Extended Overview of ChEMU 2021: Reaction Reference Resolution and Anaphora Resolution in Chemical Patents

Yuan Li[1], Biaoyan Fang[1], Jiayuan He[1,2], Hiyori Yoshikawa[1,3], Saber A. Akhondi[4], Christian Druckenbrodt[5], Camilo Thorne[5], Zubair Afzal[4], Zenan Zhai[1], Timothy Baldwin[1] and Karin Verspoor[1,2]

[1]*The University of Melbourne, Australia*
[2]*RMIT University, Australia*
[3]*Fujitsu Limited, Japan*
[4]*Elsevier BV, Netherlands*
[5]*Elsevier Information Systems GmbH, Germany*

Abstract

In this paper, we provide an overview of the Cheminformatics Elsevier Melbourne University (ChEMU) evaluation lab 2021, part of the Conference and Labs of the Evaluation Forum 2021 (CLEF 2021). The ChEMU evaluation lab focuses on information extraction over chemical reactions from patent texts. As the second instance of our ChEMU lab series, we build upon the ChEMU corpus developed for ChEMU 2020, extending it for two distinct tasks related to reference resolution in chemical patents. Task 1 — Chemical Reaction Reference Resolution — focuses on paragraph-level references and aims to identify the chemical reactions or general conditions specified in one reaction description referred to by another. Task 2 — Anaphora Resolution — focuses on expression-level references and aims to identify the reference relationships between expressions in chemical reaction descriptions. Herein, we describe the resources created for these tasks and the evaluation methodology adopted. We also provide a brief summary of the results obtained in this lab, finding that one submission achieves substantially better results than our baseline models.

Keywords

Reaction reference resolution, Anaphora resolution, Chemical patents, Text mining, Information Extraction

1. Introduction

The discovery of new chemical compounds is perceived as a key driver of the chemical industry and many other industrial sectors, and information relevant for this discovery is found in chemical synthesis descriptions in natural language texts. In particular, patents serve as a critical source of information about new chemical compounds. Compared with journal publications, patents provide more timely and comprehensive information about new chemical compounds [1, 2, 3], since they are usually the first venues where new chemical compounds are disclosed.

CLEF 2021 – Conference and Labs of the Evaluation Forum, September 21–24, 2021, Bucharest, Romania
karin.verspoor@rmit.edu.au (K. Verspoor)
0000-0002-8661-1544 (K. Verspoor)

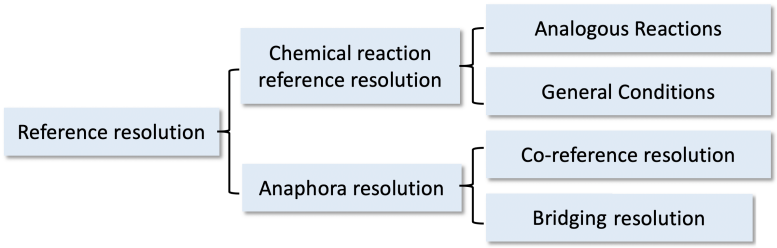


Figure 1: Illustration of the task hierarchy.

Despite the significant commercial and research value of the information in patents, manual extraction of such information is costly, considering the large volume of patents available [4, 5]. Thus, developing automatic natural language processing (NLP) systems for chemical patents, which 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[1]. ChEMU 2020 [6, 8, 9] provided two information extraction tasks, named entity recognition (NER) and event extraction, and attracted 37 teams around the world to participate. In the ChEMU 2021 lab, we provide two new information extraction tasks: chemical reaction reference resolution and anaphora resolution, focusing on reference resolution in chemical patents. Compared with previous shared tasks dealing with anaphora resolution, e.g., the CRAFT-CR task [10], our proposed tasks extend the scope of reference resolution by considering reference relationships on both paragraph-level and expression-level (see Fig. 1). Specifically, our first task aims at the identification of reference relationships between reaction descriptions. Our second task aims at the identification of reference relationships between chemical expressions, including both coreference and bridging. Moreover, we focus on chemical patents while the CRAFT-CR task focused on journal articles.

