evaluation

A new corpus for Task 2b of 788 patents containing annotated tables will be first split into training, development, and test sets according to 60%/15%/25% portion. The training and development sets will be released in December, and the test set without annotations will be released one week before the evaluation deadline.

Data for other tasks will be released following the same schedule. For the three tasks of Task 1, the data released for ChEMU 2020 and 2021 (1500 snippets) will serve as the training and development sets, while 500 new snippets will be annotated for all three tasks and used as the test set. Since no one participated in Task 2a (ChEMU 2021 Task 1), its test set is untouched. Therefore, the data for this task will be reused as is for ChEMU 2022.

For evaluation, standard precision, recall, and F1 score will be used. For each task, we will take the model from our published papers as strong baselines and make them available to all participants, as shown in Table 4.

3 Conclusion

In this paper, we have presented a brief description of the upcoming ChEMU lab at CLEF-2022 including the re-run of all four tasks from ChEMU 2020/2021 and a new table semantic classification task.

We expect participants from both academia and industry and will advertise our tasks via social media and NLP-related mailing lists. In addition, we will invite previous participants and authors who have submitted to Frontiers In Research Metrics and Analytics special issue (Information Extraction from Bio-Chemical Text) to join ChEMU 2022.

References

1. Akhondi, S.A., et al.: Automatic identification of relevant chemical compounds from patents. Database **2019**, baz001 (2019)
2. Bregonje, M.: Patents: a unique source for scientific technical information in chemistry related industry? World Patent Inf. **27**(4), 309–315 (2005)
3. Fang, B., Druckenbrodt, C., Akhondi, S.A., He, J., Baldwin, T., Verspoor, K.M.: ChEMU-Ref: a corpus for modeling anaphora resolution in the chemical domain. In: Merlo, P., Tiedemann, J., Tsarfaty, R. (eds.) Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021, Online, 19–23 April 2021, pp. 1362–1375. Association for Computational Linguistics (2021). https://www.aclweb.org/anthology/2021.eacl-main.116/
4. He, J., et al.: ChEMU 2021: reaction reference resolution and Anaphora resolution in chemical patents. In: Hiemstra, D., Moens, M.-F., Mothe, J., Perego, R., Potthast, M., Sebastiani, F. (eds.) ECIR 2021. LNCS, vol. 12657, pp. 608–615. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-72240-1_71
5. He, J., et al.: Overview of ChEMU 2020: named entity recognition and event extraction of chemical reactions from patents. In: Arampatzis, A., et al. (eds.) CLEF 2020. LNCS, vol. 12260, pp. 237–254. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-58219-7_18
6. He, J., et al.: ChEMU 2020: natural language processing methods are effective for information extraction from chemical patents. Frontiers Res. Metrics Anal. **6**, 654438 (2021). https://doi.org/10.3389/frma.2021.654438
7. Hu, M., Cinciruk, D., Walsh, J.M.: Improving automated patent claim parsing: dataset, system, and experiments. arXiv preprint arXiv:1605.01744 (2016)
8. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: CHEMDNER: the drugs and chemical names extraction challenge. J. Cheminform. **7**(1), 1–11 (2015)
9. Li, Y., et al.: Overview of ChEMU 2021: reaction reference resolution and Anaphora resolution in chemical patents. In: Candan, K.S., et al. (eds.) CLEF 2021. LNCS, vol. 12880, pp. 292–307. Springer, Cham (2021). https://doi.org/10.1007/978-3-030-85251-1_20
10. Li, Y., et al.: Extended overview of ChEMU 2021: reaction reference resolution and anaphora resolution in chemical patents. In: Faggioli, G., Ferro, N., Joly, A., Maistro, M., Piroi, F. (eds.) Proceedings of the Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum, Bucharest, Romania, 21st–24th September 2021. CEUR Workshop Proceedings, vol. 2936, pp. 693–709. CEUR-WS.org (2021). http://ceur-ws.org/Vol-2936/paper-58.pdf

11. Muresan, S., et al.: Making every SAR point count: the development of chemistry connect for the large-scale integration of structure and bioactivity data. Drug Discovery Today **16**(23–24), 1019–1030 (2011)
12. Nguyen, D.Q., et al.: ChEMU: named entity recognition and event extraction of chemical reactions from patents. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 572–579. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_74
13. Senger, S., Bartek, L., Papadatos, G., Gaulton, A.: Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. J. Cheminform. **7**(1), 1–12 (2015). https://doi.org/10.1186/s13321-015-0097-z
14. Yoshikawa, H., et al.: Chemical reaction reference resolution in patents. In: Proceedings of the 2nd Workshop on on Patent Text Mining and Semantic Technologies (2021)
15. Zhai, Z., et al.: ChemTables: dataset for table classification in chemical patents (2021). https://doi.org/10.17632/g7tjh7tbrj.3
16. Zhai, Z., et al.: ChemTables: a dataset for semantic classification on tables in chemical patents. J. Cheminform. **13**(1), 97 (2021). https://doi.org/10.1186/s13321-021-00568-2