Overview of ChEMU 2020: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents

Jiayuan He[1], Dat Quoc Nguyen[1,4], Saber A. Akhondi[7], Christian Druckenbrodt[6], Camilo Thorne[6], Ralph Hoessel[2], Zubair Afzal[7], Zenan Zhai[1], Biaoyan Fang[1], Hiyori Yoshikawa[1,5], Ameer Albahem[3], Lawrence Cavedon[3], Trevor Cohn[1], Timothy Baldwin[1], and Karin Verspoor[1(✉)]

[1] The University of Melbourne, Melbourne, Australia
{estrid.he,hiyori.yoshikawa,trevor.cohn,
tbaldwin,karin.verspoor}@unimelb.edu.au
{zenan.zhai,biaoyanf}@student.unimelb.edu.au
[2] Elsevier, Amsterdam, The Netherlands
r.hoessel@elsevier.com
[3] RMIT University, Melbourne, Australia
{ameer.albahem,lawrence.cavedon}@rmit.edu.au
[4] VinAI Research, Hanoi, Vietnam
v.datnq9@vinai.io
[5] Fujitsu Laboratories Ltd., Kawasaki, Japan
[6] Elsevier Information Systems GmbH, Frankfurt, Germany
{c.druckenbrodt,c.thorne.1}@elsevier.com
[7] Elsevier BV, Amsterdam, The Netherlands
{s.akhondi,m.afzal.1}@elsevier.com

Abstract. In this paper, we provide an overview of the Cheminformatics Elsevier Melbourne University (ChEMU) evaluation lab 2020, part of the Conference and Labs of the Evaluation Forum 2020 (CLEF2020). The ChEMU evaluation lab focuses on information extraction over chemical reactions from patent texts. Using the ChEMU corpus of 1500 "snippets" (text segments) sampled from 170 patent documents and annotated by chemical experts, we defined two key information extraction tasks. Task 1 addresses chemical named entity recognition, the identification of chemical compounds and their specific roles in chemical reactions. Task 2 focuses on event extraction, the identification of reaction steps, relating the chemical compounds involved in a chemical reaction. Herein, we describe the resources created for these tasks and the evaluation methodology adopted. We also provide a brief summary of the participants of this lab and the results obtained across 46 runs from 11 teams, finding that several submissions achieve substantially better results than our baseline methods.

Keywords: Named entity recognition · Event extraction · Information extraction · Chemical reactions · Patent text mining

A. Arampatzis et al. (Eds.): CLEF 2020, LNCS 12260, pp. 237–254, 2020.
https://doi.org/10.1007/978-3-030-58219-7_18

1 Introduction

The discovery of new chemical compounds and their synthesis processes is of great importance to the chemical industry. Patent documents contain critical and timely information about newly discovered chemical compounds, providing a rich resource for chemical research in both academia and industry. Chemical patents are often the initial venues where a new chemical compound is disclosed. Only a small proportion of chemical compounds are ever published in journals and these publications can be delayed by up to 3 years after the patent disclosure [5,15]. In addition, chemical patent documents usually contain unique information, such as reaction steps and experimental conditions for compound synthesis and mode of action. These details are crucial for the understanding of compound prior art, and provide a means for novelty checking and validation [3,4]. Due to the high volume of chemical patents [11], approaches that enable automatic information extraction from these patents are in demand. Natural language processing methods are core to meeting the need for large-scale mining of chemical information from patent texts.

The ChEMU (Cheminformatics Elsevier Melbourne University) lab provides participants with opportunities to develop automated approaches for information extraction from chemical reactions in chemical patents. The ChEMU 2020 lab, first introduced in Nguyen et al. (2020) [12], was the first running of ChEMU. Specifically, we provided two information extraction tasks. The first task, named entity recognition, requires identification of essential elements of a chemical reaction, including compounds, conditions and yields, and their specific roles in the reaction. The second task, event extraction, requires the identification of specific event steps that are involved in a chemical reaction. In collaboration with chemical domain experts, we have prepared a high-quality annotated data set of 1,500 segments of chemical patent texts specifically targeting these two tasks.

The rest of the paper is structured as follows. We first introduce the corpus we created for use in the lab in Sect. 2. Then we give an overview of the tasks in Sect. 3 and detail the evaluation framework of ChEMU in Sect. 4 including the evaluation methods and baseline models for each task. We present the evaluation results in Sect. 5 and finally conclude this paper in Sect. 6.

2 The ChEMU Chemical Reaction Corpus

The annotated corpus prepared for the ChEMU shared task consists of 1,500 patent snippets that were sampled from 170 English document patents from the European Patent Office and the United States Patent and Trademark Office. Each snippet contains a meaningful description of a chemical reaction [18].

The corpus was based on information captured in the Reaxys® database.[1] This resource contains details of chemical reactions identified through a mostly manual process of extracting key reaction details from sources including patents and scientific publications, dubbed "excerption" [9].

