ChEMU

Cheminformatics Elsevier Melbourne Universities

http://chemu.eng.unimelb.edu.au/

End-to-End Chemical Reaction Extraction from Patents

Yuan Li, Biaoyan Fang, Jiayuan He, Hiyori Yoshikawa, Saber Akhondi, Christian Druckenbrodt, Camilo Thorne, Zenan Zhai, **Karin Verspoor**

Introduction

- Chemical reaction information is relevant to *drug discovery*, an important and yet costly (and high uncertainty) process.
- Chemical patents provide *timely* and *comprehensive* information about newly discovered chemical compounds.

- We aim to build a system that focuses on chemical reaction processes described in chemical patents.
  - enable automatic identification of each reaction described in a complete patent document
  - fully characterize each reaction by extracting each relevant component

➤ Search ➤ Compare ➤ Synthesise

➤ Connect ➤ Discover ➤ Characterise

Structuring chemical reaction data

2-Phenyl-2H-imidazo[1,5-a]pyridinium tetrafluoroborate (1)

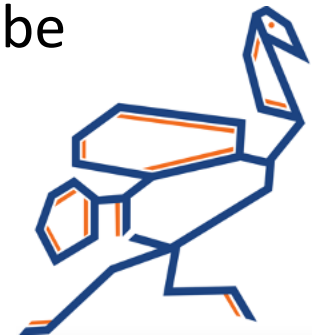
The general synthesis starts with the slow addition of excess concentrated **hydrochloric acid** to **aniline** (4.66 g, 4.6 mL, 50.0 mmol) dissolved in a small amount of **methylene chloride** under rigorous stirring. A solid immediately formed, which was collected, washed with diethyl ether and dried at 40 °C at <10 mbar for two hours. Then the **hydrochloride salt** was dissolved in 100 mL **ethanol**, and 37 wt% aqueous **formaldehyde** solution (2.25 g, 2.1 mL, 75.0 mmol) as well as **2-pyridinecarboxyaldehyde** (5.36 g, 4.8 mL, 50.0 mmol) were added.

2-(4-Methoxyphenyl)-2H-imidazo[1,5-a]pyridinium chloride monohydrate (3)

The synthesis followed *the general procedure as given for* ***1*** but without salt metathesis to *the corresponding tetrafluoroborate salt*. **4-Methoxyaniline** (6.16 g, 50.0 mmol) was used as *amine*.

Product 1: 2-Phenyl-2H-imidazo[1,5-a]pyridinium tetrafluoroborate
Stage 1:
Reactant 1: hydrochloric acid
Reactant 2: aniline
Solvent 3: methylene chloride
Product: hydrochloride salt[1]
Stage 2: collected, washed with diethyl ether
Stage 3:
Reactant 4: hydrochloride salt[1]
Solvent 5: ethanol
Solvent 6: aqueous formaldehyde solution
Reactant 7: 2-pyridinecarboxyaldehyde
Product 3: 2-(4-Methoxyphenyl)-2H-imidazo[1,5-a]pyridinium chloride monohydrate

- A chemical reaction is a process leading to the transformation of one set of chemical substances to another.
- A full reaction requires at least the starting materials and the final product to be defined, and usually includes information such as reagents, catalysts, and experiment conditions to further describe the reaction.



