End-to-End Chemical Reaction Extraction from Patents

Yuan Li
yuan.li1@unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Jiayuan He*
jiayuan.he@rmit.edu.au
RMIT University
Melbourne, Australia

Hiyori Yoshikawa*
hiyori.yoshikawa@unimelb.edu.au
Fujitsu Limited
Minato Ward, Japan

Biaoyan Fang
biaoyanf@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Zenan Zhai
zenan.zhai@student.unimelb.edu.au
The University of Melbourne
Melbourne, Australia

Christian Druckenbrodt
c.druckenbrodt@elsevier.com
Elsevier Information Systems GmbH
Frankfurt, Germany

Camilo Thorne
c.thorne.1@elsevier.com
Elsevier Information Systems GmbH
Frankfurt, Germany

Saber A. Akhondi
s.akhondi@elsevier.com
Elsevier BV
Amsterdam, Netherlands

Karin Verspoor*†
karin.verspoor@rmit.edu.au
RMIT University
Melbourne, Australia

ABSTRACT

With the rapid growth of chemical patents, there is increasing demand for automated extraction of information relating to chemical compounds and their synthesis from patents. Although there are existing models that can extract chemical entities and reaction events, these have significant practical limitations. First, they typically cannot process a full patent document, targeting short texts containing only reaction descriptions. Second, they neglect reaction texts where steps in the reaction are elided through reference to other reactions. To address these issues, we propose an integrated and comprehensive chemical reaction extraction system consisting of a pipeline of components for reaction detection, chemical named entity recognition, event extraction, anaphora resolution, reaction reference resolution, and table classification.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

information extraction, named entity recognition, event extraction, anaphora resolution, chemical reactions, patent text mining

ACM Reference Format:
Yuan Li, Jiayuan He, Hiyori Yoshikawa, Biaoyan Fang, Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, Saber A. Akhondi, and Karin Verspoor. 2022. End-to-End Chemical Reaction Extraction from Patents. In *Proceedings of PatentSemTech'22.* ACM, New York, NY, USA, 2 pages. https://doi.org/XXXXXXX.XXXXXXX

*Also with The University of Melbourne.
†Corresponding author.

PatentSemTech'22, July 15, 2022, Madrid, Spain

ACM ISBN 978-1-4503-XXXX-X/18/06…$15.00
https://doi.org/XXXXXXX.XXXXXXX

1 INTRODUCTION

The discovery of new chemical compounds is a key driver of the chemistry and pharmaceutical industries, *inter alia*. Patents serve as a critical source of information about new chemical compounds, providing timely and comprehensive information about new chemical compounds [1, 2]. Despite the significant commercial and research value of the information in patents, manual effort is still the primary mechanism for extracting and organizing this information. This is costly, considering the large volume of patents available [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 [9, 10].

In this study we consider a system that focuses on chemical reaction processes described in chemical patents. A chemical reaction is a process leading to the transformation of one set of chemical substances to another. A full reaction requires at least the starting materials and the final product to be defined, and usually includes information such as reagents, catalysts, and experiment conditions to further describe the reaction. Our overarching objective is to enable the automatic identification of each reaction described in a complete patent document, and to fully characterize each reaction by extracting each relevant component.

2 SYSTEM OVERVIEW

To perform end-to-end extraction of chemical reactions from full patents, we define a pipeline of interconnected NLP tasks.

Reaction snippet detection: We first need to locate reaction descriptions in a patent, for processing in downstream tasks. We formulate this task as a paragraph-level sequence tagging problem, where a patent is given as a sequence of paragraphs and the task is to detect a span of contiguous paragraphs describing a single chemical reaction. We train a BiLSTM-CRF model for this task on the dataset described in [13] using the same experimental settings.

Chemical NER:. Using the reaction snippets extracted from full patents, the task to identify chemical entities and their roles in a chemical reaction can be formulated as named entity recognition

(NER). We train a BERT-CRF model for this task using the annotation schema and data for chemical NER task detailed in [7, 8].

Event extraction: A chemical reaction usually consists of an ordered sequence of *event steps* that either transforms a starting material into a product or just purifies or isolates a chemical substance. An event is characterised by (a) a trigger word that flags its occurrence, and (b) a relation connecting the trigger word and chemical entities involved in the event. For this task, we use a BERT-CRF model to extract trigger words and chemical entities from snippets and borrow ideas from the span-based BERT model in [5]. In this approach, all pairs of trigger words and entities are enumerated, BERT is applied to obtain the contextualized representation of each relevant token, and a classifier decides the nature of the relation between them using pooling of token representations.

Anaphora resolution: There are rich anaphoric relations *between* and *within* event steps. We consider two main types of anaphoric relations defined in [6]: coreference, where two mentions refer to the same entity, and bridging, linking a chemical compound and its source. We decompose this task into (a) anaphor mention detection and (b) relation classification. We use a BERT-CRF model for mention detection. For relation classification, we adopt the span-based BERT model proposed in [4].

Reaction reference resolution: So far, we have assumed that a reaction snippet contains the complete information of a chemical reaction. However, chemical patents often detail several similar compounds that have a common substructure and can be synthesized in analogous ways. They contain many references connecting descriptions of similar chemical reactions, to avoid redundancy in describing common reaction conditions. This leads to the problem of identifying references from an incomplete snippet to others. Here, we use the model proposed in [12], first determining if a snippet has others that refer to it, and then enumerating possible reference pairs of snippets and classifying them.