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

2.1 Annotation Process

To prepare the gold-standard annotations for the extracted patent snippets, multiple domain experts with rich expert knowledge in chemistry were invited to assist with corpus annotation. A silver-standard annotation set was first derived by mapping details from records in the Reaxys database to the source patents from which the information was originally extracted, by scanning the texts for mentions of relevant entities. Since the original records refer only to the patent IDs of source texts and do not provide the precise locations of excerpted entities or event steps, these annotations needed to be manually reviewed to produce higher quality annotations. Two domain experts manually annotated all patent snippets independently by correcting location information and adding more annotations. Their annotations were then evaluated by measuring their inter-annotator agreement (IAA) [6], and thereafter merged by a third domain expert who acted as an adjudicator, to resolve differences. More details about the quality evaluation over the annotations and the harmonization process will be provided in a more in-depth paper to follow.

An example snippet

[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.

Fig. 1. An example snippet with key focus text highlighted.

We present an example of a patent snippet in Fig. 1. This snippet describes the synthesis of a particular chemical compound, named *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; and (6) concentration of the purified product to get the title compound. We aim to extract the synthesis process from the patent snippet. To achieve this, 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). Thus, our annotation process consists of two steps: named entity annotations and relation annotations. Next, we describe the two steps of annotations in Sect. 2.2 and Sect. 2.3, respectively.

2.2 Named Entity 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 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. For example, the chemical compound "ethanol" in Fig. 1 must be annotated with the label "SOLVENT".

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 labels are summarized in Table 1. Interested readers may find more information about the labels in [12] and examples of named entity annotations in the Task 1—NER annotation guidelines [17].

2.3 Relation Annotations

A reaction step usually involves an action (i.e., a trigger word) and chemical compound(s) on which the action takes effect. To fully quantify a reaction step, it is also crucial for us to link an action to the conditions under which the action is carried out, and resultant yields from the action. Thus, annotations in this step are performed to identify the relations between actions (trigger words) and other arguments that are involved in the reaction steps, e.g., chemical compounds and conditions.

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 [13] 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 [7]. 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 in Table 1. Detailed annotation guidelines for relation annotation are available online [17].

Table 1. Definitions of entity and relation types, i.e., labels, in Task 1 and Task 2.

Label	Definition
Entity annotations	
STARTING_MATERIAL	A substance that is consumed in the course of a chemical reaction providing atoms to products is considered as starting material
REAGENT_CATALYST	A reagent is a compound added to a system to cause or help with a chemical reaction
REACTION_PRODUCT	A product is a substance that is formed during a chemical reaction
SOLVENT	A solvent is a chemical entity that dissolves a solute resulting in a solution
OTHER_COMPOUND	Other chemical compounds that are not the products, starting materials, reagents, catalysts and solvents
TIME	The reaction time of the reaction
TEMPERATURE	The temperature at which the reaction was carried out
YIELD_PERCENT	Yield given in percent values
YIELD_OTHER	Yields provided in other units than %
EXAMPLE_LABEL	A label associated with a reaction specification
Relation annotations	
WORKUP	An event step which is a manipulation required to isolate and purify the product of a chemical reaction
REACTION_STEP	An event within which starting materials are converted into the product
Arg1	The relation between an event trigger word and a chemical compound
ArgM	The relation between an event trigger word and a temperature, time, or yield entity

2.4 Snippet Annotation Format

The gold standard annotations for the data set were delivered in the BRAT standoff format [16]. Two files were delivered for each snippet: a text file (.txt) containing the original texts in the snippet, and a paired annotation file (.ann) containing all the annotations that have been made for that text, including entities, trigger words, and event steps. Continuing with the above snippet example, we show the formatted annotations for the highlighted sentence in Tables 2 and 3. For ease of presentation, we illustrate the format of the annotated named entities and trigger words in Table 2 and the format of the annotated event steps in Table 3 separately. We can see that two entities (i.e., T1 and T2) and one trigger word are included in Table 2. Two event steps are included in Table 3.

Table 2. The annotated entities and trigger words of the snippet example in BRAT standoff format [16].

ID	Entity type	Offsets	Text span
T1	TEMPERATURE	313 329	Room temperature
T2	REAGENT_CATALYST	231 252	Hydrazine monohydrate
T3	REACTION_STEP	281 290	Dissolved

Table 3. The annotated relations of the snippet example in BRAT standoff format [16]. Building on the annotations in Table 2, we see that R6 expresses the relation between a compound participating as a reagent (T2) in the T3 "dissolved" reaction step, and R8 captures the temperature (T1) at which that step occurred.

ID	Event type	Entity 1	Entity 2
R6	Arg1	T3	T2
R8	ArgM	T3	T1

2.5 Data Partitions

We randomly partitioned the whole data set into three splits for training, development and test purposes, with a ratio of 0.6/0.15/0.25. The training and development sets were released to participants for model development. Note that participants are allowed to use the combination of training and development sets and to use their own partitions to build models. The test set is withheld for use in the formal evaluation. The statistics of the three splits including their number of snippets, total number of sentences, and number of words per snippet, are summarized in Table 4.