Unfortunately, we didn't receive any submissions to Task 1, chemical reaction reference resolution. The complexity of this task in particular combined with relatively short time periods for people to develop their systems may have made it difficult for people to participate. We plan to re-run it in 2022, to give the opportunity for more people to participate since the data and task definitions will have been around for a longer period of time. As a result, the remainder of this paper will focus on the second task, anaphora resolution.

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 task in Section 4 and detail the valuation framework of ChEMU in Section 5 including the evaluation methods and baseline models. We present the evaluation results in Section 6 and finally conclude this paper in Section 7.

[1]Our main website is http://chemu.eng.unimelb.edu.au

2. Related Shared Tasks

Several shared tasks have addressed reference resolution in scientific literature. BioNLP2011 hosted a subtask on protein coreference [11]. CRAFT 2019 hosted a subtask on coreference resolution (CRAFT-CR) in biomedical articles [10]. However, these shared tasks differ from ours in several respects.

First, previous shared tasks considered different domains of scientific literature. For example, the dataset used in BioNLP2011 is derived from the GENIA corpus [12], which primarily focuses on the biological domain, viz. gene/proteins and their regulations. The dataset used in CRAFT-CR shared task is based on biomedical journal articles in PubMed [13, 14]. Our ChEMU shared task, in contrast, focuses on the domain of chemical patents. This difference entails the critical importance for this shared task: information extraction methodologies for general scientific literature or the biomedical domain will not be effective for chemical patents [15]. It is widely acknowledged that patents are written quite differently as compared with general scientific literature, resulting in substantially different linguistic properties. For example, patent authors may trade some clarity in wording for more protection of their intellectual property.

Secondly, our reference resolution tasks include both paragraph-level and entity-level reference phenomena. Our first task aims at identification of reference relationships between reaction descriptions, i.e. paragraph-level. This task is challenging because a reaction description may refer to an extremely remote reaction and thus requires processing of very long documents. Our second task aims at anaphora resolution, similarly to previous entity-level coreference tasks. However, a key difference is that we extend the scope of this task by including both coreference and bridging phenomena. That is, we not only aim at finding expressions referring to the same entity, but also expressions that are semantically related or associated.

3. The ChEMU Chemical Reaction Corpus

In this section, we explain how the dataset is created for the anaphora resolution task. The complete annotation guidelines are made available at [16].

3.1. Corpus Selection

We build on the ChEMU corpus [17] developed for the ChEMU 2020 shared task [18]. The ChEMU corpus contains patents from the European Patent Office and the United States Patent and Trademark Office, available in English in a digital format. It is based on the Reaxys® database,[2] containing reaction entries for patent documents manually created by experts in chemistry. It consists of 'snippets' extracted from chemical patents, where each snippet corresponds to a reaction description. It is common that several snippets are extracted from the same chemical patent.

[2]Reaxys® Copyright ©2021 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

3.2. Mention Type

We aim to capture anaphora in chemical patents, with a focus on identifying chemical compounds during the reaction process. Consistent with other anaphora corpora [19, 13, 20], only mentions that are involved in referring relationships (as defined in Section 3.3) and related to chemical compounds are annotated. The mention types that are considered for anaphora annotation are listed below. 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.

3.2.1. Chemical Names

Chemical names are a critical component of chemical patents. We capture as atomic mentions the formal name of chemical compounds, e.g. *N-[4-(benzoxazol-2-yl)-methoxyphenyl]-S-methyl-N'-phenyl-isothiourea* or *2-Chloro-4-hydroxy-phenylboronic acid*. Chemical names often include nested chemical components, but for the purposes of our corpus, we consider chemical names to be atomic and do not separately annotate internal mentions. Hence *4-(benzoxazol-2-yl)-methoxyphenyl* and *acid* in the examples above will not be annotated as mentions, as they are part of larger chemical names.

3.2.2. Identifiers

In chemical patents, identifiers or labels may also be used to represent chemical compounds, in the form of uniquely-identifying sequences of numbers and letters such as *5i*. These can be abbreviations of longer expressions incorporating that identifier that occur earlier in the text, such as *chemical compound 5i*, or may refer back to an exact chemical name with that identifier. Thus, the identifier is annotated as an atomic mention as well.

