Extended Overview of ChEMU 2022 Evaluation Campaign: Information Extraction in Chemical Patents

Yuan Li[1], Biaoyan Fang[1], Jiayuan He[1,4], Hiyori Yoshikawa[1,5], Saber A. Akhondi[2], Christian Druckenbrodt[3], Camilo Thorne[3], Zubair Afzal[2], Zenan Zhai[1], Kojiro Machi[6], Masaharu Yoshioka[6], Youngrok Jang[7], Hosung Song[7], Junho Lee[8], Gyeonghun Kim[7], Yireun Kim[7], Stanley Jungkyu Choi[7], Honglak Lee[7], Kyunghoon Bae[7], Darshini Mahendran[9], Christina Tang[9], Bridget McInnes[9], Timothy Baldwin[1] and Karin Verspoor[4,1]

[1]*The University of Melbourne, Australia*
[2]*Elsevier BV, Netherlands*
[3]*Elsevier Information Systems GmbH, Germany*
[4]*RMIT University, Australia*
[5]*Fujitsu Limited, Japan*
[6]*Hokkaido University, Japan*
[7]*LG AI Research, South Korea*
[8]*LG DISPLAY, South Korea*
[9]*Virginia Commonwealth University, United States*

Abstract

In this paper, we provide an overview of the Cheminformatics Elsevier Melbourne University (ChEMU) evaluation lab 2022, part of the Conference and Labs of the Evaluation Forum 2022 (CLEF 2022). The ChEMU campaign focuses on information extraction tasks over chemical reactions in patents. The ChEMU 2020 lab provided two information extraction tasks, named entity recognition and event extraction. The ChEMU 2021 lab introduced one more task, anaphora resolution. This year, we re-run all the three tasks with new test data. Together, the tasks support comprehensive automatic chemical patent analysis. Herein, we describe the resources created for these tasks and the evaluation methodology adopted. We also provide a brief summary of the methods employed by participants of this lab and the results obtained across 22 runs from 3 teams, finding that several submissions achieve better results than the baseline methods prepared by the organizers.

Keywords

Chemical patents, Text mining, Information Extraction

1. Introduction

The discovery of new chemical compounds is a key driver of the chemistry and pharmaceutical industries. Patents serve as a critical source of information about new chemical compounds, providing timely and comprehensive information about new chemical compounds [1, 2, 3]. 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 [4, 5]. 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, 7].

The ChEMU campaign focuses on information extraction tasks over chemical reactions in patents. The ChEMU2020 lab [8, 7] provided two information extraction tasks, named entity recognition (NER) and event extraction (EE). The ChEMU 2021 lab [9, 10] introduced one more task, anaphora resolution (AR). This year, we re-run all the three tasks with new test sets. Together, the tasks support comprehensive automatic chemical patent analysis.

In collaboration with chemical domain experts, we have built upon the datasets used in ChEMU 2020/2021 (1500 snippets) and prepared 500 snippets from selected chemical patents that specifically target all three tasks. For the NER and the EE tasks, three chemical experts were hired to manually annotate the corpus, labeling named entities and event steps in these text segments. Two of them reviewed all text segments independently and the third annotator acted as an adjudicator who resolved their disagreements and merged their annotations into the final gold-standard corpus. For the AR task, two chemical experts, a PhD candidate and a final year bachelor student in Chemistry were hired to annotate the same set of snippets. The dataset was first annotated by the two annotators individually, and then their annotations were compared and combined by an adjudicator.

The ChEMU2022 lab has received considerable interest, attracting 54 registrants. Specifically, we received 8 runs from 3 teams in the NER task, 11 runs from 3 teams in the EE task, and 3 runs from 1 team in the AR task, respectively. Several submissions achieved exciting results, with a few of them outperforming baseline models significantly.

The rest of the paper is structured as follows. We first discuss related work and shared tasks in Section 2 and introduce the corpus we created for use in the lab in Section 3. Then we give an overview of the tasks in Section 4 and detail the valuation framework of ChEMU in Section 5

CLEF 2022: Conference and Labs of the Evaluation Forum, September 5–8, 2022, Bologna, Italy
yuan.li1@unimelb.edu.au (Y. Li); biaoyanf@student.unimelb.edu.au (B. Fang); jiayuan.he@rmit.edu.au (J. He); y.hiyori@jp.fujitsu.com (H. Yoshikawa); s.akhondi@elsevier.com (S. A. Akhondi); C.Druckenbrodt@elsevier.com (C. Druckenbrodt); camilo.thorne@gmail.com (C. Thorne); m.afzal.1@elsevier.com (Z. Afzal); zenan.zhai@student.unimelb.edu.au (Z. Zhai); machi@eis.hokudai.ac.jp (K. Machi); yoshioka@ist.hokudai.ac.jp (M. Yoshioka); jyrok3357@lgresearch.ai (Y. Jang); hosung.song@lgresearch.ai (H. Song); junho1126@lgdisplay.com (J. Lee); ghkayne.kim@lgresearch.ai (G. Kim); yireun.kim@lgresearch.ai (Y. Kim); stanleyjk.choi@lgresearch.ai (S. J. Choi); honglak@umich.edu (H. Lee); k.bae@lgresearch.ai (K. Bae); mahendrand@vcu.edu (D. Mahendran); ctang2@vcu.edu (C. Tang); btmcinnes@vcu.edu (B. McInnes); tb@ldwin.net (T. Baldwin); karin.verspoor@rmit.edu.au (K. Verspoor)
0000-0002-8661-1544 (K. Verspoor)

including the evaluation methods and baseline models. We present the evaluation results in Section 7 and finally conclude this paper in Section 8.

2. Related Work

To assess and advance the natural language processing (NLP) techniques in the biochemical domain, many shared tasks/labs have been organized, including n2c2[1], TREC[2], BioCreative[3], BioNLP[4], and CLEF workshops[5]. These shared tasks have covered a range of benchmark text mining tasks: information retrieval, such as document retrieval (CLEF eHealth 2014 [11]) and text classification (CoNLL 2010 [12]); word semantics, such as named entity recognition (BioCreative II [13] Task 1) and mention normalization (BioCreative III [14, 15] Gene Normalization Task); relation semantics, such as event extraction (GENIA Event Extraction [16]) and interaction extraction (Drug-Drug Interaction [17]); and high-level applications, such as question answering (Semantic QA [18]) and document summarization (Biomed-Summ [19]).

Nevertheless, most of these shared tasks/labs did not focus on the domain of chemical patents. These shared tasks mainly focused on the text mining over biomedical texts (e.g., scientific literature, such as PubMed abstracts) or clinical data (e.g., clinical health records). Text mining techniques that are developed for biomedical or biochemical texts, such as scientific journals and clinical records may not be effective for chemical patents. This is because their purpose is distinct—chemical patents are written for protection of intellectual property related to chemical compounds—and their content has different scope and characteristics, including variations in linguistic structures. Thus, it is critical to develop text mining techniques that are tailored for chemical patents.

Only two shared tasks have previously considered chemical patents. TREC 2009 [20] provided a chemical information retrieval track for the tasks of ad hoc retrieval of chemical patents and prior art search. However, this track differs significantly from the subtasks in our ChEMU lab: it addresses document-level retrieval and relevance to queries instead of considering the detailed content of each document. The ChemDNER-patents task [21] at the BioCreative V workshop was the task that is most similar with ours. It aimed at detection of chemical compounds and genes/proteins in patent text. However, the ChemDNER-patents task only considered entity detection within patent abstracts while we consider data extracted from the full texts of patents. Moreover, our definition of chemical compound entities is much richer as our label set defines not only that a chemical or drug compound is mentioned, but also what its specific role is with respect to the chemical reaction that it is related to in the description, e.g., starting material, catalyst, or product.

