Chemical Reaction Reference Resolution in Patents

Hiyori Yoshikawa*
Fujitsu Limited
Tokyo, Japan
y.hiyori@fujitsu.com

Saber Akhondi
Elsevier BV
Amsterdam, Netherlands
s.akhondi@elsevier.com

Camilo Thorne
Elsevier Information Systems GmbH
Frankfurt, Germany
c.thorne.1@elsevier.com

Christian Druckenbrodt
Elsevier Information Systems GmbH
Frankfurt, Germany
c.druckenbrodt@elsevier.com

Ralph Hoessel
Elsevier Information Systems GmbH
Frankfurt, Germany
r.hoessel@elsevier.com

Zenan Zhai
The University of Melbourne
Melbourne, Australia
zenan.zhai@student.unimelb.edu.au

Jiayuan He†
RMIT University
Melbourne, Australia
estrid.he@rmit.edu.au

Timothy Baldwin
The University of Melbourne
Melbourne, Australia
tbaldwin@unimelb.edu.au

Karin Verspoor‡
RMIT University
Melbourne, Australia
karin.verspoor@unimelb.edu.au

ABSTRACT

Many new chemical compounds are reported each year in patent documents, leading to increasing demand for methods for automatic information extraction of chemical compounds and reactions from patents. Chemical patents often detail a number of similar compounds that have a common substructure and can be synthesized in analogous ways, and therefore contain many references connecting descriptions of similar chemical reactions, to avoid redundancy in describing common reaction conditions. This leads to the problem of *reaction reference resolution*, where, given a reaction description, we need to identify links to other reaction descriptions it refers to. In this paper, we formally introduce the task and propose baseline methods to address it in analogy with co-reference resolution. To evaluate the performance, we create a large-scale silver-standard dataset based on a commercial database of chemical reactions. The experimental results show that the approach based on a state-of-the-art co-reference resolution method struggles to outperform a simple heuristic in detecting reference links, demonstrating the difficulty of the proposed task and its fundamentally different nature to co-reference resolution.

CCS CONCEPTS

• **Computing methodologies → Information extraction**.

KEYWORDS

information extraction, reaction reference resolution, natural language processing

---

*Also with The University of Melbourne.
†Also with The University of Melbourne.
‡Also with The University of Melbourne.

---

1 INTRODUCTION

Patents represent a critical source of information about new chemical compounds, and lead journal publications in both volume and time [2]. Patents provide not only the name and characteristics of new chemical compounds, but also the reaction details for their synthesis, including starting materials, product, reagents, catalysts, solvents, and the conditions of the reactions such as temperature and time. Databases such as Reaxys®[1] and CASREACT [5] store a large volume of such chemical reaction information. Given the commercial and research value of the information in patents, as well as the large numbers of available patents, developing methods for automatic information extraction from chemical patents has been a focus of recent research [17, 24].

A text mining system for chemical reaction information extraction should aim to obtain details of each reaction, requiring: (1) the identification of specific spans of text describing reactions; and (2) the extraction of the compounds participating in, and the conditions of, a reaction. The second step involves named entity recognition [6, 8, 13, 16, 24], relation extraction [15, 31], entity linking [3, 27], and event extraction [8, 24, 30], which have been explored intensively in previous studies. A relatively small number of studies have addressed the chemical reaction extraction task as a whole [1, 10, 19, 21].

The first step of locating all information relevant to a chemical reaction has received relatively little attention. This step is a challenge as it requires processing very long and semantically unstructured documents, in which a number of similar chemical reactions are presented. Yoshikawa et al. [32] proposed the chemical reaction detection task to locate descriptions of individual chemical reactions in patents, and achieved promising results using a contextualized document modeling method. Jessop et al. [10] also addressed the detection of chemical reaction descriptions in a heuristic way. However, previous work has not addressed the important step of *resolving references between reaction descriptions*. In general, a chemical patent reports a number of similar compounds that have a common chemical substructure and can be synthesized in analogous ways. Synthesis steps that are common across similar reactions may be described

---

only once, and the paragraphs describing individual reactions refer back to those descriptions to provide specific details without repeating information. In our dataset described below, approximately 17% of reaction descriptions contain references to other reaction descriptions.

To fill this gap in previous work and take a step towards an end-to-end method for chemical reaction information extraction, this paper focuses on the task of *reaction reference resolution*. Our contributions are three-fold:

- We define the new task of reaction reference resolution, where, given a reaction description, we aim to identify other reaction descriptions it refers to.
- In the absence of a large-scale gold-standard dataset, we leverage Reaxys®, a large commercial chemical reaction database, to create a silver-standard dataset, of a size that supports the development of not only traditional rule-based methods but also neural network models.
- We propose several baseline methods for the task and evaluate them on our silver-standard dataset. The baselines include an extension of the state-of-the-art end-to-end neural co-reference resolution method of Lee et al. [20]. Experimental results show that, while it successfully detects reaction descriptions that refer to other textual spans, it struggles to relate them correctly.