3.2.3. Phrases and Noun Types

Apart from chemical names and identifiers, chemical compounds are commonly presented as noun phrases (NPs). An NP consists of a noun or pronoun, and premodifiers; NPs are the most common type of compound expressions in chemical patents. Here we detail NPs that are related to compounds:

1. Pronouns: In chemical patents, pronouns (e.g. *they* or *it*) usually refer to a previously mentioned chemical compounds.
2. Definite and indefinite NPs: Commonly used to refer to chemical compounds, e.g. *the solvent*, *the title compound*, *the mixture*, and *a white solid*, *a crude product*.

Furthermore, there are a few types of NPs that need specific handling in chemical patents:

1. Quantified NPs: Chemical compounds are usually described with a quantity. NPs with quantities are considered as atomic mentions if the quantities are provided, e.g. *398.4 mg of the compound 1*.
2. NPs with prepositions: Chemical NPs connected with prepositions (e.g. *in*, *with*, *of*) can be considered as a single mention. For example, the phrase *2,4-dichloro-6-(6-triuoromethylpyridin-2-yl)-1,3,5-triazine (5.0 g, 16.9 mmol) in tetrahydrofuran (100 mL)* is a single mention,

as it describes a solvent that contains *2,4-dichloro-6-(6-triuoromethylpyridin-2-yl)-1,3,5-triazine (5.0 g, 16.9 mmol)* and *tetrahydrofuran (100 mL)*.

NPs describing chemical equipment containing a compound may also be relevant to anaphora resolution. This generally occurs when the equipment that contains the compound undergoes a process that also affects the compound. Thus, equipment expressions such as the flask and the autoclave can also be mentions if they are used to implicitly refer to a contained compound.

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.

3.2.4. Relationship to ChEMU 2020 entities

Since this dataset is built on the ChEMU 2020 corpus [17], annotation of related chemical compounds is available by leveraging existing entity annotations introduced for the ChEMU 2020 named entity recognition (NER) task. However, there are some differences in the definitions of entities for the two tasks.

In the original ChEMU 2020 corpus, 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 our 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.

3.3. Relation Types

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.

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.

3.3.1. Coreference

Coreference is defined as expressions/mentions that refer to the same entity [21, 22]. In chemistry, identifying whether two mentions refer to the same entity needs to consider various

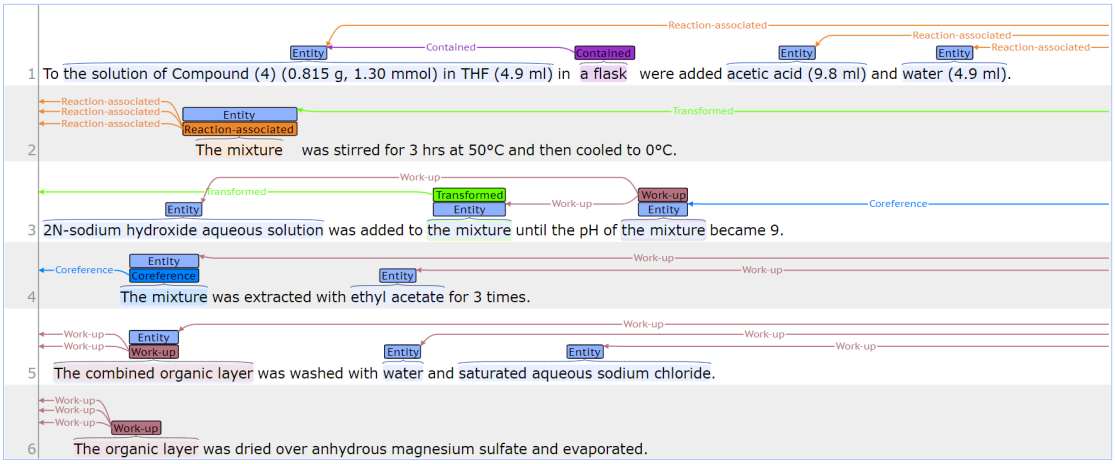


Figure 2: Annotated snippet of anaphora resolution in the chemical patents. The figure is taken from [23]. Different color of links represent different anaphora relation types.

chemical properties (e.g. temperature or pH). As such, for two mentions to be coreferent, they must share the same chemical properties. We consider two different cases of coreference:

1. Single Antecedents: the anaphor refers to a single antecedent.
2. Multiple Antecedents: the anaphor refers to multiple antecedents, e.g. *start materials* refers to all the chemical compounds or materials that are used at the beginning.

