




ChEMU 2021: Reaction Reference Resolution and Anaphora Resolution in Chemical Patents

Jiayuan He^{1,3}, Biaoyan Fang¹, Hiyori Yoshikawa^{1,4}, Yuan Li¹, Saber A. Akhondi², Christian Druckenbrodt⁵, Camilo Thorne⁵, Zubair Afzal², Zenan Zhai¹, Lawrence Cavedon³, Trevor Cohn¹, Timothy Baldwin¹, and Karin Verspoor^{1,3} 

¹ The University of Melbourne, Melbourne, Australia

² Elsevier B.V., Amsterdam, The Netherlands

³ RMIT University, Melbourne, Australia

karin.verspoor@rmit.edu.au

⁴ Fujitsu Laboratories Ltd., Kawasaki, Japan

⁵ Elsevier Information Systems GmbH, Frankfurt am Main, Germany

Abstract. Chemical patents serve as an indispensable source of information about new discoveries of chemical compounds. The ChEMU (Cheminformatics Elsevier Melbourne University) lab addresses information extraction over chemical patents, and aims to advance the state of the art on this topic. ChEMU lab 2021, as part of the 12th Conference and Labs of the Evaluation Forum (CLEF-2021), will be the second ChEMU lab. ChEMU 2021 will provide two distinct tasks related to reference resolution in chemical patents. Task 1—Chemical Reaction Reference Resolution—focuses on paragraph-level references and aims to identify the chemical reactions or general conditions specified in one reaction description referred to by another. Task 2—Anaphora Resolution—focuses on expression-level references and aims to identify the reference relationships between expressions in chemical reaction descriptions. In this paper, we introduce ChEMU 2021, including its motivation, goals, tasks, resources, and evaluation framework.

Keywords: Reaction reference resolution · Anaphora resolution · Chemical patents · Text mining

1 Introduction

The discovery of new chemical compounds is perceived as a key driver of the chemical industry and many other industrial sectors, and information relevant for this discovery is found in chemical synthesis descriptions in natural language texts. In particular, 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, 4, 18], 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 extraction of such information is costly, considering the large volume of patents available [10, 13]. Thus, developing automatic natural language processing (NLP) systems for chemical patents, which convert text corpora into structured knowledge about chemical compounds, has become a focus of recent research [11, 14].

The ChEMU campaign focuses on information extraction tasks over chemical reactions in patents¹. ChEMU 2020 [9, 14] provided two information extraction tasks, named entity recognition (NER) and event extraction, and attracted 37 teams around the world to participate. In the ChEMU 2021 lab, we will provide two new information extraction tasks: chemical reaction reference resolution and anaphora resolution, focusing on reference resolution in chemical patents. Compared with previous shared tasks dealing with anaphora resolution, e.g., the CRAFT co-reference task [3], our proposed tasks extend the scope of reference resolution by considering reference relationships on both paragraph-level and expression-level (see Fig. 1). Specifically, our first task aims at the identification of reference relationships between reaction descriptions. Our second task aims at the identification of reference relationships between chemical expressions, including both co-reference and bridging. Moreover, we focus on chemical patents while the CRAFT co-reference task focused on journal articles.

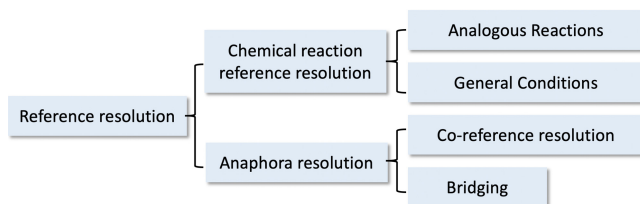


Fig. 1. Illustration of the task hierarchy.

The ChEMU lab 2021 will be a challenging opportunity for researchers to further improve the sophistication of information extraction systems for chemical patents. In this paper, we introduce our motivation and goals, a detailed description of the tasks, and our evaluation framework.

2 Related Shared Tasks

Several shared tasks have addressed reference resolution in scientific literature. BioNLP2011 hosted a subtask on protein co-reference [15]. CRAFT-CR 2019 hosted a subtask on co-reference resolution in biomedical articles [3]. However, these shared tasks differ from ours in several respects.