2 TASK DESCRIPTION

In this section, we formally introduce the reaction reference resolution task. We assume a pipeline system where, given the *Description* section of a patent, text spans corresponding to individual reaction descriptions are extracted in advance (following the methods of Jessop et al. [10] or Yoshikawa et al. [32]). Then, in a second step, all relationships between reaction descriptions that are relevant are identified to fully locate the details of each reaction. The output of this task could then be passed to a targeted information extraction system to pull out individual reaction details.

An input document is represented by a tuple $d = (P, R)$, where P is a sequence of paragraphs $P = (p_1, \ldots, p_{|P|})$ and R is a sequence of all textual spans of reaction descriptions $R = (r_1, \ldots, r_{|R|})$, dubbed *reaction spans*. A reaction span r_i is a sequence of consecutive paragraphs, represented by the first and last indices of paragraphs of the span $r_i = (\text{start}(i), \text{end}(i))$. For each reaction span r_i, the task is to: (1) predict whether the reaction refers to one or more previous reaction descriptions; and if so, (2) detect the reaction span(s) r_j (for $j < i$) that r_i refers to. Hereinafter, we call r_j the *parent reaction* and r_i the *child reaction*.

Figure 1 shows typical examples of reaction references. A reference relation is often indicated by an example ID, as in example (a), or the compound label used in the previous reaction description, as in example (b). However, it can also be the case that there is no direct referential expression between reaction descriptions, as in example (c). Therefore, we formulate the task as reference resolution between reaction descriptions (paragraphs) rather than between a referring expression such as *Example 1* and its referent, as in traditional co-reference resolution tasks. Subtasks (1) and (2) are analogous to the mention detection and mention clustering subtasks of co-reference resolution.

3 METHODS

In this section, we introduce a baseline neural model that captures the context and detects reference relations between reaction descriptions. As illustrated in Figure 2, our model is based on the end-to-end neural co-reference model of Lee et al. [20]. The original model takes a contextualized representation for each entity mention and computes the likelihood that a pair of mentions is co-referential, whereas our model extends it to a hierarchical architecture that first encodes individual paragraphs and then feeds it to a document-level encoder, so that the model can capture references between longer text spans (i.e. multiple paragraphs).

3.1 Paragraph and Span Representation

Our model has a hierarchical architecture that first encodes each paragraph and then encodes the entire document to obtain a contextualized representation of each paragraph. We first encode each input paragraph to obtain a context-independent paragraph encoding. A paragraph p_t consists of a sequence of words $p_t = \{w_1^{(t)}, \ldots, w_{|p_t|}^{(t)}\}$. Each word is represented by:

$$x_{t,k} = \text{WE}(w_k^{(t)}) \oplus \text{ELMo}(p_t)_k$$
$$\oplus \text{CC}(w_k^{(t)}) \oplus \text{NER}(p_t)_k, \tag{1}$$

where $\text{WE}(w_k^{(t)}) \in \mathbb{R}^{d_{\text{WE}}}$ is a pretrained word embedding, $\text{ELMo}(p_t)_k \in \mathbb{R}^{d_{\text{ELMo}}}$ is a contextualized word embedding [25], $\text{CC}(w_k^{(t)}) \in \mathbb{R}^{d_{\text{CC}}}$ is a CNN-based character-level word encoding, and $\text{NER}(p_t)_k \in \mathbb{R}^{d_{\text{NER}}}$ is a trainable embedding of the named entity label (here, chemical compound type). The details of the embeddings and named entity labels are provided in Section 5. We then encode p_t into a vector using a bidirectional LSTM:

$$h_t = \overrightarrow{h_{t,|p_t|}} \oplus \overleftarrow{h_{t,1}}, \tag{2}$$

$$\overrightarrow{h_{t,k}} = \text{LSTM}(x_{t,k}, \overrightarrow{h_{t,k-1}}; \theta_{\text{PF}}), \tag{3}$$

$$\overleftarrow{h_{t,k}} = \text{LSTM}(x_{t,k}, \overleftarrow{h_{t,k+1}}; \theta_{\text{PB}}), \tag{4}$$

where θ_{PF} and θ_{PB} denote model parameters of the forward and the backward LSTMs, respectively.