It is possible for there to be ambiguity as to which mention of a given antecedent an anaphor refers to (where the mention is identical); in these cases the closest mention is selected.

3.3.2. Bridging

As stated above, when we consider the anaphora relations, we take the chemical properties of the mention into consideration. Coreference is insufficient to cover all instances of anaphora in chemical patents, and bridging occurs frequently. We define four bridging types:

TRANSFORMED Links between chemical compounds that are initially based on the same components, but which have undergone a change in condition, such as pH or temperature. Such cases must be one-to-one relations (not one-to-many). As shown in Figure 2, the *mixture* in line 2 and the first-mentioned *mixture* in line 3 have the TRANSFORMED relation, as they have the same chemical components but different chemical properties.

REACTION-ASSOCIATED The relationship between a chemical compound and its immediate source compounds is via a mixing process, where the source compounds retain their original chemical structure. This relation is one-to-many from the anaphor to the source compounds (antecedents). For example, the *mixture* in line 2 has REACTION-ASSOCIATED links to three mentions on line 1 that are combined to form it: (1) *the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml)*; (2) *acetic acid (9.8 ml)*; and (3) *water (4.9 ml)).*

WORK-UP Chemical compounds are used to isolate or purify an associated output product, in a one-to-many relation, from the anaphor to the compounds (antecedents) that are used for the work-up process. As demonstrated in Figure 2, *The combined organic layer* in line 5 comes from the extraction of *The mixture and ethyl acetate* in line 4, and they are hence annotated as WORK-UP.

CONTAINED A chemical compound is contained inside equipment. It is a one-to-many relation from the anaphor (equipment) to the compounds (antecedents) that it contains. An example of this is *a flask* and *the solution of Compound (4) (0.815 g, 1.30 mmol) in THF (4.9 ml)* on line 1, where the compound is contained in the flask.

3.4. Annotation Process

For the corpus annotation, we use the BRAT text annotation tool.[3] In total 1500 snippets have been annotated by two chemical experts, a PhD candidate and a final year bachelor student in Chemistry. A draft of the annotation guideline was created and refined with chemical experts, then four rounds of annotation training were completed prior to beginning official annotation. In each round, the two annotators individually annotated the same 10 snippets (different across each round of annotation), and their annotations were compared and combined by an adjudicator; annotation guidelines were then refined based on discussion. After several rounds of training, we achieved a high inner-annotator agreement of Krippendorff's $\alpha = 0.92$ [24] at the mention annotation level,[4] and $\alpha = 0.84$ for relations. Finally, the development and test sets were double annotated by the two expert annotators, with any disagreements merged by the adjudicator.

3.5. Data Partitions

We randomly partitioned the whole dataset 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 average number of tokens per sentence, are summarized in Table 1.

To ensure the snippets included in the training, development, and test splits have similar distributions, we compare the distribution of relation types (five types of relations in total). Based on the numbers in Table 1, we confirm that the label distribution in the three splits are similar, with very little variation ($\leq 2\%$) across the three splits observed for each relation type.

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

[4] With the lowest agreement being $\alpha = 0.89$ for coreference mentions.

Table 1
Corpus annotation statistics.

	Training	Development	Test
Snippets	6392	1535	2585
Sentences	763	164	274
Tokens/Sentences	15.8	15.2	15.8
Mentions	19626	4515	7810
Discontinuous mentions	876	235	399
Coreference	3568	870	1491
Bridging	10377	2419	4135
Transformed	493	107	166
Reaction-associated	3308	764	1245
Work-up	6230	1479	2576
Contained	346	69	148

4. Task definition

This task requires the resolution of general anaphoric dependencies between expressions in chemical patents. Five types of anaphoric relationships are defined:

1. *Coreference*: two expressions/mentions that refer to the same entity.
2. *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).
3. *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.
4. *Work-up*: the relationship between chemical compounds that were used for isolation or purification purposes, and their corresponding output products.
5. *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.