To ensure the snippets included in the training, development, and test splits have similar distributions over labels, we compare the distributions of entity labels (ten classes of entities in Task 1 and two classes of trigger words in Task 2) of the three splits and summarize the results in Table 5. In Table 5, each cell represents the proportion (e.g., 0.038) of an entity label (e.g., EXAMPLE_LABEL) in the gold annotations of a data split (e.g., Train). The results in Table 5 confirm that the label distributions in the three splits are similar. Only some slight fluctuations ($\leqslant$0.004) across the three splits are observed for each label.

We further compare the International Patent Classification (IPC) [2] distributions of the training, development and test sets. The IPC information of each patent snippet reflects the application category of the original patent, e.g., "A61K" represents the category of patents that are preparations for medical, dental, or toilet purposes. Patents with different IPCs may be written in different ways and may differ in the vocabulary. Thus, they may differ in their linguistic characteristics. For each data split, we extract the primary IPC of each patent snippet included in the data split, and summarize the IPC distributions of the three splits in Table 6.

Table 4. Summary of data set statistics.

Data split	# snippets	#sentences	# words per snippet
Train	900	5,911	112.16
Dev	225	1,402	104.00
Test	375	2,363	108.63

Table 5. Distributions of entity labels in the training, development, and test sets.

Entity label	Train	Dev.	Test
EXAMPLE_LABEL	0.038	0.040	0.037
OTHER_COMPOUND	0.200	0.198	0.205
REACTION_PRODUCT	0.088	0.093	0.091
REAGENT_CATALYST	0.055	0.053	0.053
SOLVENT	0.049	0.046	0.045
STARTING_MATERIAL	0.076	0.076	0.075
TEMPERATURE	0.065	0.064	0.065
TIME	0.046	0.046	0.048
YIELD_OTHER	0.046	0.048	0.047
YIELD_PERCENT	0.041	0.042	0.041
REACTION_STEP	0.164	0.163	0.160
WORKUP	0.132	0.132	0.133

Table 6. Distributions of International Patent Classifications (IPCs) in the training, development, and test sets. Only dominating IPC groups that take up more than 1% of a data split are included in this table.

IPC	Train	Dev.	Test
A61K	0.277	0.278	0.295
A61P	0.129	0.134	0.113
C07C	0.063	0.045	0.060
C07D	0.439	0.444	0.437
C07F	0.011	0.009	0.010
C07K	0.013	0.012	0.008
C09K	0.012	0.021	0.011
G03F	0.012	0.019	0.014
H01L	0.019	0.021	0.019

3 Overview of Tasks

We provide two tasks in ChEMU lab: Task 1—Named Entity Recognition (NER), and Task 2—Event Extraction (EE). We also host a third track where participants can work on building end-to-end systems addressing both tasks jointly.

3.1 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 Sect. 2.2. The task requires the detection of the entity names in patent snippets and the assignment of correct labels to the detected entities (see Table 1). 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.

Participants in this track were provided with the patent snippets in the training and development sets and the gold standard entities of these snippets. In the evaluation phase, their models were evaluated using the snippets in the test set.

3.2 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 five reaction steps in the synthesis process of the chemical compound described in the example in Fig. 1. The event extraction task (Task 2) targets identifying these event steps.

Similarly to conventional event extraction problems [8], Task 2 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 Task 1 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 $\mathcal{S}$ of arguments that participate in e, Task 2 requires the creation of $|\mathcal{S}|$ relation entries connecting e to an argument entity in $\mathcal{S}$. Here, $|\mathcal{S}|$ represents the cardinality of the set $\mathcal{S}$. Finally, Task 2 requires the assignment of correct relation type labels (Arg1 or ArgM) to each of the detected relations.

Participants in the track for Task 2 were provided with the patent snippets in the training and development sets, along with the gold standard entity and event annotations in these snippets. In the evaluation phase, they were provided with the patent snippets in the test set as well as the gold standard entities in these snippets. Their models were evaluated against the ground truth events annotated in the test snippets. While in a real-world use of an event extraction system, gold standard entities would not typically be available, this framework allowed participants to focus on event extraction in isolation of the NER task.

This track was delayed until after both Task 1 and the end-to-end track (described below) were complete, in order to prevent any leakage of the information about gold standard entities from this track to the others.

3.3 End-to-End Systems

We also hosted a third track which allows participants to develop end-to-end systems that address both tasks simultaneously, i.e., the extraction of reaction events including their constituent entities directly from chemical patent snippets. This is a more realistic scenario for an event extraction system to be applied for large-scale annotation of events.

In the evaluation phase, participants in this track were provided only with the text of a patent, and were required to identify the named entities defined in Table 1, the trigger words defined in Sect. 3.2, and the event steps involving the entities, that is, the reaction steps. Proposed models in this track were evaluated against the events that they predict for the test snippets, which is the same as in Task 2. However, a major difference between this track and Task 2 is that the gold named entities were not provided but rather had to be predicted by the systems.