To obtain the representation of each reaction span, we first obtain a contextualized representation of each paragraph using a document-level bidirectional LSTM:

$$h_t^* = \overrightarrow{h_t^*} \oplus \overleftarrow{h_t^*}, \tag{5}$$

$$\overrightarrow{h_t^*} = \text{LSTM}(h_t, \overrightarrow{h_{t-1}^*}; \theta_{\text{DF}}), \tag{6}$$

$$\overleftarrow{h_t^*} = \text{LSTM}(h_t, \overleftarrow{h_{t+1}^*}; \theta_{\text{DB}}), \tag{7}$$

where θ_{DF} and θ_{DB} denote model parameters of the forward and the backward LSTMs, respectively. These paragraph representations are combined to obtain each reaction span representation:

$$g_i = h_{\text{start}(i)}^* \oplus h_{\text{end}(i)}^*. \tag{8}$$

3.2 Training Objectives

Our goal is to identify the parent reaction span for each child reaction span. As described in Section 2, this can be further divided into two subtasks: predicting whether the given span is a child of some reaction span, and linking the child spans to their parent spans.

ID	Text
(a) Reference with the example ID	
RX1	**Example 26** 11-Dioxo-3,3-dibutyl-5-phenyl-7-methylthio-8-(N-{(R)-α-[N-((S)-1-carboxy-(R)-hydroxypropyl)carbamoyl]-4-... 1,1-Dioxo-3,3-dibutyl-5-phenyl-7-methylthio-8-[N-((R)-α-carboxy-4-hydroxybenzyl) carbamoylmethoxy]- 2,3,4,5-tetrahydro-1,2,5-benzothiadiazepine (Example 18; 100mg, 0.152mmol) was dissolved in ...
RX2	Example 27 1,1-Dioxo-3,3-dibutyl-5-phenyl-7-methylthio-8-(N-{(R)-α-[N-((S)-1-carboxy-2-, methylpropyl)carbamoyl]-4- **The title compound was synthesized by the procedure described in Example 26** starting from ...
(b) Reference with the compound label	
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, (3.82 g, 35% yield). ...
(c) Reference with no direct referential expression	
RX1	III: 5-fluoro-N2-(4-methyl-3-propionylaminosulfonylphenyl)-N4-[4-(prop-2-ynyloxy)phenyl]-2,4- pyrimidinediamine mono-sodium salt 5-Fluoro-N2-(4-methyl-3-propionylaminosulfonylphenyl)-N4-[4-(prop-2-ynyloxy)phenyl]-2,4-pyrimidinediamine, II, (0.125 g, 0.258 mmol) was suspended in ... **The following compounds were made in a similar fashion to those above.**
RX2	IV: 5-Fluoro-N2-[4-methyl-3-(N-propionylaminosulfonyl)phenyl]-N4-[4-(2-propynyloxy)phenyl]-2,4-pyrimidinediamine Potassium Salt ...
RX3	V: 5-Fluoro-N2-[4-methyl-3-(N-propionylaminosulfonyl)phenyl]-N4-[4-(2-propynyloxy)phenyl]-2,4-pyrimidinediamine Calcium Salt ...

Figure 1: Abbreviated examples of reaction references: (a) reference with the example ID [29], (b) reference with the compound label [26], (c) reference with no direct referential expression [22].
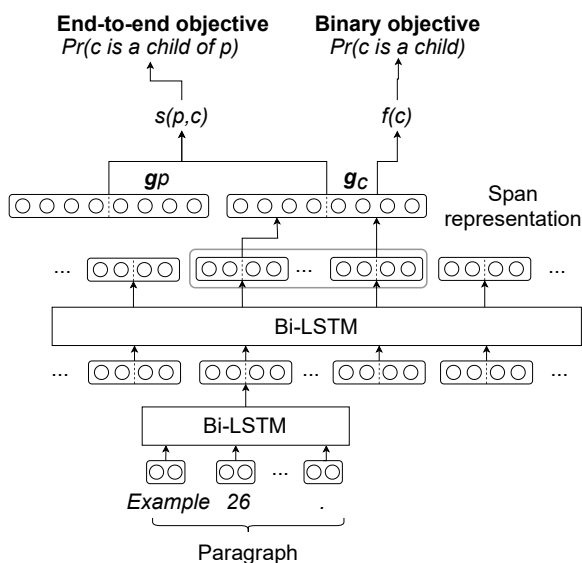

Figure 2: Illustration of our model architecture.

To evaluate the difficulty of these individual subtasks, we propose several model variants that differ in training objectives and decoding methods. Specifically, we train our model using either a binary classification objective that only models whether a reaction span is child or not, or an end-to-end objective that also models which span the child refers to. In decoding, our model uses either a heuristic method or the learned end-to-end model to identify the parent reaction for each child reaction span.