Taking the text snippet in Figure 3 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 **[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".

[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 3: 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.

5. Evaluation Framework

5.1. Evaluation Methods

We use BRATEval[5] 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.

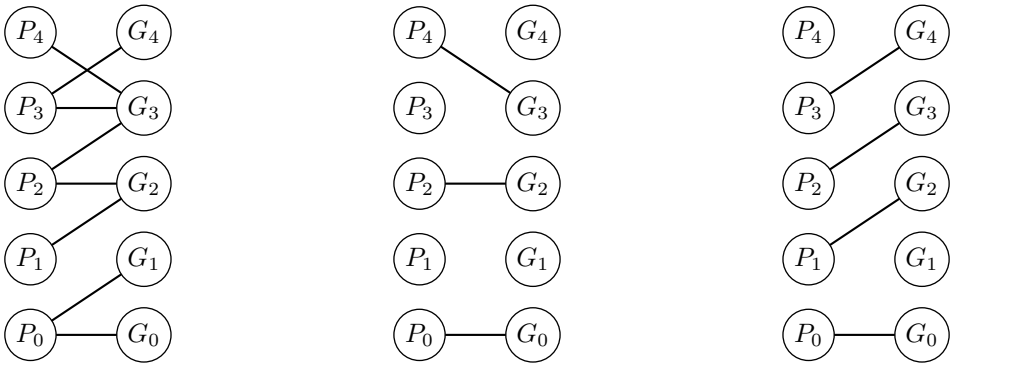
Furthermore, let $R = (RT, E^{ana}, E^{ant})$ be a relation where RT is the type of R, E^{ana} the anaphor of R, E^{ant} the antecedent of R. Then R_1 and R_2 are exactly matched ($R_1 = R_2$) if $RT_1 = RT_2$, $E_1^{ana} = E_2^{ana}$, and $E_1^{ant} = E_2^{ant}$. While R_1 and R_2 are approximately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

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

[5] 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 4: An example matching graph and two bipartite matching for it.

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 4a 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 4b shows a matching with $TP = 3$ but is not optimal. Figure 4c 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

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

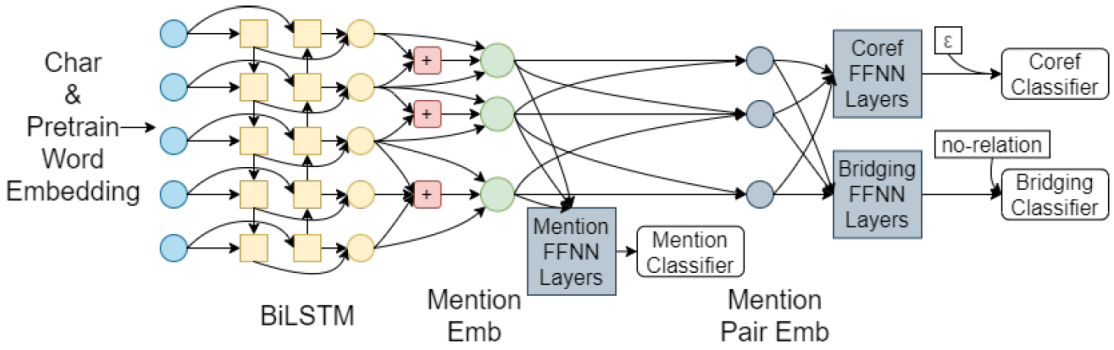


Figure 5: The architecture of our baseline model. The figure is taken from [23].

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

Our baseline model adopts an end-to-end architecture for coreference resolution [25, 26], as depicted in Figure 5. Following the methods presented in [23], we use GloVe embeddings and a 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.

Table 2
Overall performance for all runs on the test set. Here P, R, and F are short for Precision, Recall, and F_1 score. For each metric, the best result is highlighted in **bold**.