The ChEMU labs also contribute new corpus on chemical text mining for the research community[6]. Most existing benchmark datasets for biochemical text mining focus on biomedical texts, i.e., texts that consider the interaction of chemicals with molecular biology or human

[1] https://n2c2.dbmi.hms.harvard.edu/
[2] https://trec.nist.gov/
[3] https://biocreative.bioinformatics.udel.edu/
[4] https://2019.bionlp-ost.org/
[5] https://sites.google.com/site/clefehealth/
[6] https://chemu.eng.unimelb.edu.au/

disease. CHEMProt [22] consists of 1,820 PubMed[7] abstracts with chemical-protein interactions, DDI extraction 2013 corpus [17] is a collection of 792 texts selected from the DrugBank database[8] and other 233 PubMed abstracts, and BC5CDR is a collection of 1,500 PubMed titles and abstracts selected from the CTD-Pfizer corpus, just to give a few examples.

The number of public datasets that focus on the chemistry domain is limited. Further, several existing chemical datasets are based on structured/semi-structured texts rather than free, natural language, texts. For example, the ZINC 15 250k corpus[9] is a collection of 250,000 molecules with their Simplified Molecular Input Line Entry System (SMILES) strings. The Tox21 dataset contains roughly 7,000 molecules and typical 120 characteristics, such as atomic number, aromicity, donor status. There are two datasets that are constructed from free patent texts: (1) the dataset released by the ChemDNER patents task and (2) the dataset created by Akhondi et al. [23]. However, these two datasets only contain entity annotations. Our chemical reaction corpus is further enriched by the relations between the annotated entities.

Despite the limited number of shared tasks on chemical patent mining, there is an increasing interest in developing information extraction models for patents in general research communities [24, 2, 25]. Various text mining techniques have been proposed for information extraction over chemical patents [26], addressing fundamental NLP tasks, such as named entity recognition and relation extraction [24, 27, 28, 29]. Early techniques for chemical text mining, such as dictionary-based methods [30, 31, 28] and grammar-based methods [32, 33, 34], heavily rely on expert knowledge in the chemical domain. Recently, machine learning-based techniques have reported state-of-the-art effectiveness in chemical text mining [35, 29]. However, such techniques require a large amount of annotated text data, which still remains limited. Thus, ChEMU lab 2020 was hosted to provide an opportunity for NLP experts to develop information extraction systems over chemical patents. The new ChEMU reaction corpus was also made publicly available to all researchers as an important benchmark dataset for future research in this domain [36].

3. The ChEMU Chemical Reaction Corpus

In this section, we explain how the dataset is created for our shared tasks. The complete annotation guidelines are made available on our website[10].

3.1. Data Selection

The ChEMU chemical reaction corpus was built with the aid of Elsevier Reaxys® database.[11] Reaxys® is a rich information resource for chemical reactions, which contains detailed descriptions of chemical reactions that are extracted via an "excerption" process, i.e., manual selection of information from literature sources, such as patents and scientific publications.

[7] https://pubmed.ncbi.nlm.nih.gov/

[8] https://go.drugbank.com/

[9] https://github.com/aspuru-guzik-group/chemical_vae/tree/master/models/zinc

[10] http://chemu2022.eng.unimelb.edu.au/

[11] Reaxys® Copyright ©2022 Elsevier Life Sciences IP Limited except certain content provided by third parties. Reaxys is a trademark of Elsevier Life Sciences IP Limited, used under license. https://www.reaxys.com

[Step 4] Synthesis of N-((5-(hydrazinecarbonyl)pyridin-2-yl)methyl)-1-methyl-N-phenylpiperidine-4-carboxamide Methyl 6-((1-methyl-N-phenylpiperidine-4-carboxamido)methyl)nicotinate (0.120 g, 0.327 mmol), synthesized in step 3, and hydrazine monohydrate (0.079 mL, 1.633 mmol) were dissolved in ethanol (10 mL) at room temperature, and the solution was heated under reflux for 12 hours, and then cooled to room temperature to terminate the reaction. The reaction mixture was concentrated under reduced pressure to remove the solvent, and the concentrate was purified by column chromatography (SiO2, 4 g cartridge; methanol/dichloromethane = from 5% to 30%) and concentrated to give the title compound (0.115 g, 95.8%) as a foam solid.

Figure 1: An example of one patent snippet in ChEMU chemical reaction corpus.

In ChEMU 2020, we selected 180 English patents from the European Patent Office and the United States Patent and Trademark Office, for which information had been included in the Reaxys database. From these patents, 1500 text segments were sampled from chemical reaction descriptions pre-identified by expert domain annotators, available as a product of the process used to populate information in Reaxys® . We refer to each text segment as a patent "snippet" and use the two expressions interchangeably in the remainder of this paper. The 1500 snippets were annotated for the named entity recognition (NER) and the event extraction (EE) tasks. In ChEMU 2021, we annotated the same 1500 snippets for the anaphora resolution (AR) task. In ChEMU 2022, we further collect 500 snippets from the selected patents and annotate them for all three tasks.

We present an example of a patent snippet in Figure 1. This snippet describes the synthesis of a particular chemical compound, N-((5-(hydrazinecarbonyl)pyridin-2-yl)methyl)-1-methyl-N-phenylpiperidine-4-carboxamide. The synthesis process consists of an ordered sequence of reaction steps:

1. dissolving the chemical compound synthesized in step 3 and hydrazine monohydrate in ethanol;
2. heating the solution under reflux;
3. cooling the solution to room temperature;
4. concentrating the cooled mixture under reduced pressure;
5. purification of the concentrate by column chromatography;
6. concentration of the purified product to get the title compound.

Our shared tasks aim at extraction of chemical reactions from chemical patents, e.g., extracting the above synthesis steps given the patent snippet in Figure 1. To achieve this goal, it is crucial for us to first identify the entities that are involved in these reaction steps (e.g., hydrazine monohydrate and ethanol) and then determine the relations between the involved entities (e.g., hydrazine monohydrate is dissolved in ethanol).

Furthermore, our shared tasks also aim at resolving the reference in the chemical reactions. For example, the *solution* in the second step refers to the title compound (0.120 g, 0.327 mmol),

hydrazine monohydrate (0.079 mL, 1.633 mmol), and ethanol (10 mL).

3.2. Annotation Guidelines

3.2.1. NER Annotations

Four categories of entities are annotated over the corpus: (1) chemical compounds that are involved in a chemical reaction; (2) conditions under which a chemical reaction is carried out; (3) yields obtained for the final chemical product; and (4) example labels that are associated with reaction specifications. Ten labels are further defined under the above four categories. We define five different roles that a chemical compound can play within a chemical reaction, corresponding to five labels under this category: STARTING MATERIAL, REAGENT CATALYST, REACTION PRODUCT, SOLVENT, and OTHER COMPOUND. We also define two labels under the category of conditions: TIME and TEMPERATURE; and two labels under the category of yields: YIELD PERCENT and YIELD OTHER.

The definitions of all resultant labels are summarized as follows:

1. *Reaction product*: A substance that is formed during a chemical reaction.
2. *Starting material*: A substance that is consumed in the course of a chemical reaction providing atoms to products.
3. *Reagent catalyst*: 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.
4. *Solvent*: A chemical entity that dissolves a solute resulting in a solution.
5. *Other compound*: Other chemical compounds that are not the products, starting materials, reagents, catalysts and solvents.
6. *Example label*: A label associated with a reaction specification.
7. *Temperature*: The temperature at which the reaction was carried out.
8. *Time*: The reaction time of the reaction.
9. *Yield percent*: Yield given in percent values.
10. *Yield other*: Yields provided in other units than %.