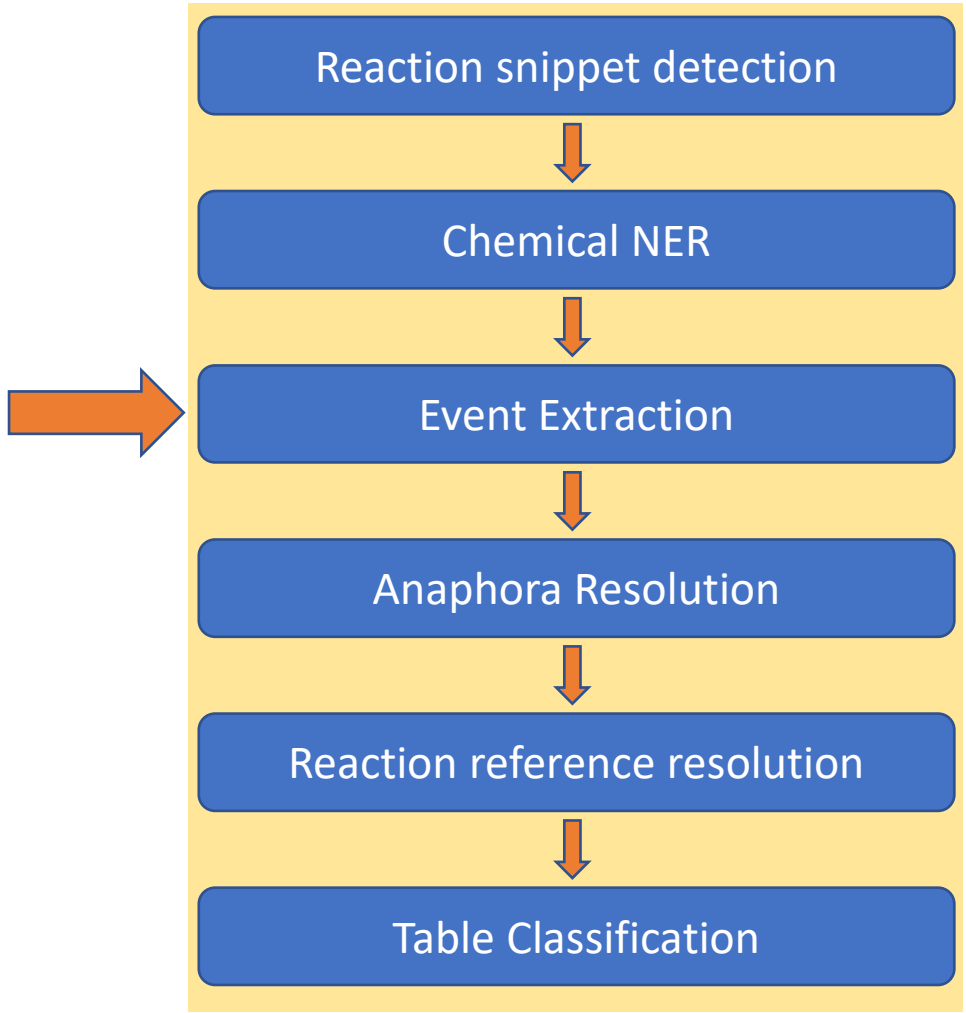
A Pipeline of NLP Tasks

2-Phenyl-2H-imidazo[1,5-a]pyridinium tetrafluoroborate (1)

The general synthesis starts with the slow addition of excess concentrated **hydrochloric acid** to **aniline** (4.66 g, 4.6 mL, 50.0 mmol) dissolved in a small amount of **methylene chloride** under rigorous stirring. A solid immediately formed, which was collected, washed with diethyl ether and dried at 40 °C at <10 mbar for two hours. Then the **hydrochloride salt** was dissolved in 100 mL **ethanol**, and 37 wt% aqueous **formaldehyde** solution (2.25 g, 2.1 mL, 75.0 mmol) as well as **2-pyridinecarboxyaldehyde** (5.36 g, 4.8 mL, 50.0 mmol) were added.

2-(4-Methoxyphenyl)-2H-imidazo[1,5-a]pyridinium chloride monohydrate (3)

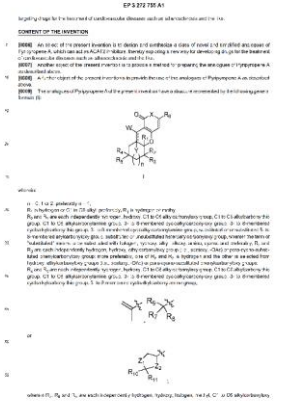
The synthesis followed *the general procedure as given for* ***1*** but without salt metathesis to *the corresponding tetrafluoroborate salt*. **4-Methoxyaniline** (6.16 g, 50.0 mmol) was used as *amine*.

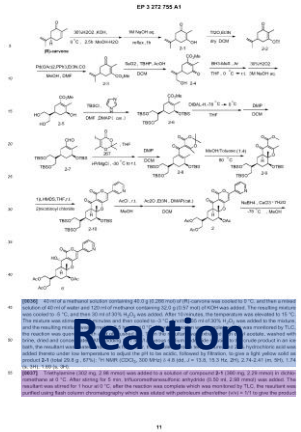


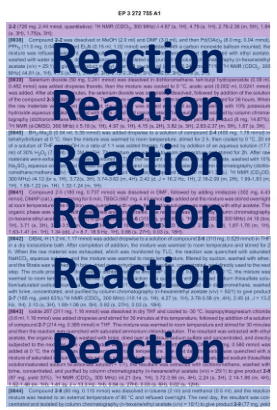
Product 1: 2-Phenyl-2H-imidazo[1,5-a]pyridinium tetrafluoroborate
Stage 1:
Reactant 1: hydrochloric acid
Reactant 2: aniline
Solvent 3: methylene chloride
Product: hydrochloride salt[1]
Stage 2: collected, washed with diethyl ether
Stage 3:
Reactant 4: hydrochloride salt[1]
Solvent 5: ethanol
Solvent 6: aqueous formaldehyde solution
Reactant 7: 2-pyridinecarboxyaldehyde
Product 3: 2-(4-Methoxyphenyl)-2H-imidazo[1,5-a]pyridinium chloride monohydrate

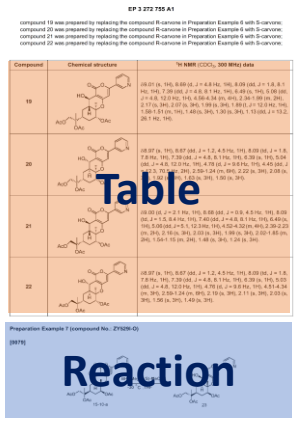
Reaction snippet detection

- We first need to locate reaction descriptions in a patent, for processing in downstream tasks.
- We formulate this task as a paragraph-level sequence tagging problem, where a patent is given as a sequence of paragraphs and the task is to detect a span of contiguous paragraphs describing a single chemical reaction.