First, previous shared tasks considered different domains of scientific literature. For example, the dataset used in BioNLP2011 is derived from the GENIA

¹ Our main website is <http://chemu.eng.unimelb.edu.au>.

corpus [16], which primarily focuses on the biological domain, viz. gene/proteins and their regulations. The dataset used in CRAFT-CR co-reference shared task is based on biomedical journal articles in PubMed [2, 5]. Our ChEMU shared task focuses by contrast on the domain of chemical patents. This difference entails the critical importance for this shared task: information extraction methodologies for general scientific literature or the biomedical domain will not be effective for chemical patents [12]. It is widely acknowledged that patents are written quite differently as compared with general scientific literature, resulting in substantially different linguistic properties. For example, patent authors may trade some clarity in wording for more protection of their intellectual property.

Secondly, our reference resolution tasks include both paragraph-level and entity-level reference phenomena. Our first task aims at identification of reference relationships between reaction descriptions, i.e., paragraph-level. This task is challenging because a reaction description may refer to an extremely remote reaction and thus requires processing of very long documents. Our second task aims at anaphora resolution, similarly to previous entity-level co-reference tasks. However, a key difference is that we extend the scope of this task by including both co-reference and bridging phenomena. That is, we not only aim at finding expressions referring to the same entity, but also expressions that are semantically related or associated.

3 Goals and Importance

The goals of ChEMU2021 are three-fold: We aim to (1) develop tasks that address fundamental challenges in automatic information extraction over chemical patents; (2) provide the community with a new dataset that can serve as benchmark datasets for future method development; and (3) advance the state-of-the-art technologies in information extraction over chemical patents together with worldwide NLP experts.

Our tasks provide an important opportunity for NLP experts to develop information extraction models for chemical patents and gain experience in analysing the linguistic properties of patent documents. The campaign will provide strong baselines as well as useful resources for future research in this area.

4 Tasks

Task 1: 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. Assume a set of reaction statements (RSs), each of which corresponds to a description of an individual chemical reaction or a general condition for the reaction. By identifying all the reference relationships amongst these reaction statements, the details of reactions can be fully specified by connecting related reaction statements. Two types of reference relationships are defined in this task, namely *Analogous Reactions* and *General Conditions*.

(a) Analogous reactions

ID	Text
RS1	Prep. 2 1(R)-Benzyl-6-methoxy-1(S)-(3-oxo-butyl)-3,4-dihydro-1H-naphthalen-2-one A solution of 62 g (0.23 mol) of the title product of Preparation 1 and 28 mL, ...
RS2	Prep. 3 1(S)-Benzyl-6-methoxy-1(R)-(3-oxo-butyl)-3,4-dihydro-1H-naphthalen-2-one The title product of this preparation was prepared using a method analogous to Prep. 2, using (R)-(+)-alphamethyl benzylamine in the initial imine formation ...

(b) General conditions

ID	Text
RS3	Preparation Examples (1) Step (A) 4-Bromobenzaldehyde and boronic acid were subjected to Suzuki cross coupling reaction using a palladium catalyst as shown in [Scheme 1a]... (2) Steps (B) and (C) ... (3) Preparation of salt ...
RS4	Example 1: Synthesis of (S)-2-(((2'-fluorobiphenyl-4-yl)methyl)amino) propanamide methane-sulfonate White solid; yield: 90%; 1H NMR ...
RS5	Example 2: Synthesis of (S)-2-(((3'-fluorobiphenyl-4-yl)methyl)amino) propanamide methane-sulfonate White solid; yield: 97%; 1H NMR ...

Fig. 2. Abbreviated examples of reaction references: (a) analogous reactions [6] and (b) general conditions [17].

Examples of the two types of reaction reference relationships are given in Fig. 2. In Fig. 2(a), the description of RS2 contains a statement “using a method analogous to Prep. 2” which is highlighted in bold. This indicates a reference relationship from RS2 to RS1. In Fig. 2(b), a standard procedure RS3 is first given. Unlike RS1 and RS2 in analogous reactions, RS3 is not associated with any specific reaction. In addition, Scheme 1a in the figure illustrates RS3 via Markush structures, with a variable X that can be replaced with several sub-structures. These indicate that the following chemical reactions RS4 and RS5 refer to this procedure. These two reactions should each be linked with their common procedure RS3.

Task 2: Anaphora Resolution. This task requires the resolution of general anaphoric dependencies between expressions in chemical patents. In this task, we define five types of anaphoric relationships, common in chemical patents:

1. *Co-reference*: two expressions/mentions that refer to the same entity.
2. *Transformed*: two chemical compound entities that are initially based on the same chemical components and have undergone possible changes through various conditions (e.g., pH and temperature).
3. *Reaction-associated*: the relationship between a chemical compound and its immediate sources via a mixing process. The immediate sources do need to be reagents, but they need to end up in the corresponding product. The source compounds retain their original chemical structure.