3.2.2. EE Annotations

A chemical reaction process is usually a sequence of steps, and these steps can be categorized into two types: (1) reaction steps, i.e., the steps required to convert the starting materials to the target reaction product; and (2) work-up steps, i.e., the manipulations required to purify or isolate a chemical product. For example, in Figure 1, the step of heating the solution under reflux for 12 hours is a reaction step while the step of cooling it to room temperature is a work-up step.

We define two types of trigger words: WORKUP which refers to an event step where a chemical compound is isolated/purified, and REACTION STEP which refers to an event step that is involved in the conversion from a starting material to an end product. When labelling event arguments, we adapt semantic argument role labels Arg1 and ArgM from the Proposition Bank to label the relations between the trigger words and other arguments. Specifically, the

label Arg1 refers to 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. 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. The definitions of trigger word types and relation types are summarized as follows:

1. *Workup*: An event step which is a manipulation required to isolate and purify the product of a chemical reaction.
2. *Reaction step*: An event within which starting materials are converted into the product.
3. *Arg1*: The relation between an event trigger word and a chemical compound.
4. *ArgM*: The relation between an event trigger word and a temperature, time, or yield entity.

3.2.3. AR Annotations - Mentions

We aim to capture anaphora in chemical patents, with a focus on identifying chemical compounds during the reaction process. Consistent with other anaphora corpora [37, 38, 39], only mentions that are involved in referring relationships (as defined in Section 3.2.4) and related to chemical compounds are annotated. The mention types that are considered for anaphora annotation are listed below.

1. *Chemical names*: the formal name of chemical compounds.
2. *Identifiers*: identifiers or labels that uniquely represent chemical compounds which occur earlier in the text.
3. *Phrases and noun types*: pronouns that refer to a previously mentioned chemical compounds, e.g. *they* or *it*, and definite and indefinite noun phrases that refer to chemical compounds, e.g. *the solvent*, *the title compound*, *the mixture*, and *a white solid*, *a crude product*.

It should be noted that verbs (e.g. *mix*, *purify*, *distil*) and descriptions that refer to events (e.g. *the same process, step 5*) are not annotated in this corpus.

Unlike many annotation schemes, our annotation allows discontinuous mentions. For example, the underlined spans of the fragment *114 mg of 4-((4aS,7aS)-6-benzyloctahydro-1-pyrrolo[3,4-b]pyridine-1-yl)-7H-pyrrolo[2,3-d]pyrimidine was obtained with a yield of about 99.1%* are treated as a single discontinuous mention. This introduces further complexity into the task and helps to capture more comprehensive anaphora phenomena.

There are some differences in the definitions of entities for the NER task and the AR task. For the NER task, entity annotations identify chemical compounds (i.e. REACTION_PRODUCT, STARTING_MATERIAL, REAGENT_CATALYST, SOLVENT, and OTHER COMPOUND), reaction conditions (i.e. TIME, TEMPERATURE), quantity information (i.e. YIELD_PERCENT, YIELD_OTHER), and example labels (i.e. EXAMPLE_LABEL). There is overlap with our definition of mention for the labels relating to chemical compounds. However, in AR annotation, chemical names are annotated along with additional quantity information, as we consider this information to be an integral part of the chemical compound description. Furthermore, the original entity annotations do not include generic expressions that corefer with chemical

compounds such as *the mixture*, *the organic layer*, or *the filtrate*, and neither do they include equipment descriptions.

3.2.4. AR Annotations - Relation

Anaphora resolution subsumes both coreference and bridging. In the context of chemical patents, we define four sub-types of bridging, incorporating generic and chemical knowledge.

1. *Coreference*: two expressions/mentions that refer to the same entity.
2. *Bridging*:
    a) *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).
    b) *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.
    c) *Work-up*: the relationship between chemical compounds that were used for isolation or purification purposes, and their corresponding output products.
    d) *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.

A referring mention which cannot be interpreted on its own, or an indirect mention, is called an *anaphor*, and the mention which it refers back to is called the *antecedent*. In relation annotation, we preserve the direction of the anaphoric relation, from the anaphor to the antecedent. Following similar assumptions in recent work, we restrict annotations to cases where the antecedent appears earlier in the text than the anaphor.

3.3. Annotation Process

To facilitate the annotation process, a silver standard set was first prepared based on information captured in the Elsevier Reaxys® database. The extracted records from Reaxys® are linked to the IDs of their source patents. However, the precise locations of the key entity and relation information in these records in source patents are needed to construct the gold-standard corpus. The silver-standard dataset was prepared by automatically mapping elements of the records in the Reaxys® database to the source patents from which the records were extracted. This mapping process was performed by scanning patent texts and searching for excerpted entity mentions.

For the NER and the EE tasks, three chemical experts were hired to prepare the gold standard corpus. They manually reviewed all texts and pre-annotations in the silver-standard dataset to add or correct precise locations of the relevant entities and relations in the texts, according to annotation guidelines in Section 3.2.1. Two of the experts first independently reviewed and updated the silver standard annotations. Then, a third chemical expert served as an adjudicator who resolved their disagreements to produce the final gold-standard corpus. For the AR task,

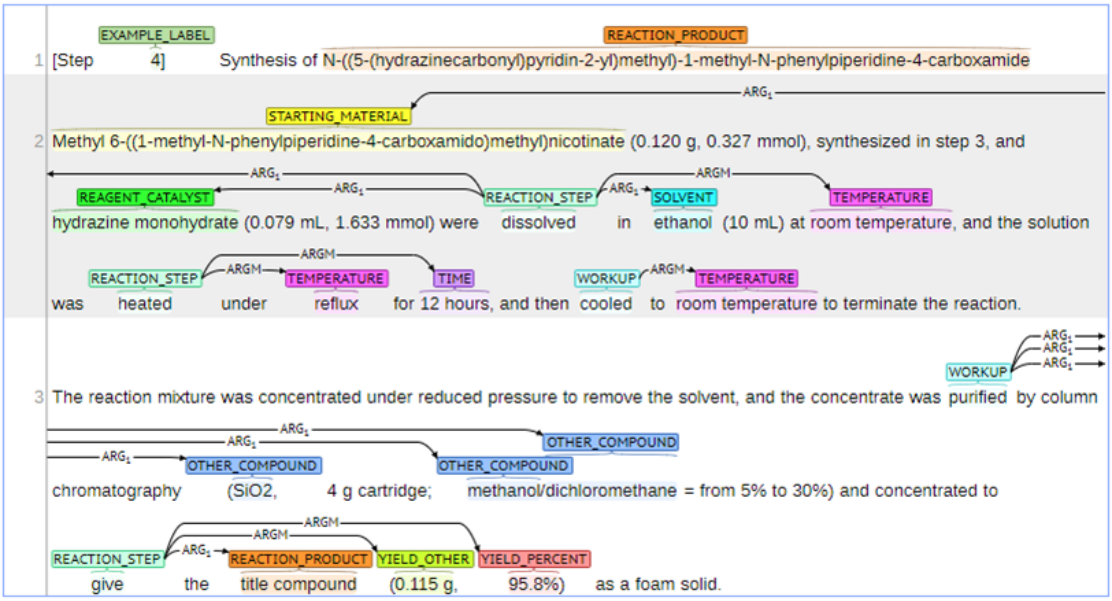


Figure 2: Visualization of the annotations in the snippet in Figure 1 for the NER and the EE tasks.

one of the chemical experts who had annotated for the anaphora resolution task in ChEMU 2021 was hired to annotate the same set of snippets.

The annotation process was conducted using the BRAT annotation tool,[12] which is an interactive web-based tool for adding annotations to input texts. Continuing with the example snippet shown in Figure 1, a visualization of the snippet after annotation is presented in Figure 2 for the NER and the EE tasks, and Figure 3 for the AR task.