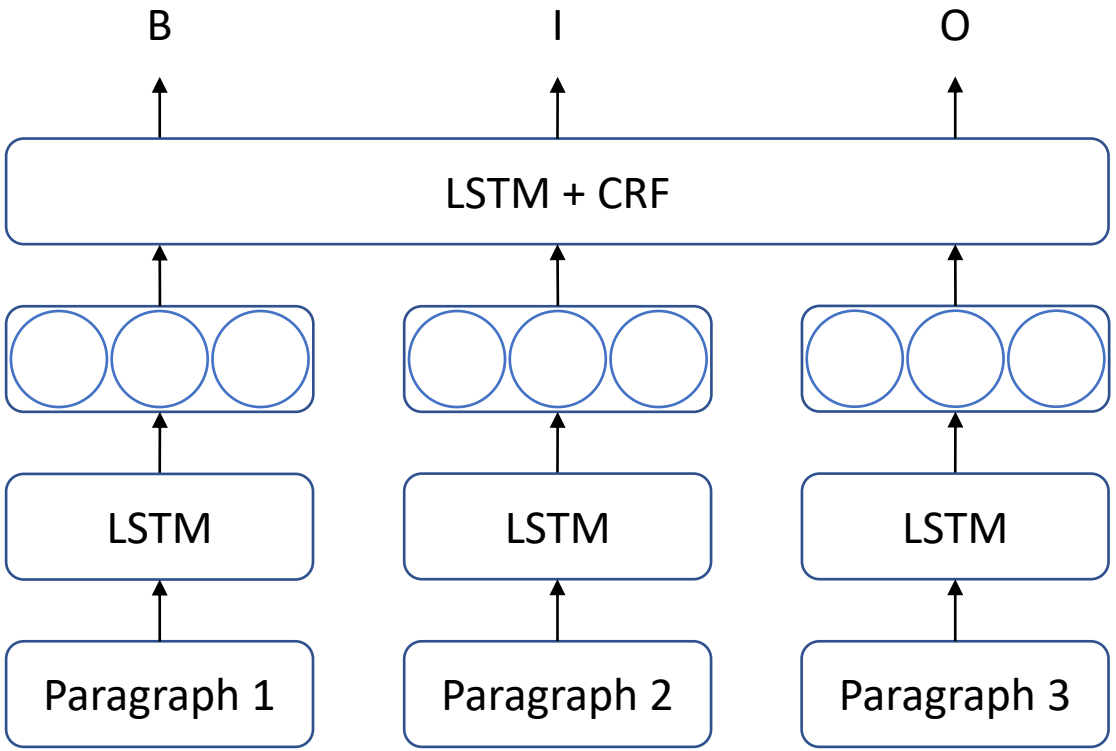






Reaction snippet detection

- We use LSTM to encode every paragraph and get paragraph representation, then use another layer of LSTM to process all the representations, and finally use CRF to decode them into tags.



Chemical NER

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

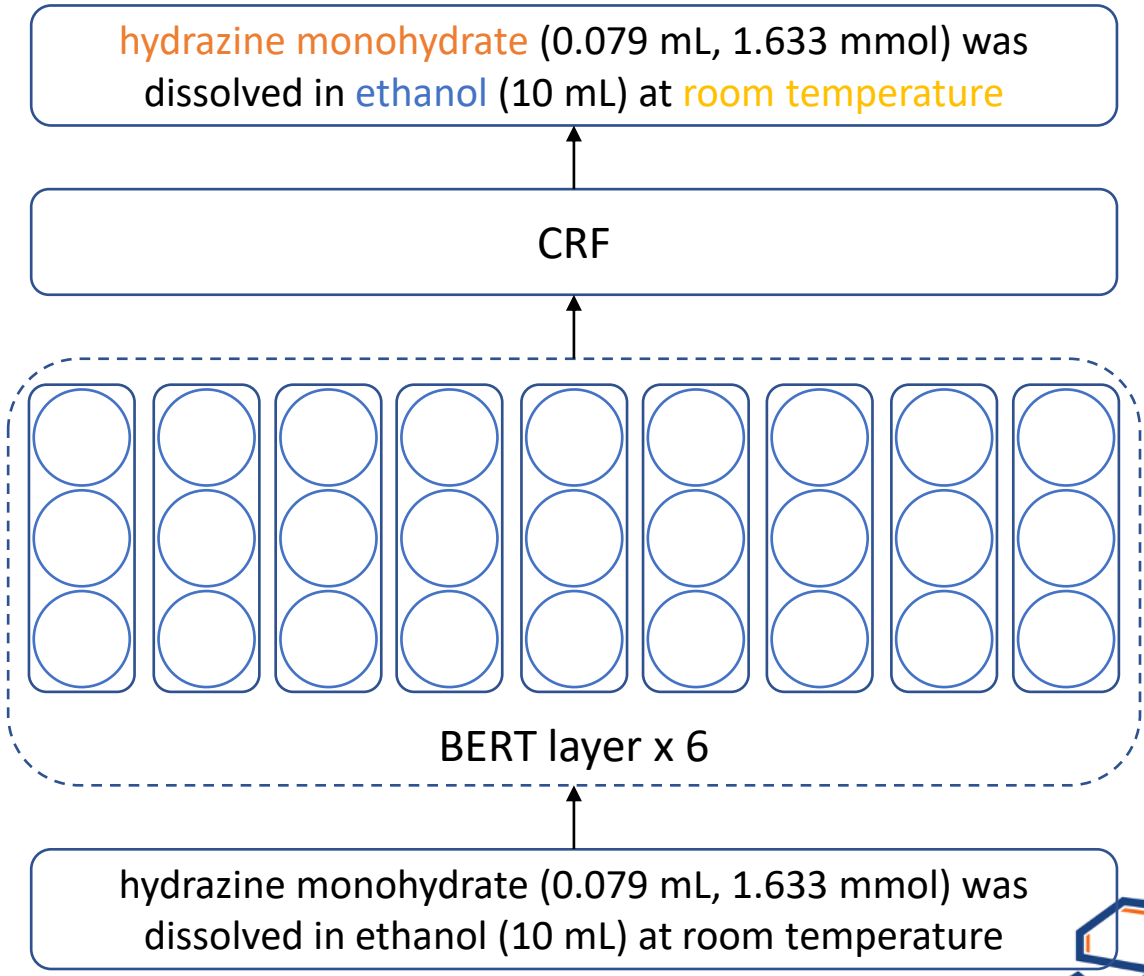
hydrazine monohydrate (0.079 mL, 1.633 mmol) was dissolved in ethanol (10 mL) at room temperature