3.4 Track Overview

We illustrate the workflows of the three tracks in Fig. 2 using as example the sentence highlighted in Fig. 1. In Task 1—NER—, participants need to identify entities that defined in Table 1, e.g., the text span "ethanol" is identified as "SOLVENT". In Task 2—EE—, participants are provided with the three gold standard entities in the sentence. They are required to firstly identify the trigger

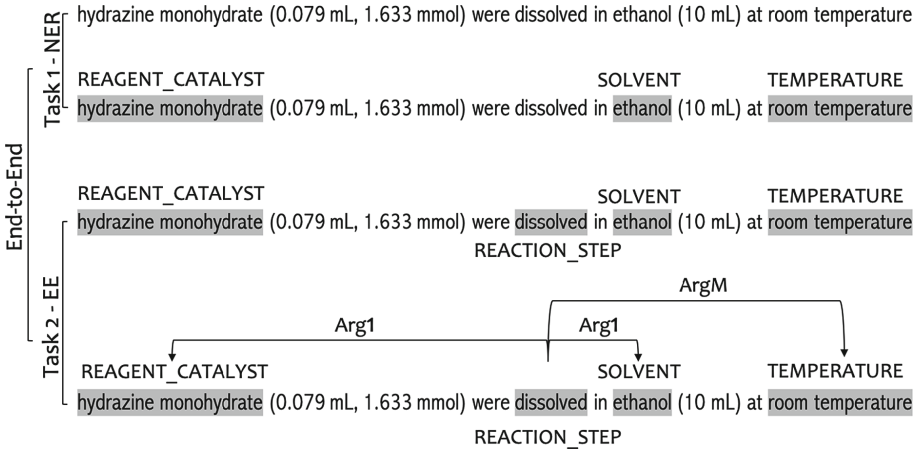

Fig. 2. Illustration of the three tasks. Shaded text spans represents annotated entities or trigger words. Arrows represent relations between entities.

words and their types (e.g., the text span "dissolved" is identified as "REACTION_STEP") and then identify the relations between the trigger words and the provided entities (e.g., a directed link from "dissolved" to "ethanol" is added and labeled as "ARG1"). In the track of end-to-end systems, participants are only provided with the original text. They are required to identify both the entities and the trigger words, and predict the event steps directly from the text.

4 Evaluation Framework

In this section, we describe the evaluation framework of the ChEMU lab. We introduce three baseline algorithms for Task 1, Task 2, and end-to-end systems, respectively.

4.1 Evaluation Methods

The evaluation process consists of two phases. In phase one, the text files of the snippets in the test set are provided to all teams participating in Task 1 and the track for end-to-end systems. Once phase one is completed, the gold standard entities of the snippets in the test set are provided to all teams participating in Task 2. For each track, each participating team is allowed to select up to 3 rounds of results (runs) as their final submissions.

We use BRATEval [1] to evaluate all the runs that we receive. Three metrics are used to evaluate the performance of all the submissions for Task 1: Precision, Recall, and F_1-score. Specifically, given a predicted entity and a ground-truth entity, we treat the two entities as a match if (1) the types associated with the two entities match; and (2) their text spans match. The overall Precision, Recall, and F_1-score are computed by micro-averaging all instances (entities).

In addition, we exploit two different matching criteria, exact-match and relaxed-match, when comparing the texts spans of two entities. Here, the exact-match criterion means that we consider that the text span of an entity matches with that of another entity if both the starting and the end offsets of their spans match. The relaxed-match criterion means that we consider that the text span of one entity matches with that of another entity as long as their text spans overlap.

The submissions for Task 2 and end-to-end systems are evaluated using Precision, Recall, and F_1-score by comparing the predicted events and gold standard events. We consider two events as a match if (1) their trigger words and event types are the same; and (2) the entities involved in the two events match. Here, we follow the method in Task 1 to test whether two entities match. This means that the matching criteria of exact-match and relaxed-match are also applied in the evaluation of Task 2 and of end-to-end systems. Note that the relaxed-match will only be applied when matching the spans of two entities; it does not relax the requirement that the entity type of predicted and ground truth entities must agree. Since Task 2 provides gold entities but not event triggers with their ground

truth spans, the relaxed-match only reflects the accuracy of spans of predicted trigger words.

To somewhat accommodate a relaxed form of entity type matching, we also evaluate submissions in Task 1—NER using a set of high-level labels shown in the hierarchical structure of entity classes in Fig. 3. The higher-level labels used are highlighted in grey. In this set of evaluations, given a predicted entity and a ground-truth entity, we consider that their labels match as long as their corresponding high-level labels match. For example, suppose we get as predicted entity "STARTING_MATERIAL, [335, 351), boron tribromide" while the (correct) ground-truth entity instead reads "REAGENT_CATALYST, [335, 351), boron tribromide", where each entity is presented in the form of "TYPE, SPAN, TEXT". In the evaluation framework described earlier this example will be counted as a mismatch. However, in this additional set of entity type relaxed evaluations we consider the two entities as a match, since both labels "STARTING_MATERIAL" and "REAGENT_CATALYST" specialize their parent label "COMPOUND".