3.4. Data Partitions

We combine the training/development/test sets for ChEMU 2020/2021 (1500 snippets) and use it as the training set for ChEMU 2022. The 500 new snippets that we annotated for ChEMU 2022 are used as the test set.

In ChEMU 2020 and 2021, the evaluation results of all submissions to the test set were only available when the shared tasks ended. This year, we run all shared tasks in a Kaggle-style where the test set (500 snippets) is randomly partitioned into two splits public/private with a ratio of 30%/70%, and the participants will get immediate feedback on the public test set (150 snippets) after making a submission, while the evaluation results on the private test set (350 snippets) remain secret until the end of the shared tasks. Note that the participants are not aware of the specific split of public and private test sets.

[12] https://brat.nlplab.org/

Figure 3: Visualization of the annotations in the snippet in Figure 1 for the AR task.

4. Task Definition

The three tasks, named entity recognition, event extraction, and anaphora resolution, are all snippet-level tasks 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.

4.1. Task 1: Named entity recognition

In order to understand and extract a chemical reaction from natural language texts, the first essential step is to identify the entities that are involved in the chemical reaction. The first task aims to accomplish this step by identifying the ten types of entities described in Section 3.2.1. The task requires the detection of the entity names in patent snippets and the assignment of correct labels to the detected entities. For example, given a detected chemical compound, the task requires the identification of both its text span and its specific type according to the role in which it plays within a chemical reaction description.

Table 1
Illustration of three tasks performed on the same snippet (NER, EE, and AR).

Raw 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%)**

4.2. Task 2: Event extraction

A chemical reaction usually consists of an ordered sequence of event steps that transforms a starting product to an end product, such as the six reaction steps in the synthesis process of the chemical compound described in the example in Figure 1. The event extraction task (Task 2) targets identifying these event steps. Similarly to conventional event extraction problems, the EE task involves three subtasks: event trigger word detection, event typing and argument prediction. First, it requires the detection of event trigger words and assignment of correct labels for the trigger words. Second, it requires the determination of argument entities that are associated with the trigger words, i.e., which entities identified in the NER task participate in event or reaction steps. This is done by labelling the connections between event trigger words and their arguments. Given an event trigger word e and a set S of arguments that participate in e, the EE task requires the creation of $|S|$ relation entries connecting e to an argument entity in S. Here, $|S|$ represents the cardinality of the set S. Finally, this task requires the assignment of correct relation type labels (Arg1 or ArgM) to each of the detected relations.

4.3. Task 3: 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, reaction-associated, work-up, contained, and transform.

Taking the text snippet in Figure 4 as an example, several anaphoric relationships can be extracted from it. **[The mixture]**$_4$ and **[the mixture]**$_3$ refer to the same "mixture" and thus, form a coreference 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

[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	Coreference	**[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]**

Figure 4: 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.

[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. Evaluation Framework

5.1. Evaluation Methods

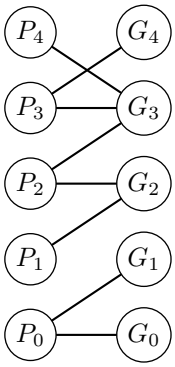
We use BRATEval[13] to evaluate all the runs that we receive. Three metrics are used to evaluate the performance of all the submissions: Precision, Recall, and F_1 score. We use two difference matching criteria, exact matching and relaxed matching (approximate matching), as in some practical applications it also makes sense to understand if the model can identify the *approximate* region of mentions.

Formally, let $E = (ET, A, B)$ denote an entity where ET is the type of E, A and B are the beginning position (inclusive) and end position (exclusive) of the text span of E. Then two entities E_1 and E_2 are exactly matched ($E_1 = E_2$), if $ET_1 = ET_2$, $A_1 = A_2$, and $B_1 = B_2$. While two entities E_1 and E_2 are approximately matched ($E_1 \approx E_2$) if $ET_1 = ET_2$, $A_2 < B_1$, and $A_1 < B_2$, i.e. the two spans $[A_1, B_1)$ and $[A_2, B_2)$ overlaps.

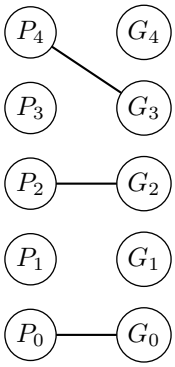
Furthermore, let $R = (RT, E^{ana}, E^{ant})$ be a relation where RT is the type of R, E^{ana} the anaphor of R, E^{ant} the antecedent of R. Then R_1 and R_2 are exactly matched ($R_1 = R_2$) if $RT_1 = RT_2$, $E_1^{ana} = E_2^{ana}$, and $E_1^{ant} = E_2^{ant}$. While R_1 and R_2 are approximately matched ($R_1 \approx R_2$) if $RT_1 = RT_2$, $E_1^{ana} \approx E_2^{ana}$, and $E_1^{ant} \approx E_2^{ant}$.

In summary, we require strict type match in both exact and relaxed matching, but are lenient in span matching.

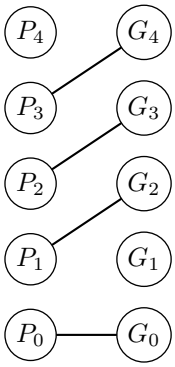
[13] https://bitbucket.org/nicta_biomed/brateval/src/master/



(a) All pairs matched in relaxed setting



(b) A non-optimal bipartite matching ($TP = 3$)



(c) A maximum bipartite matching ($TP = 4$)

Figure 5: An example matching graph and two bipartite matching for it.

5.1.1. Exact Matching

With the above definitions, the metrics for exact matching can be easily calculated. The true positives (TP) are exact matching pairs found in gold relations and predicted relations. Then false positives (FP) are the predicted relations that don't have a match, i.e. $FP = \#pred - TP$, where $\#pred$ is the number of predicted relations. Similarly, false negatives FN are the gold relations that are not matched by any predicted relations, i.e. $FN = \#gold - TP$ where $\#gold$ is the number of gold relations. Finally Precision $P = TP/(TP + FP)$, Recall $R = TP/(TP + FN)$, and $F_1 = 2/(1/P + 1/R)$.

5.1.2. Relaxed Matching

Unlike exact matching, relaxed matching is not well-defined and metrics in this setting have more than one way to calculate, therefore we need to clearly define all the metrics.

Let consider an example shown in Figure 5a where nodes $\{P_i\}_{i=1}^5$ are predicted relations, $\{G_i\}_{i=1}^5$ are gold relations, and every edge between a P node and a G node means they are approximately matched. At first glance, one may think that $FN = FP = 0$ because every gold relation has at least a match and so does every predicted relation. However, it is impossible to find 5 true positive pairs from this graph without using one node more than once. Therefore, if $FN = FP = 0$, then $FN + TP \neq \#gold = 5$ and $FP + TP \neq \#pred = 5$, which is inconsistent with the formulas in exact setting.

So, instead of defining FN as the number of gold relations that don't have a match, we just define $FN = \#gold - TP$. Similarly FP is defined as $\#pred - TP$. Then the problem remained is how to calculate TP. Actually, finding true positive pairs can be considered as bipartite matching. Figure 5b shows a matching with $TP = 3$ but is not optimal. Figure 5c shows one possible maximum bipartite matching with $TP = 4$. Another optimal matching is replacing edge $P_0 - G_0$ with $P_0 - G_1$.

In summary, we define TP as the maximum bipartite matching for the graph constructed by all approximately matched pairs, then $FN = \#gold - TP$ and $FP = \#pred - TP$, finally

Precision $P = TP/(TP + FP)$, Recall $R = TP/(TP + FN)$, and $F_1 = 2/(1/P + 1/R)$. This has been implemented in the latest BRATEval.