Table classification: Apart from text paragraphs, a large amount of information in patents is represented in tables and images. Here, we focus on identifying tables containing chemical reaction properties such as starting materials, products, yields, etc. To differentiate tables of interest from others, we train a Table-BERT classifier [3] on the ChemTables data [14]. The model first concatenates all tokens within all cells from the table and then takes the flattened table as input. For tables classified into reaction properties category, we further extract reactions based on the table header if there are sufficient information describing reactions.

3 DISCUSSION

We have introduced the essential requirements for building a comprehensive chemical reaction extraction system covering a wide range of tasks. We have proposed an initial approach for each step leveraging existing data resources from the ChEMU shared tasks, illustrating how the individual tasks can be brought together into a coherent whole. This integration addresses two key limitations of previous studies: our system can process full patent documents directly, and we can find the snippets an incomplete reaction snippet refers to. We leave performance evaluation of individual steps, as well as the complete system, to a more in-depth presentation. In the future, we plan to further develop this framework to extract complete reaction information by incorporating inference over reaction references, and to extend the scope of our system to handle images and chemical structures. Opportunities also exist to explore joint modelling or multi-task learning across the constituent tasks in this pipeline, for instance coupling NER and anaphora resolution.

REFERENCES

[1] Saber A Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, et al. 2019. Automatic identification of relevant chemical compounds from patents. *Database* 2019 (2019).

[2] Mervyn Bregonje. 2005. Patents: A unique source for scientific technical information in chemistry related industry? *World Patent Information* 27, 4 (2005), 309–315.

[3] Wenhu Chen, Hongmin Wang, Jianshu Chen, Yunkai Zhang, Hong Wang, Shiyang Li, Xiyou Zhou, and William Yang Wang. 2020. TabFact: A Large-scale Dataset for Table-based Fact Verification. In *8th International Conference on Learning Representations, ICLR 2020, Addis Ababa, Ethiopia, April 26-30, 2020*. OpenReview.net.

[4] Ritam Dutt, Sopan Khosla, and Carolyn P. Rosé. 2021. A pipelined approach to Anaphora Resolution in Chemical Patents. In *Working Notes of CLEF 2021 - Conference and Labs of the Evaluation Forum (CEUR Workshop Proceedings, Vol. 2936)*. CEUR-WS.org, 710–719.

[5] Markus Eberts and Adrian Ulges. 2020. Span-Based Joint Entity and Relation Extraction with Transformer Pre-Training. In *ECAI 2020 - 24th European Conference on Artificial Intelligence (Frontiers in Artificial Intelligence and Applications, Vol. 325)*. IOS Press, 2006–2013.

[6] Biaoyan Fang, Christian Druckenbrodt, Saber A. Akhondi, Jiayuan He, Timothy Baldwin, and Karin M. Verspoor. 2021. ChEMU-Ref: A Corpus for Modeling Anaphora Resolution in the Chemical Domain. In *Proc. 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume, EACL 2021*. Association for Computational Linguistics, 1362–1375.

[7] Jiayuan He, Biaoyan Fang, Hiyori Yoshikawa, Yuan Li, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Zubair Afzal, Zenan Zhai, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU 2021: Reaction Reference Resolution and Anaphora Resolution in Chemical Patents. In *Advances in Information Retrieval - 43rd European Conf. on IR Research, ECIR 2021, Part II (Lecture Notes in Computer Science, Vol. 12657)*. Springer, 608–615.

[8] Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2020. Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In *Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th Intl Conf of the CLEF Association, CLEF 2020 (Lecture Notes in Computer Science, Vol. 12260)*. Springer, 237–254.

[9] Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU 2020: Natural Language Processing Methods Are Effective for Information Extraction From Chemical Patents. *Frontiers Res. Metrics Anal.* 6 (2021), 654438.

[10] Martin Krallinger, Florian Leitner, Obdulia Rabal, Miguel Vazquez, Julen Oyarzabal, and Alfonso Valencia. 2015. CHEMDNER: The drugs and chemical names extraction challenge. *Journal of Cheminformatics* 7, 1 (2015), 1–11.

[11] Sorel Muresan, Plamen Petrov, Christopher Southan, Magnus J Kjellberg, Thierry Kogej, Christian Tyrchan, Peter Varkonyi, and Paul Hongxing Xie. 2011. 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 (2011), 1019–1030.

[12] Hiyori Yoshikawa, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Ralph Hoessel, Zenan Zhai, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. Chemical Reaction Reference Resolution in Patents. In *Proc. 2nd Workshop on Patent Text Mining and Semantic Technologies*.

[13] Hiyori Yoshikawa, Dat Quoc Nguyen, Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, Saber A. Akhondi, Timothy Baldwin, and Karin Verspoor. 2019. Detecting Chemical Reactions in Patents. In *Proc. 17th Annual Workshop of the Australasian Language Technology Association, ALTA 2019, Sydney, Australia, December 4-6, 2019*. 100–110.

[14] Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, Saber A. Akhondi, Dat Quoc Nguyen, Trevor Cohn, and Karin M. Verspoor. 2021. ChemTables: A dataset for semantic classification on tables in chemical patents. *J. Cheminformatics* 13, 1 (2021), 97.