

The ChEMU 2022 Evaluation Campaign: Information Extraction in Chemical Patents

Yuan Li¹, Biaoyan Fang¹, Jiayuan He^{2,1}, Hiyori Yoshikawa^{1,3}, Saber A. Akhondi⁴, Christian Druckenbrodt⁵, Camilo Thorne⁵, Zenan Zhai¹, Zubair Afzal⁴, Trevor Cohn¹, Timothy Baldwin¹, and Karin Verspoor^{2,1}

¹ The University of Melbourne, Australia ² RMIT University, Australia ³ Fujitsu Limited, Japan

⁴ Elsevier BV, Netherlands ⁵ Elsevier Information Systems GmbH, Germany

Expression-level Tasks

- Named Entity Recognition
- Event Extraction
- Anaphora Resolution

We rerun NER/EE task from ChEMU 2020 and AR task from ChEMU 2021 with new test data for evaluation.

Named Entity Recognition

This task aims to identify chemical compounds and their specific types. In addition, this task also requires identification of the temperatures and reaction times at which the chemical reaction is carried out, as well as yields obtained for the final chemical product and the label of the reaction. In total, the participants need to find 10 types of named entities.

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: <i>used</i> → REACTION_PRODUCT: title compound REACTION_STEP: <i>used</i> → YIELD_OTHER: 1.180 g REACTION_STEP: <i>used</i> → YIELD_PERCENT: 95.2%
AR	The title compound was used without purification (1.180 g, 95.2%) as <i>yellow solid</i> . COREFERENCE: <i>yellow solid</i> → The title compound (1.180 g, 95.2%)

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

Event Extraction

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

Anaphora Resolution

This task requires the resolution of anaphoric dependencies between expressions in chemical patents. The participants are required to find 5 types of anaphoric relationships in chemical patents, i.e. coreference, transformed, reaction-associated, work-up and contained.

Research Directions

Can multi-task learning benefit NER/EE/AR?

- In our ChEMU corpus, every snippet has been annotated for all three expression-level 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.

Can snippet-level models be extended to process full documents?

- Taking a full document as input increases the complexity of the tasks. The reaction references can relate reaction descriptions that are far apart, and the semantics of a table may depend on linguistic context from the document structure or content.

Key Information

- Our shared tasks are run in Kaggle-style where public leaderboards are based on a subset of 30% of the test set, while the private leaderboards are based on the remaining 70% of data and remain secret until the end of the evaluation.
- 16 May 2022: End of evaluation cycle and feedback for participants.
- 9 June 2022: Submission of CLEF 2022 Working Notes (participants)
- All training/dev/test datasets are available on our website.
- Website: <http://chemu.eng.unimelb.edu.au>
- Email: chemu.lab@gmail.com

Document-level Tasks

- Reaction Reference Resolution
- Table Semantic Classification

They are grouped since both of these tasks take a complete patent document as input rather than the short snippet extracts of Expression-level Tasks.

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

Table 2: An example for Reaction Reference Resolution, where reaction 2 (RX2) is producing Compound B13 following the procedure that reaction 1 (RX1) produces Compound B11.

Reaction Reference Resolution

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

Table Semantic Classification

This task is about categorising tables in chemical patents based on their contents, which supports identification of tables containing key information. We define 8 types of tables and Figure 1 shows an example SPECT table (spectroscopic data).

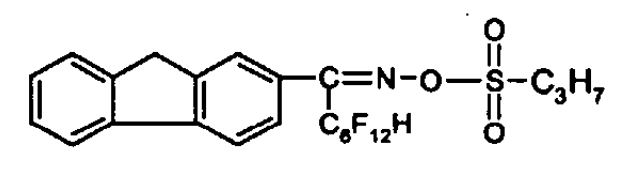
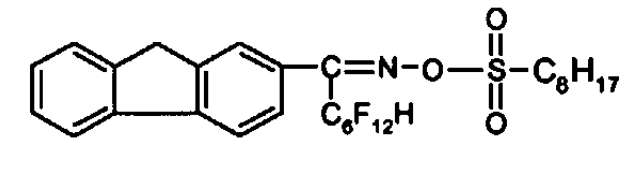
Ex.	Structure	Purification, Physical properties
3		Recrystallization from 2-propanol ¹ H-NMR and ¹⁹ F-NMR (CDCl ₃) δ [ppm]: 1.10 (t, 3H), 1.87-1.98 (m, 2H), 3.39 (t, 2H), 3.98 (s, 2H), 6.05 (t, 1H), 7.33-7.43 (m, 3H), 7.54-7.62 (m, 2H), 7.84 (d, 1H), 7.88 (d, 1H), -137.40 (d, 2F), -129.74 (s, 2F), -123.80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

Figure 1: An example table in SPECT category.