- REAGENT_CATALYST: hydrazine monohydrate
- SOLVENT: ethanol
- TEMPERATURE: room temperature

Chemical NER

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

- We train a BERT-CRF model for this task.



Event extraction

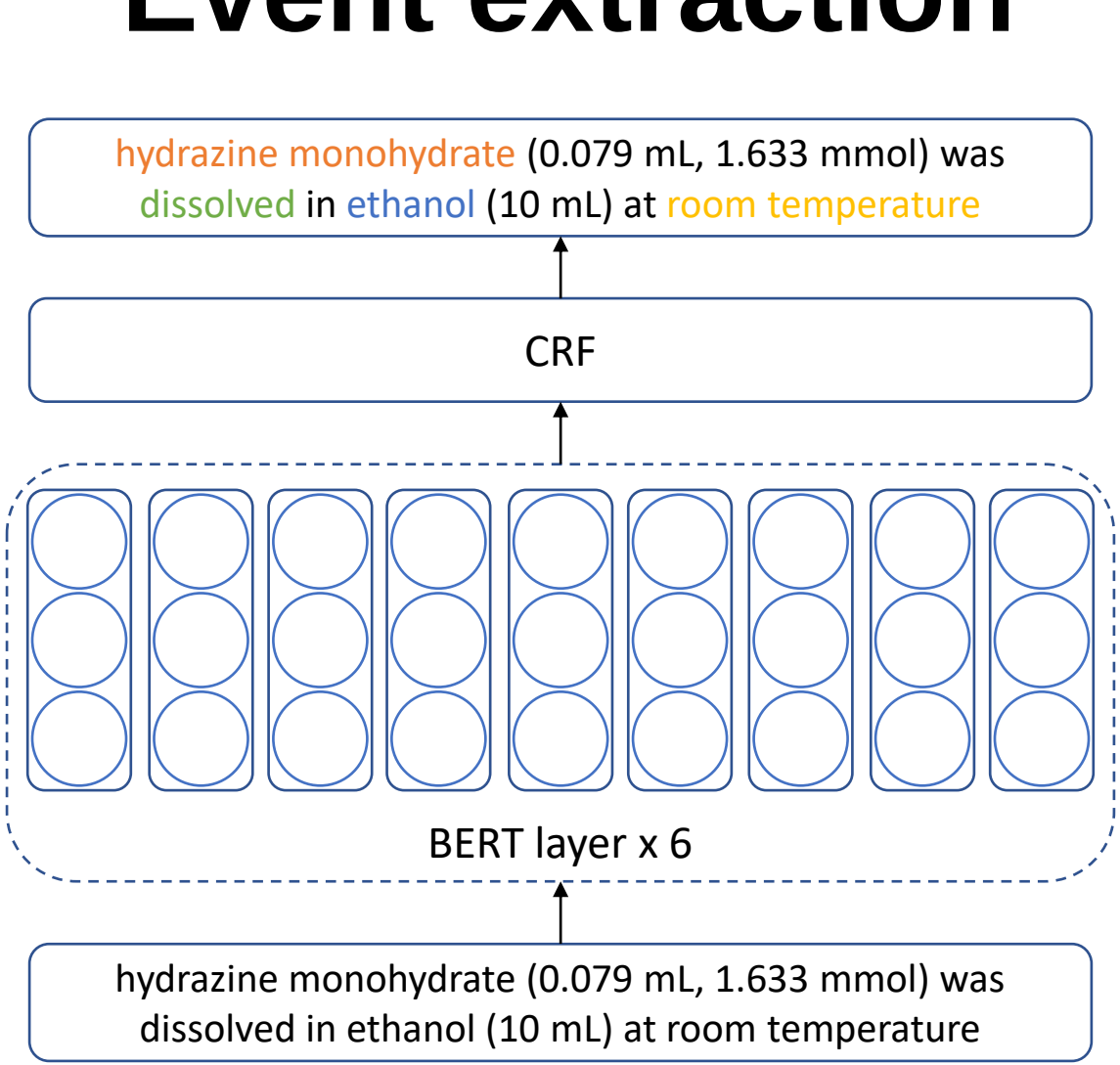
- A chemical reaction usually consists of an ordered sequence of *event steps* that
    1. transforms a starting material into a product (**REACTION_STEP**)
    2. just purifies or isolates a chemical substance (**WORKUP**)
- An event is characterised by a trigger word that flags its occurrence and a relation connecting the trigger word and chemical entities involved in the event.