5.2. Coreference Linkings in Anaphora Resolution Task

We consider two types of coreference linking, i.e. (1) surface coreference linking and (2) atomic coreference linking, due to the existence of *transitive coreference relationships*. By transitive coreference relationships we mean multi-hop coreference such as a link from an expression T1 to T3 via an intermediate expression T2, viz., "T1→T2→T3". Surface coreference linking will restrict attention to one-hop relationships, viz., to: "T1→T2" and "T2→T3". Whereas atomic coreference linking will tackle coreference 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 coreference relationships.

Note that {T1→T2,T2→T3} infers {T1→T3,T2→T3}, but the reverse is not true. This leads to a problem about how to score a prediction {T1→T3,T2→T3}, when the gold relation is {T1→T2,T2→T3}. Both T1→T3 and T2→T3 are true, but some information is missing here.

Our solution is to first expand both the prediction set and gold set where all valid relations that can be inferred will be generated and added to the set, and then to evaluate the two sets normally. In the above example, the gold set will be expanded to {T1→T2,T2→T3,T1→T3}, and then the result is $TP = 2$, $FN = 1$. Likewise, when evaluate {T1→T4,T2→T4,T3→T4} against {T1→T2,T2→T3,T3→T4}, the gold set will be expanded into 6 relations, while the prediction set won't be expanded as no new relation can be inferred. So the evaluation result will be $TP = 3$, $FN = 3$. One may worry that if there is a chain of length n then its expanded set will be in $O(n^2)$, when n is large, this local evaluation result will have too much influence on the overall result. But we find in practice that coreference chains are relatively short, with 3 or 4 being the most typical lengths, so it is unlikely to be a big issue.

5.3. Baselines

5.4. NER and EE Baseline

We use a joint model for recognizing named entities and classifying relations between them. The model first processes the input snippet using a BERT model to obtain the contextualized word representations. We adopt the BIO tagging schema for training the NER classifier which classifies every word into entity tags. Then a list of identified entities is created based on the output of the NER classifier. For each entity in the list, the contextualized word representations are max pooled to obtain the representation for the entity. Then the model enumerates all possible pairs of entities and provides them to a relation classifier which classifies every pair of entities by concatenating the representations of both entities.

5.4.1. AR Baseline

Our baseline model adopts an end-to-end architecture for coreference resolution [41, 42], as depicted in Figure 6. Following the methods presented in [40], we use GloVe embeddings and a

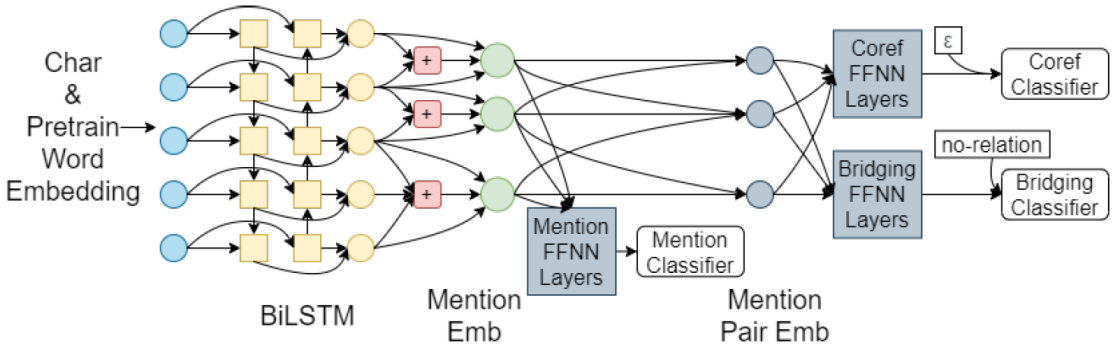


Figure 6: The architecture of our baseline model for Task 3: Anaphora Resolution. This figure is taken from [40].

character-level CNN as input to a BiLSTM to obtain contextualized word representations. Then all possible spans are enumerated and fed to a mention classifier which detects if the input is a mention. Based on the same mention representations, pairs of mentions are fed to a coreference classifier and a bridging classifier, where the coreference classifier does binary classification and the bridging one classifies pairs into 4 bridging relation types and a special class for no relation. Training is done jointly with all losses added together.

6. Overview of Participants' Approaches

We received paper submissions from all the participating teams, i.e. the LG team, the HUKB team, and the VCU team.

6.1. LG Team

The LG team developed context-aware NER and RE models based on the domain-specific language model with pipeline approach. For the domain-specific language model, they post-train the BiolinkBert[43] model with various chemical corpora and pre-processing methods, then select the best performing model from domain-specific benchmark datasets consisting of BLURB (Biomedical Language Understanding & Reasoning Benchmark)[44] and ChEMU 2020[7]. Based on this language model, they develop the NER model to predict both the entity and trigger word, and the RE model to predict the relation between them. Among the pipeline approach and the joint approach, they choose the pipeline approach because PURE[45] reports it gets higher performance than the joint approach. For the NER model, they experiment with two popular approaches, the sequence tagging approach[46, 47] and the span-based approach[45, 48]. Finally, they choose the sequence tagging approach that shows higher performance in the NER task. For the RE model, they train the model to classify the relation types or no relation between every pair of trigger word and entity in the snippet. Furthermore, they train both models using inputs that contain multiple sentences rather than a single sentence so that the model can utilize contextual information. For the ensemble, they train the best performing model with 10-fold cross validation and then predict the results with soft-voting. Finally, they apply rule-based

post-processing to the prediction results. Their best public exact match f1 score of task 1a and 1b was 96.26 and 92.56, respectively, before the submission deadline. After that, they further experimented with various post-processing and the improved final scores are 96.33 and 92.82

6.2. HUKB Team

The HUKB team participated in all three tasks. For the NER task, they used ChemBERT[49], a pre-trained language model for chemistry-related documents, and a set of post-processing rules that considers document-level context. First, ChemBERT predicted mentions using a sentence as the input. Then, two post-processing methods were applied. They submitted a result and it obtained an exact match F-score of 0.9412 and a relaxed match F-score of 0.9572.

For the EE task, they adopted a pipeline approach for relation extraction. In addition to named entities detected by the NER task, they also used ChemBERT for event detection as the NER task. After detecting those mentions, they also used ChemBERT for relation classification. In this classification process, all candidate pairs in a sentence were classified and positive relations were annotated by the system. They fine-tuned one ChemBERT model for ARG1 and a second for ARGM relations. In the training stage, they used not only gold-standard entities but also predicted events that were generated by five systems trained on 80% of the training set, similarly to five-fold cross-validation. They submitted a result and it obtained an exact match F-score of 0.8868 and a relaxed match F-score of 0.9028. They also submitted a result after the evaluation phase that corrected the sequence length of an input and the result obtained an exact match F-score of 0.8865 and a relaxed match F-score of 0.9027.

For the AR task, they adopted a pipeline approach for relation extraction, where the candidates of mentions are first detected and then relations between them are classified. In addition, they used a set of post-processing rules that considers document-level context. For mention detection, they fine-tuned one Chem-BERT model for coreference and a second for bridging relations. They augmented the number of positive examples for coreference relations in the training set by reusing sentences that contained one or more mentions five times because the number of mentions in the training set for coreference was smaller than for other datasets used for mention detection. For relation classification, Chem-BERT detected a relation for each pair of mentions in two continuous sentences. They trained two models for coreference and bridging relations. They applied two post-processing methods for each relations that consider document-level context. Both of them used results in the NER task; for example, REACTION PRODUCT in the NER task was used for the detection of corefenrence relations. They submitted three results: (1) without post-processing rules, (2) with a post-processing rule for bridging relations and (3) with post-processing rules both coreference and bridging relations. The third result obtained the best performance that was an exact match F-score of 0.7085 and an F-score of 0.7893. They also submitted a result after the evaluation phase that corrected the sequence length of an input and the result obtained an exact match F-score of 0.8865 and a relaxed match F-score of 0.9027.