Binary classification objective. The notation $y_i = 0$ indicates the special case where the span r_i does not refer to any preceding spans. The binary classification objective models whether or not a reaction span is a child of at least one preceding reaction span:

$$\pi^{\text{NN}}(i) := \Pr(y_i \neq 0 \mid d) = \sigma(f(i)), \quad (9)$$

where σ denotes the sigmoid operator, and $f(i)$ is a span scoring function that computes the likelihood of the i-th span being a child, using a feed-forward network:

$$f(i) = \boldsymbol{w}_{\text{C-NN}}^{\top} \text{FFNN}(\boldsymbol{g}_i) + b_{\text{C-NN}}. \quad (10)$$

Given the training dataset $D = (d_1, \ldots, d_{|D|})$, the model is trained to minimize the following binary cross-entropy loss:

$$\mathcal{L}_{\text{BIN}} = - \sum_{d=(P,R) \in D} \sum_{i=1}^{|R|} \Big(\mathbb{1}_{y_i \neq 0} \log \pi^{\text{NN}}(i)$$
$$+ \mathbb{1}_{y_i = 0} \log(1 - \pi^{\text{NN}}(i)) \Big). \quad (11)$$

End-to-end objective. The end-to-end objective models not only the likelihood of a span being a child, but also which span the child span refers to. Analogous to Lee et al. [20], we aim to model the conditional probability distribution for document d as a product of multinomials for individual spans:

$$\Pr(y_1, \ldots, y_{|R|} \mid d) = \prod_{i=1}^{|R|} \Pr(y_i \mid d). \quad (12)$$

The probability of each preceding span being the parent of the given span is computed as follows:

$$\pi^{\text{NN}}(j, i) := \Pr(y_i = j \mid d)$$
$$= \exp(s(j, i)) / \sum_{j'=0}^{i-1} \exp(s(j', i)), \quad (13)$$

where $s(j, i)$ denotes the score of the likelihood that span r_i refers to preceding span r_j. The score is computed using the pair of corresponding span representations:

$$s(j, i) = \boldsymbol{w}_{\text{P-NN}}^{\top} \phi_{j,i} + b_{\text{P-NN}}, \quad (14)$$
$$\phi_{j,i} = \text{FFNN}(\boldsymbol{g}_j \oplus \boldsymbol{g}_i \oplus (\boldsymbol{g}_j \circ \boldsymbol{g}_i) \oplus \text{Dist}(j, i)) \quad (15)$$

for $j > 0$, where $\circ$ denotes element-wise multiplication. $\text{Dist}(j, i)$ denotes the distance embedding, whose definition follows that of Lee et al. [20]. We define $s(0, i) = 0$ for any i, corresponding to

the case where r_i has no reference. As we assume that the reaction spans are provided as part of the input, our model does not require span likelihood scores as used in the original co-reference resolution formulation. In another departure from the co-reference resolution setting, in our case the number of span pairs to be considered is tractable,[2] meaning we do not need to perform candidate pruning.

Given the training dataset $D = (d_1, \ldots, d_{|D|})$, the model is trained to minimize the negative log-likelihood loss against the true parent spans, where Y_i^* denotes the set of all correct parent span indices:

$$\mathcal{L}_{E2E} = - \sum_{d=(P,R) \in D} \sum_{i=1}^{|R|} \log \sum_{j \in Y_i^*} \Pr(y_i = j \mid d). \quad (16)$$

Note that our model is trained to predict only one parent for each child, although a child can have multiple parent spans. We observed that more than 90% of the child reaction spans in our dataset have only a single parent, and this simplification has very limited effect on performance. We can easily extend our model to allow multiple parent spans by modeling the link probability for each candidate parent span independently. Specifically, we can replace the softmax operation in Equation (13) with the sigmoid activation, and the loss function Equation (16) with the binary cross-entropy loss.

Decoding. To decode with the model trained with the end-to-end objective, we use the maximum index (including 0, i.e. no reference) as the predicted parent index of a reaction span; we refer to this method as "E2E". For the model trained with the binary classification objective (Equation (11)), we use a simple heuristic to predict the parent reaction spans, called "IMMPREV": reaction spans that are classified as a child reaction span are linked to the immediately-preceding previous span that is not classified as a child span.

4 SILVER-STANDARD DATASET

To the best of our knowledge, there is no publicly available dataset for reaction reference. Manually annotating a dataset of chemical reaction reference requires expert knowledge of chemistry and involves document-wise annotation of patents, with a single document often containing hundreds of paragraphs. As a more viable alternative, we create a silver-standard dataset from Reaxys®, a large-scale commercial chemical reaction database which contains a large quantity of reaction information associated at the document level with patents. While our dataset relies on a proprietary database, we believe that our methodology can generalize to similar patent silver standards generated using comparable reaction databases.