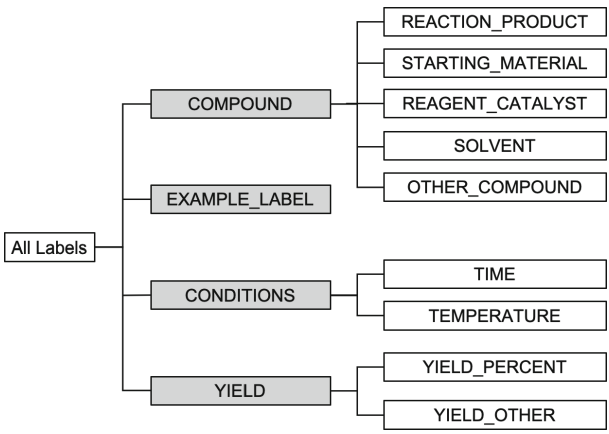

Fig. 3. Illustration of the hierarchical NER class structure used in evaluation.

4.2 Baselines

We released one baseline method for each task as a benchmark method. Specifically, the baseline for Task 1 is based on retraining **BANNER** [10] on the training and development data; the baseline for Task 2 is a co-occurrence method; and the baseline for end-to-end systems is a two-stage algorithm that first uses BANNER to identify entities in the input and then uses the co-occurrence method to extract events.

BANNER. BANNER is a named entity recognition tool for bio-medical data. In this baseline, we first use the GENIA Sentence Splitter (GeniaSS) [14] to

split input texts into separate sentences. The resulting sentences are then fed into BANNER, which predicts the named entities using three steps, namely tokenization, feature generation, and entity labelling. A simple tokenizer is used to break sentences into either a contiguous block of letters and/or digits or a single punctuation mark. BANNER uses a conditional random field (CRF) implementation derived from the MALLET toolkit[2] for feature generation and token labelling. The set of machine learning features used consist primarily of orthographic, morphological and shallow syntax features.

Co-occurrence Method. This method first creates a dictionary D_e for the observed trigger words and their corresponding types from the training and development sets. For example, if a word "added" is annotated as a trigger word with the label of "WORKUP" in the training set, we add an entry $\langle \text{added}, \text{WORKUP} \rangle$ to D_e. In the case where the same word has been observed to appear as both types of "WORKUP" and "REACTION_STEP", we only keep as entry in D its most frequent label. The method also creates an event dictionary D_r for the observed event types in the training and development sets. For example, if an event $\langle \text{ARG1}, \text{E1}, \text{E2} \rangle$ is observed where "E1" corresponds to trigger word "added" of type "WORKUP" and "E2" corresponds to entity "water" of type "OTHER_COMPOUND", we add an entry $\langle \text{ARG1}, \text{WORKUP}, \text{OTHER\_COMPOUND} \rangle$ to D_r.

To predict events, this method first identifies all trigger words in the test set using D_e. It then extracts two events $\langle \text{ARG1}, \text{T1}, \text{T2} \rangle$ and $\langle \text{ARGM}, \text{T1}, \text{T2} \rangle$ for a trigger word "E1" and an entity "E2" if (1) they co-occur in the same sentence; and (2) the relation type $\langle \text{ARGx}, \text{T1}, \text{T2} \rangle$ is included in D_r. Here, "ARGx" can be "ARG1" or "ARGM", and "T1" and "T2" are the entity types of "E1" and "E2" respectively.

BANNER + Co-occurrence Method. The above two baselines are combined to form a two-stage method for end-to-end systems. This baseline first uses BANNER to identify all the entities in Task 1. Then it utilizes the co-occurrence method to predict events, except that gold standard entities are replaced with the entities predicted by BANNER in the first stage.

4.3 Submission Website

We developed a submission website which allows participants to submit their predictions for each task during the evaluation phase.[3] In addition, the website offers several important functions to facilitate organizing the lab.

First, it hosts the download links for the training, development, and test data sets so that participants can access the data sets conveniently. Second, it allows participants to test the performance (against the development set) of their models before the evaluation phase starts, which also offers a chance for participants to familiarize themselves with the evaluation tool BRATEval [1].

[2] http://mallet.cs.umass.edu/.
[3] http://chemu.eng.unimelb.edu.au/.

The website also hosts a private leaderboard for each team that ranks all runs submitted by each team, and a public leaderboard that ranks all runs that have been made public by teams.

5 Results and Discussions

A total of 39 teams registered for the ChEMU shared task. Among them, 36 teams registered for Task 1, 31 teams registered for Task 2, and 28 teams registered for both tasks. The 39 teams are spread across 13 different countries, from both the academic and industry research communities. In this section, we report the results of all the runs that we received for each task.

5.1 Task 1—Named Entity Recognition

Task 1 received considerable interest with the submission of 25 runs from 11 teams. The 11 teams include 1 team from Germany (OntoChem), 3 teams from India (AUKBC, SSN_NLP and JU_INDIA), 1 team from Switzerland (BiTeM), 1 team from Portugal (Lasige_BioTM), 1 team from Russia (KFU_NLP), 1 team from the United Kingdom (NextMove Software/Minesoft), 2 teams from the United States of America (Melaxtech and NLP@VCU), and 1 team from Vietnam (VinAI). We evaluate the performance of all 25 runs, comparing their predicted entities with the ground-truth entities of the patent snippets in the test set. We report the performances of all runs under both matching criteria in terms of three metrics, namely Precision, Recall, and F_1-score.