4. *Work-up*: the relationship between chemical compounds that were used for isolation or purification purposes, and their corresponding output products.
5. *Contained*: the association holding between chemical compounds and the related equipment in which they are placed. The direction of the relation is from the related equipment to the previous chemical compound.

[Acetic acid (9.8 ml)] and [water (4.9 ml)] were added to [the solution] in [a flask]. [The mixture]₁ was stirred for 3 hrs at 50°C and then cooled to 0°C. 2N-sodium hydroxide aqueous solution was added to [the mixture]₂ until the pH of [the mixture]₃ became 9. [The mixture]₄ was extracted with [ethyl acetate] for 3 times. [The combined organic layer] was washed with water and saturated aqueous sodium chloride.

ID	Relation type	Anaphor	Antecedent
AR1	Co-reference	[The mixture] ₄	[the mixture] ₃
AR2	Transformed	[the mixture] ₂	[The mixture] ₁
AR3	Reaction-associated	[The mixture] ₁	[water (4.9 ml)]
AR4	Work-up	[The combined organic layer]	[ethyl acetate]
AR5	Contained	[a flask]	[the solution]

Fig. 3. Text snippet containing a chemical reaction, with its anaphoric relationships. The expressions that are involved are highlighted in **bold**. In the cases where several expressions have identical text form, subscripts are added according to their order of appearance.

Several anaphoric relationships can be extracted from the text snippet in Fig. 3. [The mixture]₄ and [the mixture]₃ refer to the same “mixture” and thus, form a co-reference relationship. The two expressions [The mixture]₁ and [the mixture]₂ are initially based on the same chemical components but the property of [the mixture]₂ changes after the “stir” and “cool” action. Thus, the two expressions should be linked as “Transformed”. The expression [The mixture]₁ comes from mixing the chemical compounds prior to it, e.g., [water (4.9 ml)]. Thus, the two expressions are linked as “Reaction-associated”. The expression [The combined organic layer] comes from the extraction of [ethyl acetate]. Thus, they are linked as “Work-up”. Finally, the expression [the solution] is contained by the entity [a flask], and the two are linked as “Contained”.

5 Data, Resources, and Evaluation

Dataset. A corpus extending the ChEMU 2020 dataset [19] is in development. The corpus contains patents from the European Patent Office and the United States Patent and Trademark Office, available in English in a digital format. The corpus is based on the Reaxys[®] database,² containing reaction entries for patent documents manually created by experts in chemistry.

² Reaxys[®] Copyright ©2020 Elsevier Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Limited. <https://www.reaxys.com>.

For Task 1, a collection of reaction descriptions will be provided with annotated reference relationships. A reaction entry in Reaxys has “locations” of the reaction in the corresponding patent document, mostly in terms of paragraph IDs. To date, a silver-standard dataset has been constructed using these locations and will be the foundation of a higher-quality gold-standard set. Existing work has established several baselines for this data [20].

For Task 2, the ChEMU-Ref corpus is in development [7, 8]. This a collection of reaction snippets with the expression-level reference relationships annotated. A detailed annotation guideline has been developed; two chemical experts have been trained for the annotation task and annotation is in progress. Several baselines will also be made available, following [7].

Resources. A number of existing resources can be utilized by participants to develop their approaches to these tasks. These include the ChELMo pre-trained ELMo embeddings for chemical patents [21] and datasets for the ChEMU 2020 Named Entity Recognition and Event Extraction tasks [9, 19].

Evaluation. For Task 1, we will use standard precision, recall, and F-score as our primary evaluation metrics. In addition, scores will be calculated for their *referrer detection* performance as well. This measure should reflect how well the model detects reactions that refer to at least one reaction description or a general condition.

In Task 2, we consider two types of co-reference linking, i.e., (1) surface co-reference linking and (2) atomic co-reference linking, due to the existence of *transitive co-reference relationships*. By transitive co-reference relationships we mean multi-hop co-reference such as a link from an expression T1 to T3 via an intermediate expression T2, viz., “T1→T2→T3”. Surface co-reference linking will restrict attention to one-hop relationships, viz., to: “T1→T2” and “T2→T3”. Whereas atomic co-reference linking will tackle co-reference between an anaphoric expression and its first antecedent, i.e., intermediate antecedents will be collapsed. Thus, these two links will be used for the above example, “T1→T3” and “T2→T3”. Note that we only consider transitive linking in co-reference relationships. In addition, the criteria of both exact and relaxed text-span matching will be used. We will use F-score in terms of exact text-span matching and surface linking as the primary system ranking metric for Task 2.