Reaxys® contains information of chemical reactions associated with their location information, i.e., the patents and the paragraphs where their reaction processes are described. To avoid exhaustive manual extraction of the full reaction details for all compounds described in a patent, construction process of Reaxys® database is streamlined through reuse of the conditions of a reaction for similar reactions described in the same patent. For this reason, the database includes internal reference links between similar reactions that share reaction conditions. As such, the mapping process from the database to our silver-standard dataset is straightforward: we simply map each reaction into one or more paragraph sequences (reaction spans) using

# Documents	143
# Paragraphs	39,437
# Reaction spans	3,072
# Child reaction spans	549
# Tokens / Paragraph	73.7

Table 1: Evaluation dataset statistics. Tokenization is based on OSCAR4 [11].

the location information and then use the reference links to label corresponding pairs of reaction spans. There are some limitations due to the fact that the database is not originally intended to be used for document processing. For example, the link information is not always complete, especially when multiple references are involved.[3] It is for this reason that we consider the annotations a silver-standard, with moderate recall and high precision.

We follow the work of Yoshikawa et al. [32] for patent document selection and obtaining reaction span information. We use the same set of patents, and extract the text from the description section of each patent. We split the dataset into five partitions, and perform five-fold cross validation, using three partitions for training, one for validation, and one for testing in each fold. We report the average performance across the five folds. An example and the statistics of the dataset are presented in Figure 3 and Table 1, respectively. We also sampled a small portion of our dataset and classified the reaction references based on the three types described in Figure 1. Among the sampled reaction references, 72% are references with the example ID, 20% are references with the compound label, and 8% are references with no direct referential expression.

5 EXPERIMENTAL DETAILS

5.1 Baseline Methods

In addition to the neural model described in Section 3, we introduce additional baselines, including heuristics and a traditional machine learning method using bag-of-words features.

Pattern matching baseline. We observe that there are high-precision patterns in child reactions that indicate their parent reaction spans. Namely, the parent reaction is often mentioned by its "example ID" such as *Example 1*, *Preparation 1*, or *Step 1*. An example is given in Figure 1(a). Based on this observation, we develop a rule-based baseline using regular expressions to detect such example ID patterns. For each reaction span that contains such a pattern, we search the preceding spans in reverse chronological order until we find a span that contains the example label in its heading or at the beginning of the text body.

Immediate previous baseline. Another naïve baseline is to link all reaction spans except the first, to the immediately previous span, i.e. $y_t = t - 1$. We label this method "NAIVE-IMMPREV". We also include an oracle variant, "ORACLE-IMMPREV", which uses the true labels for the subtask of child span detection (i.e. the oracle provides a perfect decision as to whether a reaction span is a child of other reaction spans or not), and performs parent–child linking by linking

---

[2]$O(|R|^2)$, where $|R| < 50$ for more than 90% of the dataset.

[3]Reactions that refer to multiple reaction steps usually have a link to only the final step of the reaction sequence.

par_id	RX	parent_id	text
117	9	-1	General procedure for the synthesis of aryl N-(guanidino)imines N'-arylhydrazones 7Aa-Ag and 8Aa. ...
118	-1		Characterization of compounds 7Aa-Ag and 8Aa:
119 120	10	-1	E)-2-((4'-((E and Z)-(2-(2-Chlorophenyl)hydrazono)methyl)-[1,1'-biphenyl]-4-yl)methylene)hydrazine- ... Yield: 70%; mp 328-330°C. 1H NMR, (400 MHz, DMSO-d6) δ 6.81 (dt, J=8.4 and 1.6 Hz, 1H, H-4"), ...
121 122	11	10	(E)-2-((4'-((E and Z)-(2-(2-Bromophenyl)hydrazono)methyl)-[1,1'-biphenyl]-4-yl)methylene)hydrazine- ... Yield: 64%; mp >380°C. 1H NMR (400 MHz, DMSO-d6) δ 6.75 (dt, J=8.4 and 1.6 Hz, 1H, H-4"), ...
123 124	12	10	(E)-2-((4'-((E)-(2-(4-Fluorophenyl)hydrazono)methyl)-[1,1'-biphenyl]-4-yl)methylene)hydrazine- ... Yield: 88%; mp >380°C. 1H NMR (400 MHz, DMSO-d6) δ 7.07 (d, J=7.6 Hz, 2H, H-2"), ...

Figure 3: An example of our silver-standard dataset. Only a fraction of the patent document is shown due to the space limitation. The columns represent paragraph ID (`par_id`), reaction ID (`RX`), parent reaction ID (`parent_id`), and paragraph text (`text`). "-1" in the reaction ID means that the paragraph is not a part of a reaction description. "-1" in the parent reaction ID means that the reaction span does not have a parent reaction span.

each given child span to its immediately preceding non-child span, similarly to the IMMPREV decoding method from Section 3.2. By assuming the annotations for child spans as given, we can measure the complexity of the parent–child linking task separately from the child span detection task.