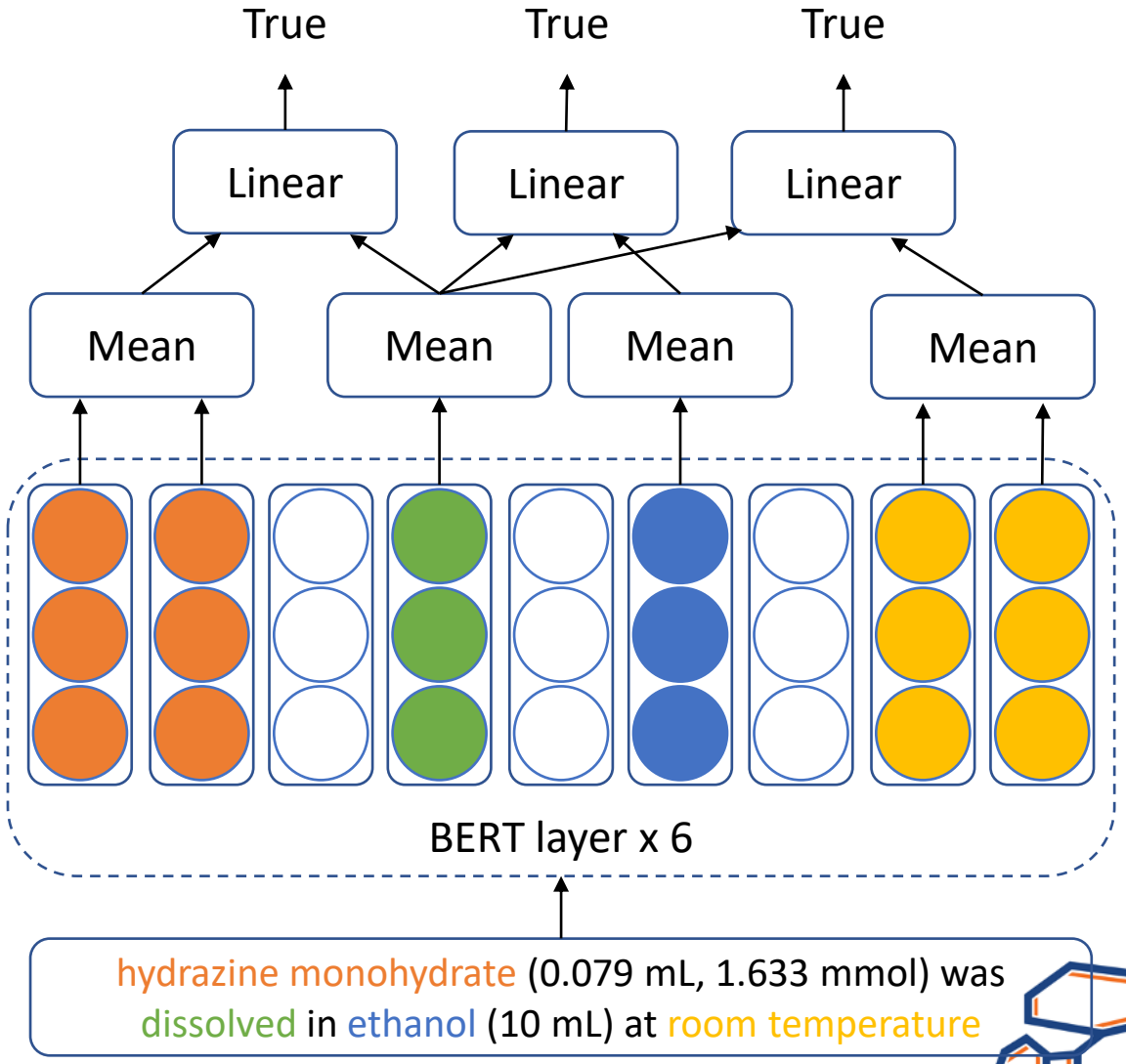
hydrazine monohydrate (0.079 mL, 1.633 mmol) was dissolved in ethanol (10 mL) at room temperature

REACTION_STEP: dissolved

- dissolved -> hydrazine monohydrate
- dissolved -> ethanol
- dissolved -> room temperature