	Exact-Match			Relaxed-Match		
	P	R	F	P	R	F
CMU	0.8177	**0.7542**	**0.7847**	**0.909**	**0.8384**	**0.8723**
Baseline-ChELMO	**0.8566**	0.6882	0.7633	0.9024	0.725	0.8041
Baseline-ELMO	0.8435	0.6676	0.7453	0.8875	0.7025	0.7842
HUKB	0.7132	0.6696	0.6907	0.7702	0.7231	0.7459

We released the code for training our baseline models to help the participants to get started on the shared task.[6] Two variants of the baseline model are evaluated on the test set, one using the ELMO embeddings as input to the BiLSTM component, while the other used pretrained ChELMO, based on the embeddings of [27] pre-trained on chemical patents, with the hope of benefiting more from domain-specific pretraining.

6. Results and Discussions

A total of 19 teams registered on our submission website for the shared task. Among them, we finally received 2 submissions on the test set. One team is from Carnegie Mellon University, US (CMU) and the other one is from Hokkaido University, Japan (HUKB). More details about their systems are provided in Section 7. In this section, we report their results along with the performance of our two baseline systems.

We report the overall performance of all runs in Table 2. The rankings of different systems are fully consistent across all metrics. The CMU team achieves an F_1 score of 0.7847 in exact matching, outperforming our two baselines which get 0.7633 and 0.7453, followed by the HUKB team who obtains 0.6907. The lead of the CMU team is even larger in relaxed matching, with an F_1 score of 0.8723, about 7 points higher than our baselines. This shows the potential of the CMU model 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 baselines have higher precision in the exact setting and precision in relaxed setting is also very close to the best, which indicates that our models are more conservative and could possibly be enhanced by making more aggressive predictions to improve recall. The use of domain-pretrained embeddings (ChELMO vs. ELMO) does, as expected, benefit performance.

Table 3 provides more details about the performance of all models for each relation type. The CMU team outperforms others on TRANSFORMED relation by a large margin. While our baselines performs the best on CONTAINED relation type. For the other three relation types, the CMU model wins F_1 score and recall, while our models achieve the highest precision, which is similar to our observation on the overall results. Given that the models perform very differently, it would be very interesting to do more analysis when the details of all the models

[6]Code available at https://github.com/biaoyanf/ChEMU-Ref

Table 3
Performance per relation type for all runs on the test set. Here P, R, and F are short for Precision, Recall, and F_1 score. For each metric, the best result is highlighted in **bold**.

		Exact-Match			Relaxed-Match		
		P	R	F	P	R	F
COREFERENCE	CMU	0.7568	**0.5822**	**0.6581**	0.8945	**0.6881**	**0.7779**
	Baseline-ChELMO	0.8476	0.4661	0.6015	**0.9244**	0.5084	0.656
	Baseline-ELMO	**0.8497**	0.4474	0.5861	0.9185	0.4836	0.6336
	HUKB	0.6956	0.5319	0.6028	0.7868	0.6016	0.6819
CONTAINED	CMU	0.7727	0.6892	0.7286	0.8561	**0.7635**	0.8071
	Baseline-ChELMO	**0.9211**	**0.7095**	**0.8015**	0.9386	0.723	**0.8168**
	Baseline-ELMO	0.9175	0.6014	0.7265	**0.9794**	0.6419	0.7755
	HUKB	0.7214	0.6824	0.7014	0.7929	0.75	0.7708
REACTION_ ASSOCIATED	CMU	0.8037	**0.7631**	0.7829	**0.9019**	**0.8562**	**0.8785**
	Baseline-ChELMO	**0.8381**	0.7357	**0.7836**	0.8673	0.7614	0.8109
	Baseline-ELMO	0.8145	0.7229	0.766	0.8498	0.7542	0.7991
	HUKB	0.668	0.6803	0.6741	0.7224	0.7357	0.729
TRANSFORMED	CMU	**0.9423**	**0.8855**	**0.913**	**0.9423**	**0.8855**	**0.913**
	Baseline-ChELMO	0.7935	0.8795	0.8343	0.7935	0.8795	0.8343
	Baseline-ELMO	0.7877	0.8494	0.8174	0.7877	0.8494	0.8174
	HUKB	0.6611	0.7169	0.6879	0.6611	0.7169	0.6879
WORK_UP	CMU	0.846	**0.8447**	**0.8454**	**0.9195**	**0.9181**	**0.9188**
	Baseline-ChELMO	**0.8705**	0.7803	0.8229	0.9181	0.823	0.868
	Baseline-ELMO	0.8566	0.7605	0.8057	0.899	0.7981	0.8456
	HUKB	0.7467	0.7403	0.7435	0.7929	0.7861	0.7895

are disclosed, and hopefully every team can borrow ideas from others and further improve the performance.