Bag-of-words baseline. We also implement a simple bag-of-words classifier. For each reaction span r_i, we construct a vector x_i of vocabulary size $|V|$, where each element is the frequency of the corresponding term in the paragraph. We calculate the score of each span being a child as follows:

$$\pi^{\text{BOW}}(i) := \Pr(y_i \neq 0 \mid d)$$
$$= \sigma(w_{\text{C-BOW}}^{\top} x_i + b_{\text{C-BOW}}), \tag{17}$$

where σ denotes the sigmoid function. For each span r_i that is classified as a child, we calculate the score of every preceding span r_j being the parent of r_i as:

$$\pi^{\text{BOW}}(j, i) := \Pr(y_i = j \mid y_i \neq 0, d) \tag{18}$$
$$= \exp(s^{\text{BOW}}(j, i)) / \sum_{j'=1}^{i-1} \exp(s^{\text{BOW}}(j', i)),$$
$$s^{\text{BOW}}(j, i) = w_{\text{P-BOW}}^{\top} (x_i \oplus x_j \oplus x_{j,i}) + b_{\text{P-BOW}}, \tag{19}$$

where $x_{j,i}$ is a one-hot vector of the words appearing in both r_j and r_i. The model is trained to optimize the following joint loss function, with an L2 regularization term:

$$\mathcal{L}_{\text{BOW}} = \sum_{(P,R) \in D} (L_{\text{bin}}(P, R) + L_{\text{link}}(P, R)), \tag{20}$$

$$L_{\text{bin}}(P, R) = - \sum_{i=1}^{|R|} \Big(\mathbb{1}_{y_i \neq 0} \pi^{\text{BOW}}(i)$$
$$+ \mathbb{1}_{y_i = 0} (1 - \pi^{\text{BOW}}(i)) \Big), \tag{21}$$

$$L_{\text{link}}(P, R) = - \sum_{i : y_i \neq 0} \log \sum_{j \in Y_i^*} \pi^{\text{BOW}}(j, i). \tag{22}$$

5.2 Implementation Details and Optimization

Each paragraph is tokenized using the OSCAR4 tokenizer [11], which was developed for the chemical domain. We use embeddings from word2vec skip-gram [23] ($d_{\text{WE}} = 200$) and ELMo [25] ($d_{\text{ELMo}} = 1024$) for the WE and ELMo embeddings in Equation (1), respectively, both trained on chemical patent documents [33]. For the character-level word encodings, we apply a character CNN with

Method	$\mathcal{P}$	$\mathcal{R}$	$\mathcal{F}_1$
NAIVE-IMMPREV	.189	1.00	.315
Pattern matching	.876	.545	.660
Bag-of-words	.882	.617	.688
RREF$_{\text{BINARY}}$	.926	.866	.893
RREF$_{\text{E2E}}$	.897	.791	.839

Table 2: Results of the baseline methods on child reaction span detection in terms of precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F-score ($\mathcal{F}_1$). The scores are the average of five-fold cross validation. "RREF$_{\text{BINARY}}$" and "RREF$_{\text{E2E}}$" indicate our neural model with the binary classification objective and the end-to-end objective, respectively.

30 trigram filters to generate 25d character embeddings. For the NER features, we use a named entity recognizer based on the Reaxys® Gold chemical tag set [33], and assign 10d trainable embeddings to the token-level labels. We use 5d trainable embeddings for the distance embeddings in $s(j, i)$. In the FFNN layer in $f(i)$ and $s(j, i)$, we use a 2-layer feed-forward network with 150d hidden layers and a ReLU activation. We optimized the number of LSTM layers and hidden state dimensionalities for the two bidirectional models via grid search on the validation set.

The models are trained using the Adam optimizer [14], with dropout and gradient clipping ≤ 5.0. The dropout rate, applied evenly across the model, is selected from $\{0.3, 0.5\}$ based on the validation data. The minibatch size is 1, and the model is trained for up to 150 epochs, with early stopping.