We report the overall performance of all runs in Table 7. The baseline of Task 1 achieves 0.8893 in F_1-score under exact match. Nine runs outperform the baseline in terms of F_1-score under exact match. The best run was submitted by team Melaxtech, achieving a high F_1-score of 0.9570. There were sixteen runs with an F_1-score greater than 0.90 under relaxed-match. However, under exact-match, only seven runs surpassed 0.90 in F_1-score. 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.

Next, we evaluate the performance of all 25 runs using the high-level labels in Fig. 3 (highlighted in grey). We report the performances of all runs in terms of Precision, Recall, and F_1-score in Table 8.

5.2 Task 2—Event Extraction

We received 10 runs from five teams. Specifically, the five teams include 1 team from Portugal (Lasige_BioTM), 1 team from Turkey (BOUN_REX), 1 team from the United Kingdom (NextMove Software/Minesoft) and 2 teams from the United States of America (Melaxtech and NLP@VCU). We evaluate all runs using the metrics Precision, Recall, and F_1-score. Again, we utilize the

Table 7. Overall performance of all runs in Task 1—Named Entity Recognition. Here, P, R, and F represents the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italic*. The results are ordered by their performance in terms of F_1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	0.9571	**0.9570**	**0.9570**	0.9690	**0.9687**	**0.9688**
Melaxtech-run2	**0.9587**	*0.9529*	*0.9558*	0.9697	0.9637	0.9667
Melaxtech-run3	*0.9572*	0.9510	0.9541	0.9688	0.9624	0.9656
VinAI-run1	0.9462	0.9405	0.9433	**0.9707**	0.9661	*0.9684*
Lasige_BioTM-run1	0.9327	0.9457	0.9392	0.9590	0.9671	0.9630
BiTeM-run1	0.9378	0.9087	0.9230	0.9692	0.9558	0.9624
BiTeM-run2	0.9083	0.9114	0.9098	0.9510	*0.9684*	0.9596
NextMove/Minesoft-run1	0.9042	0.8924	0.8983	0.9301	0.9181	0.9240
NextMove/Minesoft-run2	0.9037	0.8918	0.8977	0.9294	0.9178	0.9236
Baseline	0.9071	0.8723	0.8893	0.9219	0.8893	0.9053
NLP@VCU-run1	0.8747	0.8570	0.8658	0.9524	0.9513	0.9518
KFU_NLP-run1	0.8930	0.8386	0.8649	*0.9701*	0.9255	0.9473
NLP@VCU-run2	0.8705	0.8502	0.8602	0.9490	0.9446	0.9468
NLP@VCU-run3	0.8665	0.8514	0.8589	0.9486	0.9528	0.9507
KFU_NLP-run2	0.8579	0.8329	0.8452	0.9690	0.9395	0.9540
NextMove/Minesoft-run3	0.8281	0.8083	0.8181	0.8543	0.8350	0.8445
KFU_NLP-run3	0.8197	0.8027	0.8111	0.9579	0.9350	0.9463
BiTeM-run3	0.8330	0.7799	0.8056	0.8882	0.8492	0.8683
OntoChem-run1	0.7927	0.5983	0.6819	0.8441	0.6364	0.7257
AUKBC-run1	0.6763	0.4074	0.5085	0.8793	0.5334	0.6640
AUKBC-run2	0.4895	0.1913	0.2751	0.6686	0.2619	0.3764
SSN_NLP-run1	0.2923	0.1911	0.2311	0.8633	0.4930	0.6276
SSN_NLP-run2	0.2908	0.1911	0.2307	0.8595	0.4932	0.6267
JU_INDIA-run1	0.1411	0.0824	0.1041	0.2522	0.1470	0.1857
JU_INDIA-run2	0.0322	0.0151	0.0206	0.1513	0.0710	0.0966
JU_INDIA-run3	0.0322	0.0151	0.0206	0.1513	0.0710	0.0966

two matching criteria, namely exact-match and relaxed-match, when comparing the trigger words in the submitted runs and ground-truth data.

The overall performance of each run is summarized in Table 9.[4] The baseline (co-occurrence method) scored relatively high in Recall, i.e, 0.8861. This was expected, since the co-occurrence method aggressively extracts all possible events

[4] The run that we received from team Lasige_BioTM is not included in the table due to a technical issue found in this run.

within a sentence. However, the F_1-score was low due to its low Precision score. Here, all runs outperform the baseline in terms of F_1-score under exact-match. Melaxtech ranks first among all official runs in this task, with an F_1-score of 0.9536.