6 Conclusions

In this paper, we introduced our upcoming lab ChEMU 2021. As the second instance of our ChEMU lab series, ChEMU 2021 will provide two new tasks focusing on reference resolution in chemical patents. Our first task aims at identification of reference relationships between chemical reaction descriptions, and our second task aims at identification of reference relationships between expressions in chemical reactions. We look forward to seeing innovative approaches to these complex tasks.

Acknowledgements. Funding for the ChEMU project is provided by an Australian Research Council Linkage Project, project number LP160101469, and Elsevier.

References

1. Akhondi, S.A., et al.: Automatic identification of relevant chemical compounds from patents. In: Database (2019)
2. Bada, M., et al.: Concept annotation in the CRAFT corpus. *BMC Bioinform.* **13**, 161 (2012). <https://doi.org/10.1186/1471-2105-13-161>. <https://www.ncbi.nlm.nih.gov/pubmed/22776079>
3. Baumgartner Jr, W.A., et al.: CRAFT shared tasks 2019 overview—integrated structure, semantics, and coreference. In: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, pp. 174–184 (2019)
4. Bregonje, M.: Patents: a unique source for scientific technical information in chemistry related industry? *World Patent Inf.* **27**(4), 309–315 (2005)
5. Cohen, K.B., et al.: Coreference annotation and resolution in the Colorado Richly Annotated Full Text (CRAFT) corpus of biomedical journal articles. *BMC Bioinform.* **18**(1), 1–14 (2017)
6. Dow, R.L., Liu, K.K.C., Morgan, B.P., Swick, A.G.: Glucocorticoid receptor modulators. European patent no. EP1175383B1 (2018)
7. Fang, B., Druckenbrodt, C., Akhondi, S.A., He, J., Baldwin, T., Verspoor, K.: ChEMU-Ref: a corpus for modeling anaphora resolution in the chemical domain. In: Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics. Association for Computational Linguistics, April 2021
8. Fang, B., et al.: ChEMU-ref dataset for modeling anaphora resolution in the chemical domain (2021). <https://doi.org/10.17632/r28xxr6p92>
9. 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
10. Hu, M., Cinciruk, D., Walsh, J.M.: Improving automated patent claim parsing: dataset, system, and experiments. arXiv preprint [arXiv:1605.01744](https://arxiv.org/abs/1605.01744) (2016)
11. Krallinger, M., Leitner, F., Rabal, O., Vazquez, M., Oyarzabal, J., Valencia, A.: CHEMDNER: the drugs and chemical names extraction challenge. *J. Cheminform.* **7**(S1), S1 (2015)
12. Lupu, M., Mayer, K., Kando, N., Trippe, A.J.: Current Challenges in Patent Information Retrieval, vol. 37. Springer, Heidelberg (2017). <https://doi.org/10.1007/978-3-662-53817-3>
13. 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 Discov. Today* **16**(23-24), 1019–1030 (2011)
14. 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
15. Nguyen, N., Kim, J.D., Tsujii, J.: Overview of BioNLP 2011 protein coreference shared task. In: Proceedings of BioNLP Shared Task 2011 Workshop, pp. 74–82 (2011)

16. Ohta, T., Tateisi, Y., Kim, J.D., Mima, H., Tsujii, J.: The GENIA corpus: an annotated research abstract corpus in molecular biology domain. In: Proceedings of the Second International Conference on Human Language Technology Research, pp. 82–86 (2002)
17. Park, K.D., et al.: Alpha-aminoamide derivative compound and pharmaceutical composition comprising same. European patent no. EP3202759A1 (2017)
18. 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)
19. Verspoor, K., et al.: ChEMU dataset for information extraction from chemical patents (2020). <https://doi.org/10.17632/wy6745bjfj>
20. Yoshikawa, H., et al.: Detecting chemical reactions in patents. In: Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association, pp. 100–110. Australasian Language Technology Association, Sydney, Australia, 4–6 December 2019. <https://www.aclweb.org/anthology/U19-1014>
21. Zhai, Z., et al.: Improving chemical named entity recognition in patents with contextualized word embeddings. In: Proceedings of the 18th BioNLP Workshop and Shared Task. pp. 328–338. Association for Computational Linguistics, Florence, Italy, August 2019. <https://doi.org/10.18653/v1/W19-5035>. <https://www.aclweb.org/anthology/W19-5035>