6.3. VCU Team

The VCU team participated in the NER and the EE tasks. For the NER task, they evaluated two methods for identifying the experimental parameters and triggers. The first was a transformer-

based architecture and the second a bidirectional Long Short Term Memory (biLSTM) units, both with a Conditional Random Field (CRF) output layer. The input to their biLSTM model were pre-trained word embeddings[8] in combination with character embeddings. These embeddings are concatenated and then passed through the network. The input to their transformer model used Bidirectional Encoder Representations from Transformers (BERT)[46]. The BERT embeddings are then passed to an additional transformer encoder layer and a CRF output layer which predicts a sequence of labels corresponding to entity types. The results showed that our biLSTM model obtained higher scores overall with a exact precision, recall and F1 score of 0.73, 0.81, and 0.77 respectively, and relaxed precision, recall and F1 score of 0.83, 0.92, and 0.87 respectively.

They treat the EE task as a binary classification task building a separate model for each trigger word-entity type to determine whether a relationship exists between them. They first identify the sentence where the trigger word-entity pair is located, then they replace the non-targeted trigger word-entity pairs with 'X' from the input sentence except for the targeted trigger word-entity pair. Here, BERT captures the contextual information within a sentence, whereas GCN captures the global information. They use the BERT tokenizer for word tokenization, and then they generate a vocabulary graph $G = (V, E)$ where the word nodes in the graph are denoted by mapped integers. Next, the combined input of mapped word indices with the generated graph embeddings is passed through BERT, and the final embedding representation is fed into a fully connected layer for classification. This method obtained an overall with an exact precision, recall and F1 score of 0.82, 0.68, and 0.75 respectively, and a relaxed precision, recall and F1 score of 0.88, 0.73, and 0.79 respectively.

7. Results and Discussions

A total of 54 participants registered on our submission website for the shared tasks. Among them, we finally received 22 submissions from 3 teams on the test set. The 3 teams are LG AI Research (LG), Hokkaido University (HUKB), and Virginia Commonwealth University (VCU). In this section, we report their results along with the performance of our baseline systems.

7.1. Task 1: Named Entity Recognition

We report the overall performance of all runs in Table 2. The baseline achieves 0.9367 in F1-score under exact-match. Four runs outperform the baseline in terms of F1-score under exact-match. The best run was submitted by team LG AI Research, achieving a high F1-score of 0.9673. The F1-scores for submissions from team VCU in relaxed match are 10%-15% higher than those in exact-match. This difference between exact-match and relaxed-match may be related to the long text spans of chemical compounds, which is one of the main challenges in NER tasks in the domain of chemical documents.

7.2. Task 2: Event Extraction

The overall performance of all runs is summarized in Table 3 in terms of Precision, Recall, and F1-score under both exact-match and relaxed-match. The rankings of different systems are

Table 2
Overall performance of all runs in Task 1 Named Entity Recognition on private test set. Here, P, R, and F represents the Precision, Recall, and F1-score, respectively. For each metric, we highlight the best result in bold. The results are ordered by their performance in terms of F1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
LG-run1	**0.9663**	**0.9683**	**0.9673**	**0.9782**	**0.9803**	**0.9793**
LG-run2	0.9627	0.9655	0.9641	0.9758	0.9787	0.9772
LG-run3	0.9628	0.9652	0.964	0.9758	0.9782	0.977
HUKB	0.9401	0.9422	0.9412	0.9561	0.9583	0.9572
Baseline	0.947	0.9267	0.9367	0.964	0.9432	0.9535
VCU-run1	0.7335	0.8072	0.7686	0.8345	0.9185	0.8745
VCU-run2	0.734	0.7501	0.742	0.8802	0.8996	0.8898
VCU-run3	0.695	0.7869	0.7381	0.7944	0.8994	0.8436
VCU-run4	0.7263	0.7501	0.738	0.8726	0.9012	0.8867

Table 3
Overall performance of all runs in Task 2 Event Extraction on private test set. Here, P, R, and F represents the Precision, Recall, and F1-score, respectively. For each metric, we highlight the best result in bold. The results are ordered by their performance in terms of F1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
LG-run1	**0.9258**	0.9141	**0.9199**	0.9403	0.9284	0.9343
LG-run2	0.9251	**0.9147**	0.9198	**0.9416**	**0.9309**	**0.9362**
LG-run3	0.9241	0.9129	0.9185	0.9403	0.929	0.9346
LG-run4	0.9234	0.9135	0.9184	0.9398	0.9298	0.9348
LG-run5	**0.9258**	0.907	0.9163	0.942	0.9229	0.9323
Baseline	0.9087	0.9089	0.9088	0.9244	0.9246	0.9245
HUKB	0.9058	0.8685	0.8868	0.9222	0.8842	0.9028
VCU-run1	0.8249	0.6831	0.7473	0.8771	0.7264	0.7946
VCU-run2	0.826	0.6776	0.7445	0.8775	0.7199	0.7909
VCU-run3	0.7533	0.6883	0.7193	0.8015	0.7323	0.7653
VCU-run4	0.64	0.6238	0.6318	0.703	0.6852	0.694
VCU-run5	0.2675	0.6263	0.3749	0.3075	0.7199	0.4309

almost fully consistent across all metrics. Our baseline obtains 0.9088 F1-score under exact-match and the best run is from team LG AI Research which archieves 0.9199 F1-score under exact-match. The performance gap between our baseline and the best run indicates the difficulty of the event extraction task comparing to the NER task. We also notice that recall scores of most runs are consistently lower than their precision scores, which may reveal that the task of identifying a relation from a chemical patent is harder than the task of typing an identified relation.

Table 4
Overall performance of all runs in Task 3 Anaphora Resolution on private test set. Here, P, R, and F represents the Precision, Recall, and F1-score, respectively. For each metric, we highlight the best result in bold. The results are ordered by their performance in terms of F1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
HUKB-run1	0.6876	**0.7307**	**0.7085**	0.766	**0.814**	**0.7893**
HUKB-run2	0.729	0.6838	0.7057	0.8107	0.7604	0.7848
HUKB-run3	0.7393	0.6616	0.6983	**0.8222**	0.7358	0.7766
Baseline	**0.7418**	0.6398	0.687	0.7867	0.6784	0.7286

7.3. Task 3: Anaphora Resolution

The evaluation results of all submission to the anaphora resolution task are shown in Table 4. The first run from the Hokkaido University team achieves an F1-score of 0.7085 in exact-match, outperforming our baseline which gets 0.687. The lead of the best run is even larger in relaxed matching, with an F1-score of 0.7893, about 6 points higher than our baseline. This shows the potential of the model built by the Hokkaido University team and indicates that the performance in exact matching may be further boosted if the boundary errors of their model could be corrected in a post-processing step. Our baseline has higher precision in the exact setting, which indicates that our model is more conservative and could possibly be enhanced by making more aggressive predictions to improve recall.

8. Conclusions

This paper presents a general overview of the activities and outcomes of the ChEMU 2022 evaluation lab. As the third instance of our ChEMU lab series, ChEMU 2022 targets three tasks focusing on chemical reaction information extraction from chemical patents. The evaluation result includes different approaches to tackling the shared task, with several submissions outperforming our baseline methods. We look forward to fruitful discussion and deeper understanding of the methodological details of these submissions at the workshop.

Acknowledgments

We are grateful for the detailed excerption and annotation work of the domain experts that support Reaxys, and the support of Ivan Krstic, Director of Chemistry Solutions at Elsevier. Funding for the ChEMU project is provided by an Australian Research Council Linkage Project, project number LP160101469, and Elsevier. We acknowledge the support of annotators for the anaphora resolution task, Dr. Sacha Novakovic and Colleen Hui Shiuan Yeow at the University of Melbourne.