Table 8. Overall performance of all runs in Task 1—Named Entity Recognition where the set of high-level labels in Fig. 3 is used. Here, P, R, and F represents the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italic*. The results are ordered by their performance in terms of F_1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	*0.9774*	**0.9774**	**0.9774**	0.9906	0.9901	*0.9903*
Melaxtech-run2	**0.9789**	*0.9732*	*0.9760*	*0.9910*	0.9849	0.9879
Melaxtech-run3	0.9775	0.9714	0.9744	0.9905	0.9838	0.9871
Lasige_BioTM-run1	0.9571	0.9706	0.9638	0.9886	*0.9943*	**0.9915**
VinAI-run1	0.9635	0.9579	0.9607	0.9899	0.9854	0.9877
Baseline	0.9657	0.9288	0.9469	0.9861	0.9519	0.9687
BiTeM-run1	0.9573	0.9277	0.9423	0.9907	0.9770	0.9838
NextMove/Minesoft-run2	0.9460	0.9330	0.9394	0.9773	0.9611	0.9691
NextMove/Minesoft-run1	0.9458	0.9330	0.9393	0.9773	0.9610	0.9691
BiTeM-run2	0.9323	0.9357	0.9340	0.9845	**0.9962**	*0.9903*
NextMove/Minesoft-run3	0.9201	0.8970	0.9084	0.9571	0.9308	0.9438
NLP@VCU-run1	0.9016	0.8835	0.8925	0.9855	0.9814	0.9834
NLP@VCU-run2	0.9007	0.8799	0.8902	0.9882	0.9798	0.9840
NLP@VCU-run3	0.8960	0.8805	0.8882	0.9858	0.9869	0.9863
KFU_NLP-run1	0.9125	0.8570	0.8839	**0.9911**	0.9465	0.9683
BiTeM-run3	0.9073	0.8496	0.8775	0.9894	0.9355	0.9617
KFU_NLP-run2	0.8735	0.8481	0.8606	0.988	0.9569	0.9722
KFU_NLP-run3	0.8332	0.8160	0.8245	0.9789	0.9516	0.9651
OntoChem-run1	0.9029	0.6796	0.7755	0.9611	0.7226	0.8249
AUKBC-run1	0.7542	0.4544	0.5671	0.9833	0.5977	0.7435
AUKBC-run2	0.6605	0.2581	0.3712	0.9290	0.3612	0.5201
SSN_NLP-run2	0.3174	0.2084	0.2516	0.9491	0.5324	0.6822
SSN_NLP-run1	0.3179	0.2076	0.2512	0.9505	0.5304	0.6808
JU_INDIA-run1	0.2019	0.1180	0.1489	0.5790	0.3228	0.4145
JU_INDIA-run2	0.0557	0.0262	0.0357	0.4780	0.2149	0.2965
JU_INDIA-run3	0.0557	0.0262	0.0357	0.4780	0.2149	0.2965

5.3 End-to-end Systems

We received 10 end-to-end system runs from four teams. The four teams include 1 team from Turkey (BOUN_REX), 1 team from the United Kingdom (NextMove Software/Minesoft) and 2 teams from the United States of America (Melaxtech and NLP@VCU).

Table 9. Overall performance of all runs in Task 2—Event Extraction. Here, P, R, and F represent the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italics*. The results are ordered by their performance in terms of F_1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	*0.9568*	**0.9504**	**0.9536**	*0.9580*	**0.9516**	**0.9548**
Melaxtech-run2	**0.9619**	0.9402	*0.9509*	**0.9632**	0.9414	*0.9522*
Melaxtech-run3	0.9522	*0.9437*	0.9479	0.9534	*0.9449*	0.9491
NextMove/Minesoft-run1	0.9441	0.8556	0.8977	0.9441	0.8556	0.8977
NextMove/Minesoft-run2	0.8746	0.7816	0.8255	0.8909	0.7983	0.8420
BOUN_REX-run1	0.7610	0.6893	0.7234	0.7610	0.6893	0.7234
NLP@VCU-run1	0.8056	0.5449	0.6501	0.8059	0.5451	0.6503
NLP@VCU-run2	0.5120	0.7153	0.5968	0.5125	0.7160	0.5974
NLP@VCU-run3	0.5085	0.7126	0.5935	0.5090	0.7133	0.5941
Baseline	0.2431	0.8861	0.3815	0.2431	0.8863	0.3816

Table 10. Overall performance of all runs in end-to-end systems. Here, P, R, and F represent the Precision, Recall, and F_1-score, respectively. For each metric, we highlight the best result in **bold** and the second best result in *italics*. The results are ordered by their performance in terms of F_1-score under exact-match.

Run	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
Melaxtech-run1	**0.9201**	**0.9147**	**0.9174**	**0.9319**	**0.9261**	**0.929**
NextMove/Minesoft-run1	*0.8492*	*0.7609*	*0.8026*	*0.8663*	*0.7777*	*0.8196*
NextMove/Minesoft-run2	0.8486	0.7602	0.8020	0.8653	0.7771	0.8188
NextMove/Minesoft-run3	0.8061	0.7207	0.7610	0.8228	0.7371	0.7776
OntoChem-run1	0.7971	0.3777	0.5126	0.8407	0.3984	0.5406
OntoChem-run2	0.7971	0.3777	0.5126	0.8407	0.3984	0.5406
OntoChem-run3	0.7971	0.3777	0.5126	0.8407	0.3984	0.5406
Baseline	0.2104	0.7329	0.3270	0.2135	0.7445	0.3319
Melaxtech-run2	0.2394	0.2647	0.2514	0.2429	0.2687	0.2552
Melaxtech-run3	0.2383	0.2642	0.2506	0.2421	0.2684	0.2545

The overall performance of all runs is summarized in Table 10 in terms of Precision, Recall, and F_1-score under both exact-match and relaxed-match.[5] Since gold entities are not provided in this task, the average performance of the runs in this task are slightly lower than those in Task 2. Note that the Recall scores of most runs are substantially lower than their Precision scores. This may reveal that the task of identifying a relation from a chemical patent is harder than the task of typing an identified relation. The first run from Melaxtech team ranks best among all runs received for this task.