6 RESULTS AND DISCUSSION

6.1 Main Results

Child span detection. Table 2 shows the results of the baseline methods on child span detection, which is the performance in terms of binary classification predicting whether each reaction span is a child of another reaction span or not. The neural methods obtain high F-scores over 0.8, whereas the bag-of-words classifier shows only a small gain over the rule-based baseline. This suggests that this

Method	$\mathcal{P}$	$\mathcal{R}$ ($\mathcal{R}^{\mathrm{m}}$)	$\mathcal{F}_1$ ($\mathcal{F}_1^{\mathrm{m}}$)
Naive-ImmPrev	.026	.145 (.133)	.044±.011 (.043)
Pattern matching	.332	.207 (.188)	.252±.096 (.237)
Bag-of-words	.302	.236 (.217)	.258±.173 (.244)
RREF$_{\text{Binary-ImmPrev}}$	.595	.560 (.506)	.576±.130 (.545)
RREF$_{\text{E2E}}$	.390	.347 (.316)	.367±.155 (.349)
Oracle-ImmPrev	.706	.706 (.643)	.706±.088 (.672)

Table 3: End-to-end performance of the baseline methods in terms of precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F-score ($\mathcal{F}_1$). $\mathcal{R}^{\mathrm{m}}$ and $\mathcal{F}_1^{\mathrm{m}}$ are calculated taking multiple references into account. The scores are the average performance of five-fold cross validation. The standard deviations are also displayed for $\mathcal{F}_1$. "RREF$_{\text{Binary-ImmPrev}}$" model is trained on the binary classification only and performs ImmPrev decoding. "RREF$_{\text{E2E}}$" model is trained with the end-to-end objective. Note that Oracle-ImmPrev uses the true labels for detecting the child spans and thus the evaluation is not end-to-end in a strict sense.

subtask is ML-feasible, noting that rich document representations are needed to achieve reasonable performance.

End-to-end performance. Table 3 shows the performance on the end-to-end task. As our methods assume a single parent for each child span, we evaluate the performance with two different measures. As the primary measure, we calculate the precision ($\mathcal{P}$), recall ($\mathcal{R}$), and F-scores ($\mathcal{F}_1$) by regarding the prediction as correct if the model predicts one of the true parents of the target reaction span. We also report the recall ($\mathcal{R}^{\mathrm{m}}$) and F-score ($\mathcal{F}_1^{\mathrm{m}}$) based on the multiple reference scenario, in which, if the target reaction span has multiple true parents, the recall is calculated based on the total number of parents in the denominator.[4]

For the end-to-end task, RREF$_{\text{Binary-ImmPrev}}$ performs better than RREF$_{\text{E2E}}$ model, even though it is trained only with the binary classification objective and not trained on parent–child linking. The RREF$_{\text{E2E}}$ model struggles to detect parent–child links effectively, and achieves lower performance than the simple ImmPrev decoding. Considering the fact that the RREF$_{\text{E2E}}$ model performs comparably with RREF$_{\text{Binary}}$ at child span detection, we hypothesize that even if the model is trained with the end-to-end objective, it primarily learns to detect child reaction spans from non-child ones, largely ignoring relations between other reaction spans.

These results reveal that it is the parent–child linking step rather than the child detection step that makes reaction reference resolution difficult. ImmPrev decoding is a strong baseline (0.706 F-score given oracle child reaction spans), but there is significant room for improvement. Despite its success in the traditional co-reference resolution task, the RREF$_{\text{E2E}}$ method struggles to learn the pairwise relations between reaction spans. One of the main differences between the traditional co-reference resolution task and the reaction reference resolution task is text length: the target spans are paragraph sequences in a long document rather than short word

---

[4] $\mathcal{P}$ = #(correct refs) / #(predicted refs),
$\mathcal{R}$ = #(correct refs) / #(spans with at least one true refs),
$\mathcal{F}_1$ = $2\mathcal{P}\mathcal{R}$ / ($\mathcal{P}$ + $\mathcal{R}$),
$\mathcal{R}^{\mathrm{m}}$ = #(correct refs) / #(all true refs),
$\mathcal{F}_1^{\mathrm{m}}$ = $\mathcal{F}_1$ = $2\mathcal{P}\mathcal{R}^{\mathrm{m}}$ / ($\mathcal{P}$ + $\mathcal{R}^{\mathrm{m}}$)

sequences within a few sentences. A model that can more effectively learn to capture relations between long text spans is left for future work.

6.2 Error Analysis

Examples of the system output are shown in Figure 4, which corresponds to the typical reference cases introduced in Figure 1, namely (a) reference with the example ID, (b) reference with the compound label, and (c) reference with no direct referential expression. Obviously, the pattern matching baseline fails when the target span does not mention the example label of its parent span. While the performance of Oracle-ImmPrev and RREF$_{\text{E2E}}$ is less tied to specific reference cases, they can fail in many "easy" cases where there is an explicit mention of the parent example ID, indicating that it is difficult to learn such direct references without explicit training signal that forces the model to attend to particular patterns. The example in Figure 4 (c) is a special case that involves *general conditions*: the first paragraphs describe the general procedure to synthesize multiple similar compounds, and the examples of specific compound names are listed in subsequent paragraphs. We can see that RREF$_{\text{E2E}}$ fails to detect references to general conditions that Oracle-ImmPrev detects successfully. In some cases, RREF$_{\text{E2E}}$ even fails to recognize them as a child reaction, which might be because reference to general conditions is likely to be implicit, i.e. the child description does not explicitly mention the example label of its parent description.