7. Overview of Participants' Approaches

We received paper submissions from both the participating teams, i.e. the HUKB team and the CMU team. We first describe their approaches, then summarize the same and different aspects of them.

7.1. HUKB

The HUKB team used a two-step approach for the anaphora resolution task, where mentions including both antecedent and anaphor are first detected, then mentions are classified into different types and relations between them are determined. For step one, they found that although existing parsers such as Chemical Tagger can generate useful features, they are not enough for mention detection. Therefore they trained a BioBERT model to find candidate mentions. This is done by treating mention detection as a NER task, where a BIOHD format is

used to convert gold spans into sequence of labels. The BIOHD format is an extension to the well-known BIO format to support discontinuous spans which exist in this amphora resolution task. In the second step, different types of relations are determined based on different rules. COREFERENCE is first detected by 5 regular expression rules. The remaining relations are detected based on the features generated by ChemicalTagger. When no more relations can be found, a post-processing step is carried out to handle the transitivity property of COREFERENCE relations, i.e. enumerating all valid COREFERENCE relations based on the transitivity property.

7.2. CMU

The CMU team proposed a pipelined system for anaphora resolution, where mentions are first extracted and then relations between them are determined. The first step is done by a BERT-CRF model which is trained using BIO tagging. In the second step, for each pair of mentions, the sequence of sentences that contain the pair is fed to a BERT model to obtain the encoded representation of every token, then the representation of a mention is simply the mean of all tokens in it, and finally the representations of two mentions are concatenated and classified into 6 classes using a linear layer (5 relations + a class for no relation). To correct boundary errors in mention detection, a rule-based post processing is done between step 1 and 2. Furthermore, ensembling of 5 models are used in both step 1 and 2 to improve performance.

7.3. Summary

Both of them adopted a two-step approach where mentions are first detected and then relations between them are determined. They also both relied on BERT-like models to extract contextualized representations for mention detection. While the CMU team used a BERT-like model in the relation extraction, the HUKB team chose a rule-based method. In addition, the CMU team used ensembling of 5 models with majority voting in both mention detection and relation extraction, as well as a post-processing step between them to correct potential boundary errors in mention detection, where both techniques contribute to their superior overall performance.

8. Conclusions

This paper presents a general overview of the activities and outcomes of the ChEMU 2021 evaluation lab. As the second instance of our ChEMU lab series, ChEMU 2021 targets two new tasks focusing on reference resolution in chemical patents. Our first task aims at identification of reference relationships between chemical reaction descriptions, and our second task aims at identification of reference relationships between expressions in chemical reactions. The evaluation result includes different approaches to tackling the shared task, with one submission clearly outperforming our baseline methods. We look forward to fruitful discussion and deeper understanding of the methodological details of these submissions at the workshop.

Acknowledgements

Funding for the ChEMU project is provided by an Australian Research Council Linkage Project, project number LP160101469, and Elsevier. We acknowledge the support of our ChEMU-Ref annotators, Dr. Sacha Novakovic and Colleen Hui Shiuan Yeow at the University of Melbourne, and the annotation teams supporting the reaction reference task annotation.

References

[1] 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).

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

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

[5] 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.

[6] 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. URL: https://doi.org/10.3389/frma.2021.654438. doi:10.3389/frma.2021.654438.

[7] 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) S1.

[8] 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, Overview of ChEMU 2020: Named entity recognition and event extraction of chemical reactions from patents, in: Experimental IR Meets Multilinguality, Multimodality, and Interaction. Proceedings of the Eleventh International Conference of the CLEF Association (CLEF 2020), volume 12260, Lecture Notes in Computer Science, 2020.

[9] D. Q. Nguyen, Z. Zhai, H. Yoshikawa, B. Fang, C. Druckenbrodt, C. Thorne, R. Hoessel, S. A. Akhondi, T. Cohn, T. Baldwin, et al., ChEMU: Named entity recognition and event extraction of chemical reactions from patents, in: European Conference on Information Retrieval, Springer, 2020, pp. 572–579.