6 Conclusions

This paper presents a general overview of the activities and outcomes of the ChEMU 2020 evaluation lab. The ChEMU lab targets two important information extraction tasks applied to chemical patents: (1) named entity recognition, which aims to identify chemical compounds and their specific roles in chemical reactions; and (2) event extraction, which aims to identify the single event steps that form a chemical reaction.

We received registrations from 39 teams and 46 runs from 11 teams across all tasks and tracks. The evaluation results show that many effective solutions have been proposed, achieving high accuracy on each task. We look forward to fruitful discussions and exploring the methodological details of these submissions at the workshop.

Acknowledgements. 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.

References

1. BRATEval evaluation tool. https://bitbucket.org/nicta_biomed/brateval/src/master/. Accessed 23 June 2020
2. International Patent Classification. https://www.wipo.int/classifications/ipc/en/. Accessed 23 June 2020
3. Akhondi, S.A., et al.: Annotated chemical patent corpus: a gold standard for text mining. PLoS ONE **9**(9), e107477 (2014)
4. Akhondi, S.A., et al.: Automatic identification of relevant chemical compounds from patents. Database **2019** (2019)
5. Bregonje, M.: Patents: a unique source for scientific technical information in chemistry related industry? World Patent Inf. **27**(4), 309–315 (2005)

[5] The run that we received from the Lasige_BioTM team is not included in the table as there was a technical issue in this run. Two runs from Melaxtech, Melaxtech-run2 and Melaxtech-run3, had very low performance, due to an error in their data pre-processing step.

6. Carletta, J.: Assessing agreement on classification tasks: the Kappa statistic. Comput. Linguist. **22**(2), 249–254 (1996). https://www.aclweb.org/anthology/J96-2004
7. Jurafsky, D., Martin, J.H.: Semantic role labeling and argument structure. In: Speech & Language Processing, 3rd edn. Pearson Education India (2009)
8. Kim, J.D., Ohta, T., Pyysalo, S., Kano, Y., Tsujii, J.: Overview of BioNLP 2009 shared task on event extraction. In: Proceedings of the BioNLP 2009 Workshop Companion Volume for Shared Task, pp. 1–9 (2009)
9. Lawson, A.J., Roller, S., Grotz, H., Wisniewski, J.L., Goebels, L.: Method and software for extracting chemical data. German patent no. DE102005020083A1 (2011)
10. Leaman, R., Gonzalez, G.: BANNER: an executable survey of advances in biomedical named entity recognition. In: Pacific Symposium on Biocomputing 2008, pp. 652–663. World Scientific (2008)
11. Muresan, S., et al.: Making every SAR point count: the development of Chemistry Connect for the large-scale integration of structure and bioactivity data. Drug Discov. Today **16**(23–24), 1019–1030 (2011)
12. Nguyen, D.Q., et al.: ChEMU: named entity recognition and event extraction of chemical reactions from patents. In: Jose, J.M., et al. (eds.) ECIR 2020. LNCS, vol. 12036, pp. 572–579. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-45442-5_74
13. Palmer, M., Gildea, D., Kingsbury, P.: The proposition bank: an annotated corpus of semantic roles. Comput. Linguist. **31**(1), 71–106 (2005)
14. Sætre, R., Yoshida, K., Yakushiji, A., Miyao, Y., Matsubayashi, Y., Ohta, T.: AKANE system: protein-protein interaction pairs in BioCreAtIvE2 challenge, PPI-IPS subtask. In: Proceedings of the second BioCreative challenge workshop, Madrid, vol. 209, p. 212 (2007)
15. Senger, S., Bartek, L., Papadatos, G., Gaulton, A.: Managing expectations: assessment of chemistry databases generated by automated extraction of chemical structures from patents. J. Cheminform. **7**(1), 1–12 (2015). https://doi.org/10.1186/s13321-015-0097-z
16. Stenetorp, P., Pyysalo, S., Topić, G., Ohta, T., Ananiadou, S., Tsujii, J.: BRAT: a web-based tool for NLP-assisted text annotation. In: Proceedings of the Demonstrations at the 13th Conference of the European Chapter of the Association for Computational Linguistics, pp. 102–107 (2012)
17. Verspoor, K., et al.: ChEMU dataset for information extraction from chemical patents. https://doi.org/10.17632/wy6745bjfj.1
18. Yoshikawa, H., et al.: Detecting chemical reactions in patents. In: Proceedings of the 17th Annual Workshop of the Australasian Language Technology Association, pp. 100–110 (2019)