References

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

[2] S. A. Akhondi, H. Rey, M. Schwörer, M. Maier, J. Toomey, H. Nau, G. Ilchmann, M. Sheehan, M. Irmer, C. Bobach, et al., Automatic identification of relevant chemical compounds from patents, Database 2019 (2019).

[3] 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.

[4] 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.

[5] M. Hu, D. Cinciruk, J. M. Walsh, Improving automated patent claim parsing: Dataset, system, and experiments, arXiv preprint arXiv:1605.01744 (2016).

[6] M. Krallinger, F. Leitner, O. Rabal, M. Vazquez, J. Oyarzabal, A. Valencia, CHEMDNER: The drugs and chemical names extraction challenge, Journal of Cheminformatics 7 (2015) 1–11.

[7] J. He, D. Q. Nguyen, S. A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, Z. Afzal, Z. Zhai, B. Fang, H. Yoshikawa, A. Albahem, L. Cavedon, T. Cohn, T. Baldwin, K. Verspoor, ChEMU 2020: Natural language processing methods are effective for information extraction from chemical patents, Frontiers Res. Metrics Anal. 6 (2021) 654438.

[8] D. Q. Nguyen, Z. Zhai, H. Yoshikawa, B. Fang, C. Druckenbrodt, C. Thorne, R. Hoessel, S. A. Akhondi, T. Cohn, T. Baldwin, K. Verspoor, ChEMU: Named entity recognition and event extraction of chemical reactions from patents, in: Advances in Information Retrieval - 42nd European Conference on IR Research, ECIR 2020, Lisbon, Portugal, April 14-17, 2020, Proceedings, Part II, volume 12036 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 572–579.

[9] J. He, B. Fang, H. Yoshikawa, Y. Li, S. A. Akhondi, C. Druckenbrodt, C. Thorne, Z. Afzal, Z. Zhai, L. Cavedon, T. Cohn, T. Baldwin, K. Verspoor, 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, volume 12657 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 608–615.

[10] Y. Li, B. Fang, J. He, H. Yoshikawa, S. A. Akhondi, C. Druckenbrodt, C. Thorne, Z. Afzal, Z. Zhai, T. Baldwin, K. Verspoor, Overview of ChEMU 2021: Reaction reference resolution and anaphora resolution in chemical patents, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction - 12th International Conference of the CLEF Association, CLEF 2021, volume 12880 of *Lecture Notes in Computer Science*, Springer, 2021, pp. 292–307.

[11] L. Kelly, L. Goeuriot, H. Suominen, T. Schreck, G. Leroy, D. L. Mowery, S. Velupillai, W. W. Chapman, D. Martinez, G. Zuccon, et al., Overview of the share/clef ehealth evaluation lab 2014, in: International Conference of the Cross-Language Evaluation Forum for European Languages, Springer, 2014, pp. 172–191.

[12] R. Farkas, V. Vincze, G. Móra, J. Csirik, G. Szarvas, The conll-2010 shared task: learning to

detect hedges and their scope in natural language text, in: Proceedings of the fourteenth conference on computational natural language learning–Shared task, 2010, pp. 1–12.
[13] A. A. Morgan, Z. Lu, X. Wang, A. M. Cohen, J. Fluck, P. Ruch, A. Divoli, K. Fundel, R. Leaman, J. Hakenberg, et al., Overview of biocreative ii gene normalization, Genome biology 9 (2008) 1–19.
[14] C. N. Arighi, Z. Lu, M. Krallinger, K. B. Cohen, W. J. Wilbur, A. Valencia, L. Hirschman, C. H. Wu, Overview of the biocreative iii workshop, BMC bioinformatics 12 (2011) 1–9.
[15] Z. Lu, H.-Y. Kao, C.-H. Wei, M. Huang, J. Liu, C.-J. Kuo, C.-N. Hsu, R. T.-H. Tsai, H.-J. Dai, N. Okazaki, et al., The gene normalization task in biocreative iii, BMC bioinformatics 12 (2011) 1–19.
[16] J.-D. Kim, Y. Wang, Y. Yasunori, The genia event extraction shared task, 2013 edition-overview, in: Proceedings of the BioNLP Shared Task 2013 Workshop, 2013, pp. 8–15.
[17] M. Herrero-Zazo, I. Segura-Bedmar, P. Martínez, T. Declerck, The ddi corpus: An annotated corpus with pharmacological substances and drug–drug interactions, Journal of biomedical informatics 46 (2013) 914–920.
[18] G. Tsatsaronis, M. Schroeder, G. Paliouras, Y. Almirantis, I. Androutsopoulos, É. Gaussier, P. Gallinari, T. Artières, M. R. Alvers, M. Zschunke, A. N. Ngomo, Bioasq: A challenge on large-scale biomedical semantic indexing and question answering, in: Information Retrieval and Knowledge Discovery in Biomedical Text, Papers from the 2012 AAAI Fall Symposium, Arlington, Virginia, USA, November 2-4, 2012, volume FS-12-05 of *AAAI Technical Report*, AAAI, 2012. URL: http://www.aaai.org/ocs/index.php/FSS/FSS12/paper/view/5600.
[19] K. Jaidka, M. K. Chandrasekaran, S. Rustagi, M.-Y. Kan, Overview of the cl-scisumm 2016 shared task, in: Proceedings of the joint workshop on bibliometric-enhanced information retrieval and natural language processing for digital libraries (BIRNDL), 2016, pp. 93–102.
[20] M. Lupu, J. Zhao, J. Huang, H. Gurulingappa, J. Fluck, M. Zimmermann, I. V. Filippov, J. Tait, Overview of the trec 2011 chemical ir track., in: TREC, 2011.
[21] M. Krallinger, O. Rabal, A. Lourenço, M. P. Perez, G. P. Rodriguez, M. Vazquez, F. Leitner, J. Oyarzabal, A. Valencia, Overview of the chemdner patents task, in: Proceedings of the fifth BioCreative challenge evaluation workshop, 2015, pp. 63–75.
[22] M. Krallinger, O. Rabal, S. A. Akhondi, M. P. Pérez, J. Santamaría, G. P. Rodríguez, G. Tsatsaronis, A. Intxaurrondo, J. A. López, U. Nandal, et al., Overview of the biocreative vi chemical-protein interaction track, in: Proceedings of the sixth BioCreative challenge evaluation workshop, volume 1, 2017, pp. 141–146.
[23] S. A. Akhondi, A. G. Klenner, C. Tyrchan, A. K. Manchala, K. Boppana, D. Lowe, M. Zimmermann, S. A. Jagarlapudi, R. Sayle, J. A. Kors, et al., Annotated chemical patent corpus: a gold standard for text mining, PloS one 9 (2014) e107477.
[24] Y.-H. Tseng, C.-J. Lin, Y.-I. Lin, Text mining techniques for patent analysis, Information processing & management 43 (2007) 1216–1247.
[25] H. Yoshikawa, D. Q. Nguyen, Z. Zhai, C. Druckenbrodt, C. Thorne, S. A. Akhondi, T. Baldwin, K. Verspoor, Detecting chemical reactions in patents, in: Proc. 17th Annual Workshop of the Australasian Language Technology Association, ALTA 2019, Sydney, Australia, December 4-6, 2019, 2019, pp. 100–110.
[26] M. Krallinger, O. Rabal, A. Lourenco, J. Oyarzabal, A. Valencia, Information retrieval and text mining technologies for chemistry, Chemical reviews 117 (2017) 7673–7761.