[10] W. A. Baumgartner Jr, M. Bada, S. Pyysalo, M. R. Ciosici, N. Hailu, H. Pielke-Lombardo, M. Regan, L. Hunter, CRAFT shared tasks 2019 overview—integrated structure, semantics,

and coreference, in: Proceedings of The 5th Workshop on BioNLP Open Shared Tasks, 2019, pp. 174–184.

[11] N. Nguyen, J.-D. Kim, J. Tsujii, Overview of BioNLP 2011 protein coreference shared task, in: Proceedings of BioNLP Shared Task 2011 Workshop, 2011, pp. 74–82.

[12] T. Ohta, Y. Tateisi, J.-D. Kim, H. Mima, J. Tsujii, The GENIA corpus: An annotated research abstract corpus in molecular biology domain, in: Proceedings of the Second International Conference on Human Language Technology Research, 2002, pp. 82–86.

[13] 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.

[14] M. Bada, M. Eckert, D. Evans, K. Garcia, K. Shipley, D. Sitnikov, J. Baumgartner, W. A., K. B. Cohen, K. Verspoor, J. A. Blake, L. E. Hunter, Concept annotation in the CRAFT corpus, BMC Bioinformatics 13 (2012) 161. URL: https://www.ncbi.nlm.nih.gov/pubmed/22776079. doi:10.1186/1471-2105-13-161.

[15] M. Lupu, K. Mayer, N. Kando, A. J. Trippe, Current challenges in patent information retrieval, volume 37, Springer, 2017.

[16] B. Fang, C. Druckenbrodt, C. Yeow Hui Shiuan, S. Novakovic, R. Hössel, S. A. Akhondi, J. He, M. Mistica, T. Baldwin, K. Verspoor, Chemu-ref dataset for modeling anaphora resolution in the chemical domain, 2021. doi:10.17632/r28xxr6p92.

[17] 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, 2020. doi:10.17632/wy6745bjfj.

[18] 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, Overview of chemu 2020: Named entity recognition and event extraction of chemical reactions from patents, in: A. Arampatzis, E. Kanoulas, T. Tsikrika, S. Vrochidis, H. Joho, C. Lioma, C. Eickhoff, A. Névéol, L. Cappellato, N. Ferro (Eds.), Experimental IR Meets Multilinguality, Multimodality, and Interaction - 11th International Conference of the CLEF Association, CLEF 2020, Thessaloniki, Greece, September 22-25, 2020, Proceedings, volume 12260 of *Lecture Notes in Computer Science*, Springer, 2020, pp. 237–254. URL: https://doi.org/10.1007/978-3-030-58219-7_18. doi:10.1007/978-3-030-58219-7_18.

[19] 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/.

[20] 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.

[21] V. Ng, Machine learning for entity coreference resolution: A retrospective look at two decades of research, in: S. P. Singh, S. Markovitch (Eds.), Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence, February 4-9, 2017, San Francisco, California, USA, AAAI Press, 2017, pp. 4877–4884. URL: http://aaai.org/ocs/index.php/AAAI/AAAI17/paper/view/14995.

[22] K. Clark, C. D. Manning, Entity-centric coreference resolution with model stacking, in: Proceedings of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing of the Asian Federation of Natural Language Processing, ACL 2015, July 26-31, 2015, Beijing, China, Volume 1: Long Papers, The Association for Computer Linguistics, 2015, pp. 1405–1415. URL: https://doi.org/10.3115/v1/p15-1136. doi:10.3115/v1/p15-1136.

[23] 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.

[24] K. Krippendorff, Measuring the reliability of qualitative text analysis data, Quality and quantity 38 (2004) 787–800.

[25] 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.

[26] 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.

[27] Z. Zhai, D. Q. Nguyen, S. Akhondi, C. Thorne, C. Druckenbrodt, T. Cohn, M. Gregory, K. Verspoor, Improving chemical named entity recognition in patents with contextualized word embeddings, in: Proceedings of the 18th BioNLP Workshop and Shared Task, Association for Computational Linguistics, Florence, Italy, 2019, pp. 328–338. URL: https://www.aclweb.org/anthology/W19-5035. doi:10.18653/v1/W19-5035.