Event extraction





Anaphora resolution

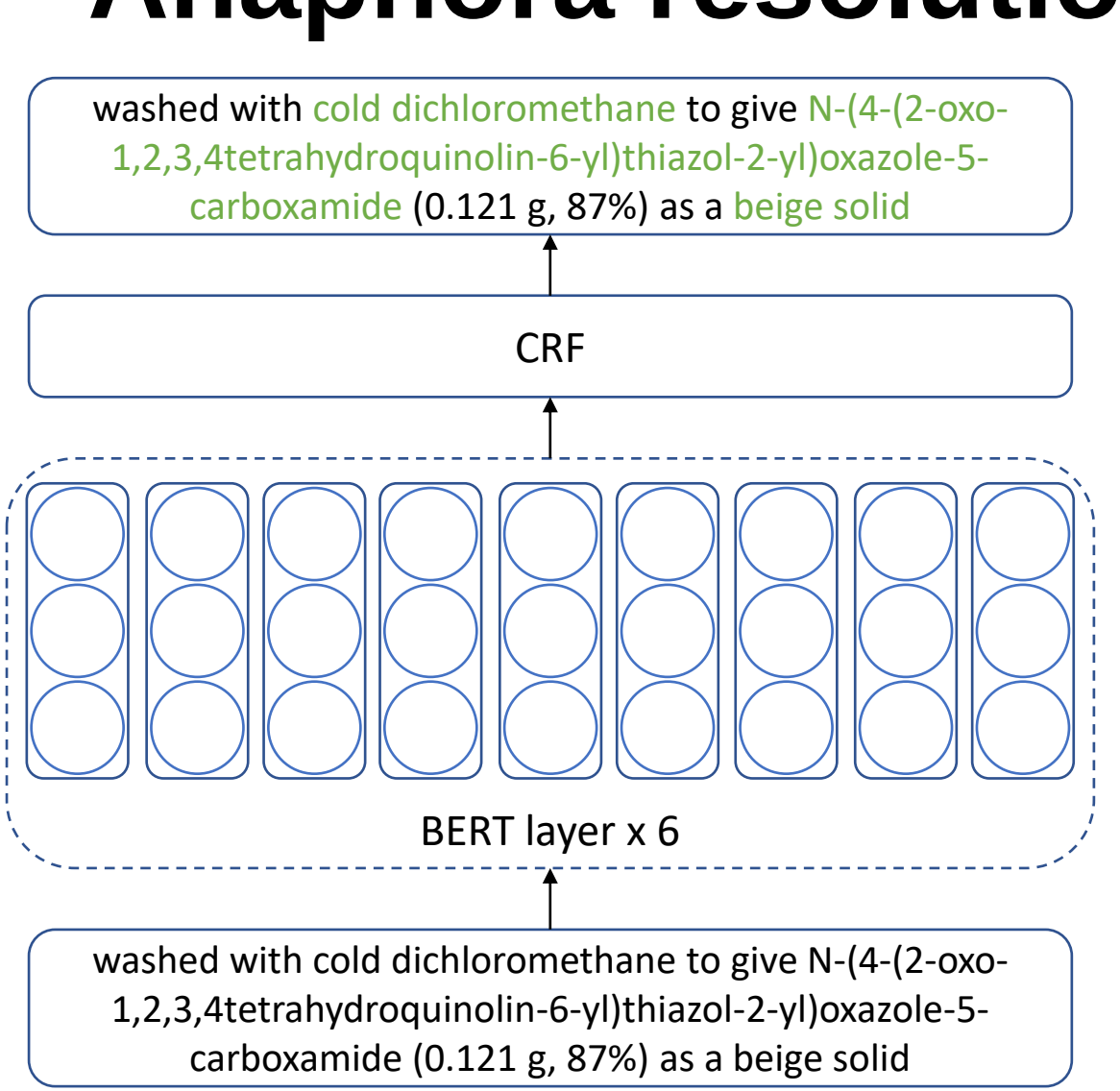
- There are rich anaphoric relations between and within event steps. We consider two main types of anaphoric relations:
  - coreference, where two mentions refer to the same entity, and
  - bridging, linking a chemical compound and its source.

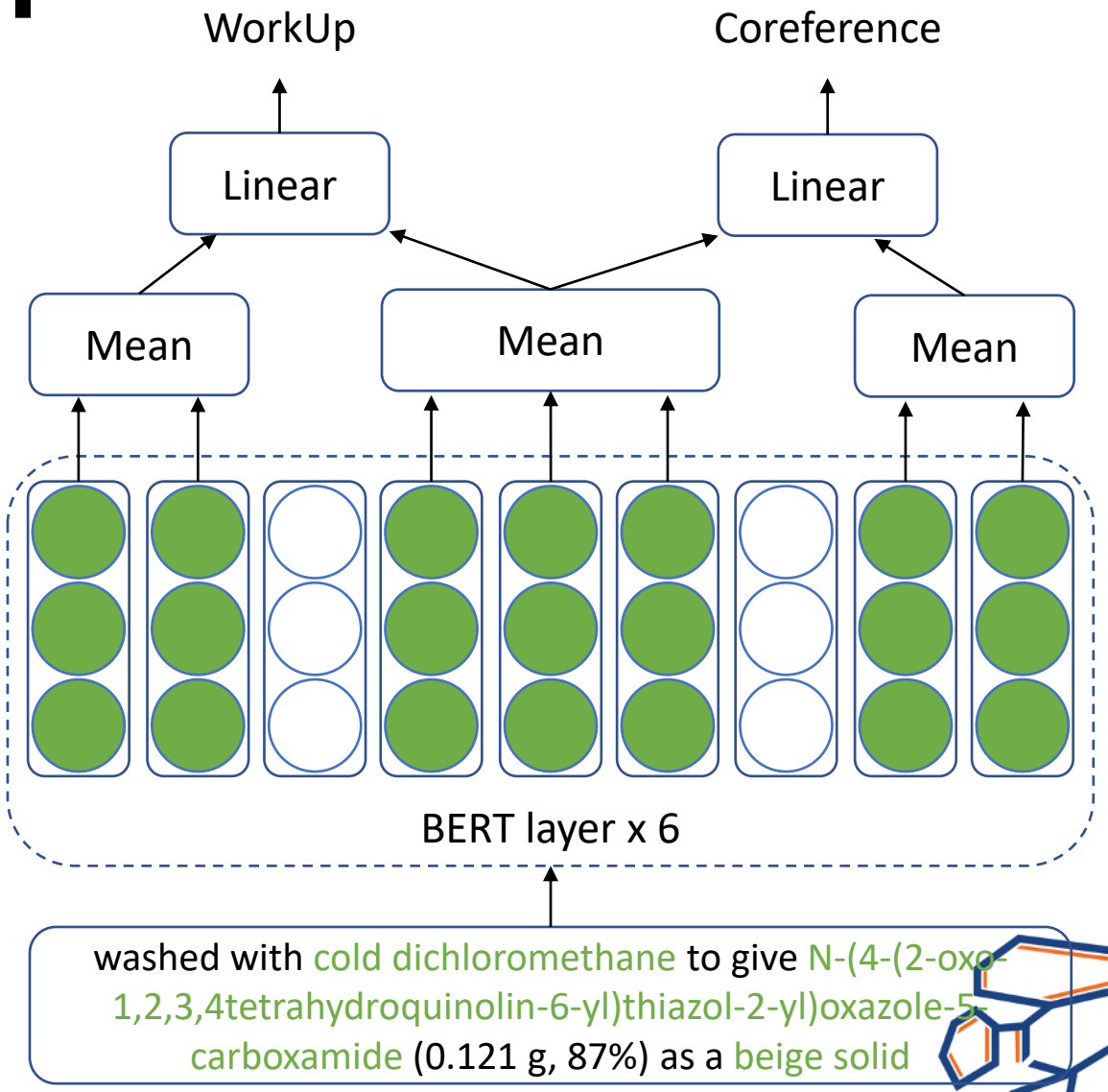
- We use a similar approach to NER+EE for this task.