[27] M. Vazquez, M. Krallinger, F. Leitner, A. Valencia, Text mining for drugs and chemical compounds: methods, tools and applications, Molecular Informatics 30 (2011) 506–519.
[28] S. A. Akhondi, E. Pons, Z. Afzal, H. van Haagen, B. F. Becker, K. M. Hettne, E. M. van Mulligen, J. A. Kors, Chemical entity recognition in patents by combining dictionary-based and statistical approaches, Database 2016 (2016).
[29] Z. Zhai, D. Q. Nguyen, S. A. Akhondi, C. Thorne, C. Druckenbrodt, T. Cohn, M. Gregory, K. Verspoor, Improving chemical named entity recognition in patents with contextualized word embeddings, arXiv preprint arXiv:1907.02679 (2019).
[30] D. Rebholz-Schuhmann, H. Kirsch, M. Arregui, S. Gaudan, M. Riethoven, P. Stoehr, Ebimed—text crunching to gather facts for proteins from medline, Bioinformatics 23 (2007) e237–e244.
[31] K. M. Hettne, R. H. Stierum, M. J. Schuemie, P. J. Hendriksen, B. J. Schijvenaars, E. M. v. Mulligen, J. Kleinjans, J. A. Kors, A dictionary to identify small molecules and drugs in free text, Bioinformatics 25 (2009) 2983–2991.
[32] M. Narayanaswamy, K. Ravikumar, K. Vijay-Shanker, A biological named entity recognizer, in: Biocomputing 2003, World Scientific, 2002, pp. 427–438.
[33] H. Liu, T. Christiansen, W. A. Baumgartner, K. Verspoor, Biolemmatizer: a lemmatization tool for morphological processing of biomedical text, Journal of biomedical semantics 3 (2012) 1–29.
[34] S. A. Akhondi, K. M. Hettne, E. Van Der Horst, E. M. Van Mulligen, J. A. Kors, Recognition of chemical entities: combining dictionary-based and grammar-based approaches, Journal of cheminformatics 7 (2015) 1–11.
[35] W. Hemati, A. Mehler, Lstmvoter: chemical named entity recognition using a conglomerate of sequence labeling tools, Journal of cheminformatics 11 (2019) 1–7.
[36] K. Verspoor, D. Q. Nguyen, S. A. Akhondi, C. Druckenbrodt, C. Thorne, R. Hoessel, J. He, Z. Zhai, ChEMU dataset for information extraction from chemical patents, Mendeley Data 2 (2020) 10–17632.
[37] S. Pradhan, A. Moschitti, N. Xue, O. Uryupina, Y. Zhang, Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes, in: S. Pradhan, A. Moschitti, N. Xue (Eds.), Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning - Proceedings of the Shared Task: Modeling Multilingual Unrestricted Coreference in OntoNotes, EMNLP-CoNLL 2012, July 13, 2012, Jeju Island, Korea, ACL, 2012, pp. 1–40. URL: https://www.aclweb.org/anthology/W12-4501/.
[38] K. B. Cohen, A. Lanfranchi, M. J. Choi, M. Bada, W. A. B. Jr., N. Panteleyeva, K. Verspoor, M. Palmer, L. E. Hunter, Coreference annotation and resolution in the colorado richly annotated full text (CRAFT) corpus of biomedical journal articles, BMC Bioinform. 18 (2017) 372:1–372:14. URL: https://doi.org/10.1186/s12859-017-1775-9. doi:10.1186/s12859-017-1775-9.
[39] A. Ghaddar, P. Langlais, Wikicoref: An english coreference-annotated corpus of wikipedia articles, in: N. Calzolari, K. Choukri, T. Declerck, S. Goggi, M. Grobelnik, B. Maegaard, J. Mariani, H. Mazo, A. Moreno, J. Odijk, S. Piperidis (Eds.), Proceedings of the Tenth International Conference on Language Resources and Evaluation LREC 2016, Portorož, Slovenia, May 23-28, 2016, European Language Resources Association (ELRA), 2016. URL: http://www.lrec-conf.org/proceedings/lrec2016/summaries/192.html.

[40] B. Fang, C. Druckenbrodt, S. A. Akhondi, J. He, T. Baldwin, K. Verspoor, 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, 2021.

[41] K. Lee, L. He, M. Lewis, L. Zettlemoyer, End-to-end neural coreference resolution, in: M. Palmer, R. Hwa, S. Riedel (Eds.), Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing, EMNLP 2017, Copenhagen, Denmark, September 9-11, 2017, Association for Computational Linguistics, 2017, pp. 188–197. URL: https://doi.org/10.18653/v1/d17-1018. doi:10.18653/v1/d17-1018.

[42] K. Lee, L. He, L. Zettlemoyer, Higher-order coreference resolution with coarse-to-fine inference, in: M. A. Walker, H. Ji, A. Stent (Eds.), Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT, New Orleans, Louisiana, USA, June 1-6, 2018, Volume 2 (Short Papers), Association for Computational Linguistics, 2018, pp. 687–692. URL: https://doi.org/10.18653/v1/n18-2108. doi:10.18653/v1/n18-2108.

[43] M. Yasunaga, J. Leskovec, P. Liang, Linkbert: Pretraining language models with document links, in: S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 8003–8016. URL: https://aclanthology.org/2022.acl-long.551.

[44] Y. Gu, R. Tinn, H. Cheng, M. Lucas, N. Usuyama, X. Liu, T. Naumann, J. Gao, H. Poon, Domain-specific language model pretraining for biomedical natural language processing, ACM Trans. Comput. Heal. 3 (2022) 2:1–2:23. URL: https://doi.org/10.1145/3458754. doi:10.1145/3458754.

[45] Z. Zhong, D. Chen, A frustratingly easy approach for entity and relation extraction, in: K. Toutanova, A. Rumshisky, L. Zettlemoyer, D. Hakkani-Tür, I. Beltagy, S. Bethard, R. Cotterell, T. Chakraborty, Y. Zhou (Eds.), Proceedings of the 2021 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2021, Online, June 6-11, 2021, Association for Computational Linguistics, 2021, pp. 50–61. URL: https://doi.org/10.18653/v1/2021.naacl-main.5. doi:10.18653/v1/2021.naacl-main.5.

[46] J. Devlin, M. Chang, K. Lee, K. Toutanova, BERT: pre-training of deep bidirectional transformers for language understanding, in: J. Burstein, C. Doran, T. Solorio (Eds.), Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, NAACL-HLT 2019, Minneapolis, MN, USA, June 2-7, 2019, Volume 1 (Long and Short Papers), Association for Computational Linguistics, 2019, pp. 4171–4186. URL: https://doi.org/10.18653/v1/n19-1423. doi:10.18653/v1/n19-1423.

[47] J. Wang, Y. Ren, Z. Zhang, Y. Zhang, Melaxtech: A report for CLEF 2020 - ChEMU task of chemical reaction extraction from patent, in: L. Cappellato, C. Eickhoff, N. Ferro, A. Névéol (Eds.), Working Notes of CLEF 2020 - Conference and Labs of the Evaluation Forum, Thessaloniki, Greece, September 22-25, 2020, volume 2696 of *CEUR Workshop Proceedings*, CEUR-WS.org, 2020. URL: http://ceur-ws.org/Vol-2696/paper_238.pdf.

[48] D. Ye, Y. Lin, P. Li, M. Sun, Packed levitated marker for entity and relation extraction, in:

S. Muresan, P. Nakov, A. Villavicencio (Eds.), Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), ACL 2022, Dublin, Ireland, May 22-27, 2022, Association for Computational Linguistics, 2022, pp. 4904–4917. URL: https://aclanthology.org/2022.acl-long.337.

[49] J. Guo, A. S. Ibanez-Lopez, H. Gao, V. Quach, C. W. Coley, K. F. Jensen, R. Barzilay, Automated chemical reaction extraction from scientific literature, J. Chem. Inf. Model. 62 (2022) 2035–2045. URL: https://doi.org/10.1021/acs.jcim.1c00284. doi:10.1021/acs.jcim.1c00284.