washed with cold dichloromethane to give N-(4-(2-oxo-1,2,3,4tetrahydroquinolin-6-yl)thiazol-2-yl)oxazole-5-carboxamide (0.121 g, 87%) as a beige solid

- WorkUp: N-(4-(2-oxo-1,2,3,4tetrahydroquinolin-6-yl)thiazol-2-yl)oxazole-5-carboxamide -> cold dichloromethane
- Coreference: beige solid -> N-(4-(2-oxo-1,2,3,4tetrahydroquinolin-6-yl)thiazol-2-yl)oxazole-5-carboxamide

Anaphora resolution





Reaction reference resolution

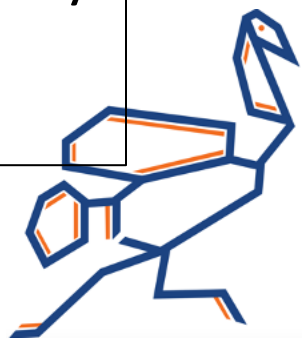
- Chemical patents often detail several similar compounds that have a common substructure and can be synthesized in analogous ways.
- They contain many references connecting descriptions of similar chemical reactions, to avoid redundancy in describing common reaction conditions.
- This leads to the problem of identifying references from an incomplete snippet to others.

Preparation 2

A solution of 62 g (0.23 mol) of the title product of Preparation 1 and 28 ml, (0.23 mol) of freshly distilled (S)-(-)-alpha-methyl benzylamine in 100 mL of toluene was heated to reflux, over a Dean-Stark trap…

Preparation 3

The title product of this preparation was prepared using a method analogous to **Preparation 2**, using (R)-(+)-alphamethyl benzylamine in the initial imine formation…



Reaction reference resolution

- We first determine if a snippet refers to others, and then enumerating possible reference pairs of snippets and classifying them.

- Since BERT runs quadratically in the length of input sequence, we use LSTM to encode the input reactions.

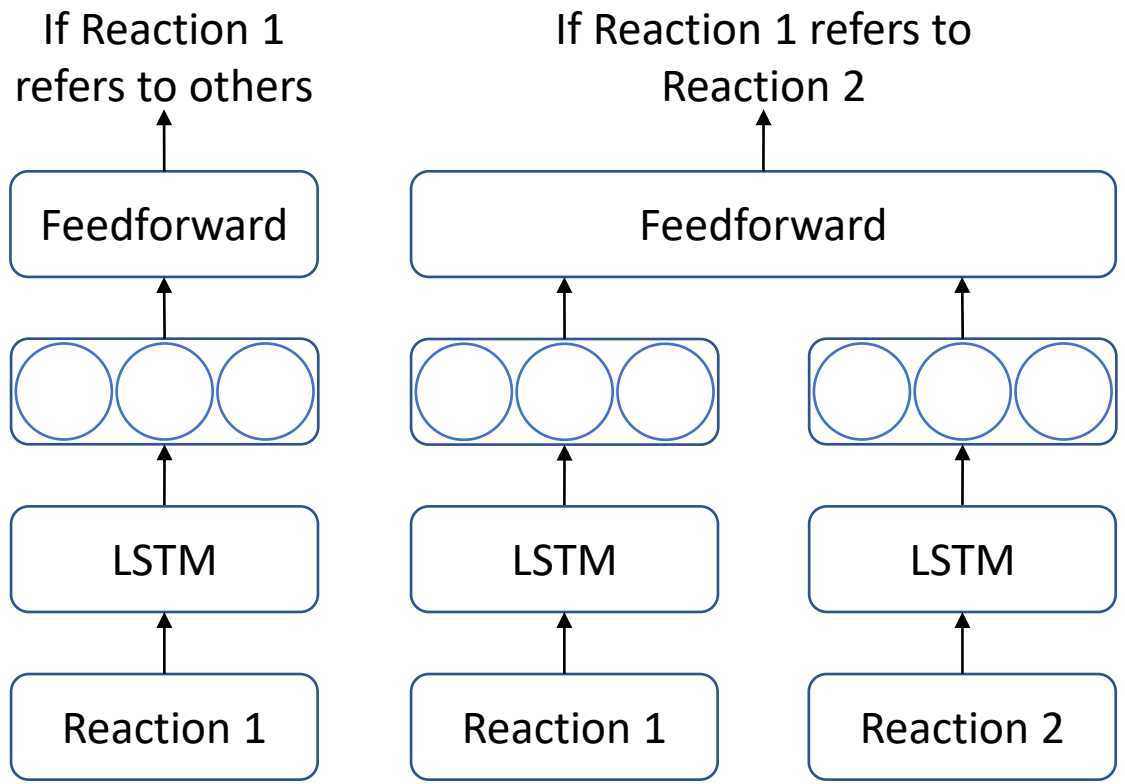


Table Classification

- Apart from text paragraphs, a large amount of information in patents is represented in tables and images.

- We focus on identifying tables containing chemical reaction properties such as starting materials, products, yields, etc.

Label	Description
SPECT	Spectroscopic data
PHYS	Physical data
IDE	Identification of compounds
RX	All properties of reactions
PHARM	Pharmacological data
CHEM	Chemical data
COMPOSITION	Compositions of mixtures
PROPERTY	Properties of chemicals
OTHER	Other tables

Table Classification

- We first linearize the table by concatenating tokens within all cells with [EOS] inserted between cells, [CLS] at the beginning, and [SEP] at the end.

- We then take the flatten table as input, encode it using BERT layers, and use a linear layer to classify the representation of [CLS] at the last layer into different categories.

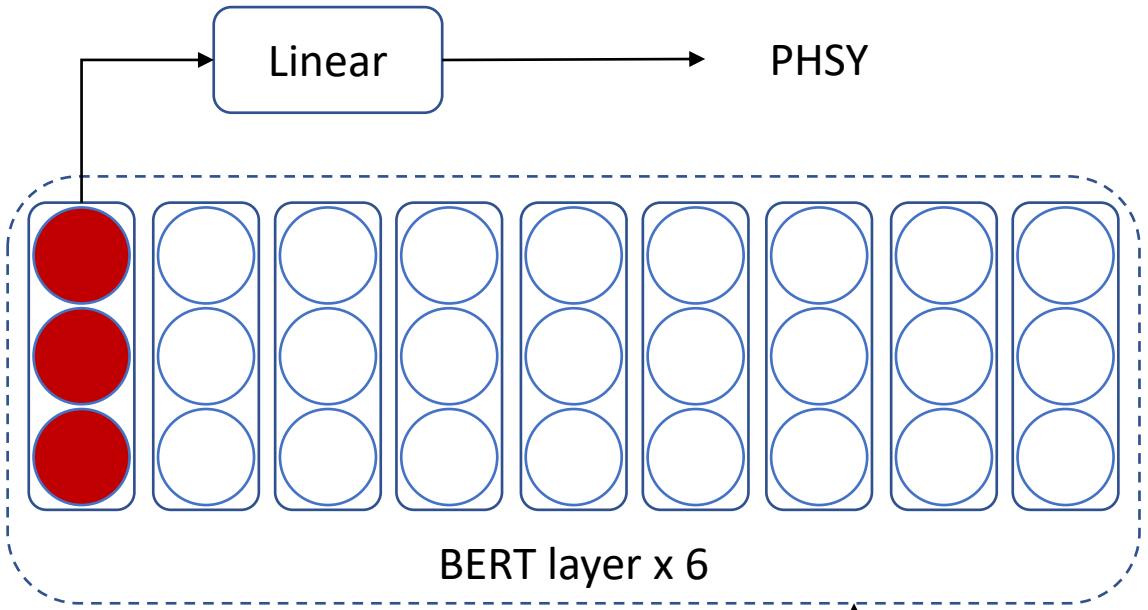
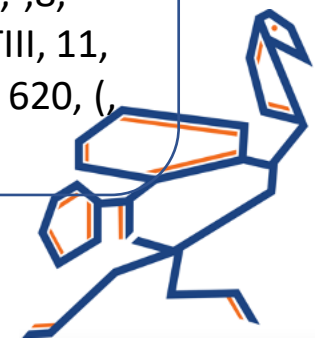


Table 1. Affinities to Heparin

Protein	Kd nM (ref)
PF4	27 (44)
IL-8	<5 (43)
ATIII	11 (42)
ApoE	620 (45)

[CLS] Table, 1, ., Affinities, to, Heparin, [EOS] Protein, Kd, nM, (, ref,) [EOS] PF4, 27, (, 44,), [EOS] IL,-,8, <,5,(, 43,) [EOS] ATIII, 11, (, 42,) [EOS] ApoE, 620, (, 45,) [SEP]



Conclusions

- We have introduced the essential requirements for building a comprehensive chemical reaction extraction system covering a wide range of tasks.
- We have proposed an initial approach for each step leveraging existing data resources from the ChEMU shared tasks, illustrating how the individual tasks can be brought together into a coherent whole.
- This integration addresses two key limitations of previous studies: our system can process full patent documents directly, and we can find the snippets an incomplete reaction snippet refers to.

Next steps

- We leave performance evaluation of individual steps, as well as the complete system, to a more in-depth presentation.
- In the future, we plan to further develop this framework to extract complete reaction information by incorporating inference over reaction references, and to extend the scope of our system to handle images and chemical structures.
- Opportunities also exist to explore joint modelling or multi-task learning across the constituent tasks in this pipeline, for instance coupling NER and anaphora resolution.