7 RELATED WORK

Krallinger et al. [17] provide an extensive survey of information retrieval and text mining for chemical literature, including extraction of chemical reaction information from patents. Several systems have been proposed that specifically target text mining from chemical patents [9, 19, 28]. Previous work has mainly focused on extracting chemical concepts and relations from individual paragraphs. In the ChEMU 2020 evaluation lab [8], state-of-the-art methods for chemical named entity recognition and event extraction were evaluated on patent documents. Jessop et al. [10] propose a broader-scope, multi-stage system called *PatentEye*, which first detects experimental sections in the description part of a patent, and then passes individual sections into downstream modules that extract reaction details (e.g. title compound, reagents, and analytical data) from the input text. Although they mention the presence of references between reaction descriptions, they do not present specific ways to resolve them.

To our knowledge, the first study to take into account references in reaction descriptions was by Ai et al. [1]. Their method automatically extracts chemical reactions from the experimental section of papers in chemistry journals, representing a reaction as a synthesis frame containing arguments such as *Product, Yield, Role*, and *Substance*. They employ template matching against parsed sentences to detect phrases related to predefined reaction events. To facilitate copying of relevant information between synthesis frames, they define two types of inter-paragraph references, namely general procedures and analogous syntheses. The paragraphs are linked based on certain referring expressions such as example IDs and compound labels. Overall, their system solely relies on handcrafted patterns and rules, and thus the cases covered by the system are limited. Although their

(a) Example ID

RX	Text
23	Example 24 ...
	...
25	Example 26 ...
26	Example 27 ... **The title compound was synthesized by the procedure described in Example 26** starting from ...

True	$26 \rightarrow 25$
Pattern matching	$26 \rightarrow 25$
ORACLE-IMMPREV	$26 \rightarrow 25$
RREF$_{E2E}$	$26 \rightarrow 23$

(b) Compound label

RX	Text
18	A mixture of the obtained ester, ... Column chromatography of the residue (silica gel-hexane/ethyl acetate, 9:1) gave **Compound B11**, ...
	...
20	Using 2-ethoxyethanol and **following the procedure for Compound B11** gave Compound B13, ...

True	$20 \rightarrow 18$
Pattern matching	no reference
ORACLE-IMMPREV	$20 \rightarrow 18$
RREF$_{E2E}$	$20 \rightarrow 18$

(c) No direct referential expression

RX	Text
9	**General procedure** for the synthesis of aryl N'-(guanidino)imines N'-arylhydrazones 7Aa-Ag and 8Aa. ...
	Characterization of compounds 7Aa-Ag and 8Aa:
10	(E)-2-((4'-((E and Z)-(2-(2-Chlorophenyl) hydrazono) methyl)-... (7Aa) Yield: 70%; mp 328-330° C. 1H NMR, ...
11	(E)-2-((4'-((E and Z)-(2-(2-Bromophenyl) hydrazono) methyl)-... (7Ab) Yield: 64%; mp > 380° C. 1H NMR ...

True	$11 \rightarrow 10$
Pattern matching	no reference
ORACLE-IMMPREV	$11 \rightarrow 10$
RREF$_{E2E}$	$11 \rightarrow 7$

Figure 4: Abbreviated system output examples. Only parts of the patent documents are shown due to the space limitation. "RX" is the reaction ID, and "True" is the label in the silver set. "X → Y" indicates that the child reaction X refers to the parent reaction Y.

system is reported to produce results in the range of 80–90% for simple synthesis paragraphs, and 60–70% for complex paragraphs, details of their data set and evaluation procedures are not provided. It is unclear whether these results would generalise to larger data sets with more variability, and we cannot directly compare their approach with ours.

Finally, we compare our task formulation with text alignment task. Text alignment task [34] is arguably one of the most relevant tasks to reaction reference resolution, as it requires understanding of long documents and it has an application to citation recommendation [4, 12]. A key difference between text alignment task and our reaction reference task is that text alignment task mainly focuses on alignment of text portions in different documents, whereas our reaction reference task targets to find inner-document references between reaction descriptions. Our task requires not only understanding of the content of each reaction description based on its contextual information but also considering the document structure such as relative position of reaction description pairs, as exemplified in Figure 1 (c). A child reaction often omits a large part of the reaction description (sometimes it is just a name of the target compound), unlike typical inter-document references where the referer shares some common context with the source document it refers to. Applying core ideas from existing methods for text alignment task to reaction reference task would be an interesting direction for future work.

8 CONCLUSION

We have introduced the chemical reaction reference resolution problem, a largely unexplored yet critical step in information extraction of complete reaction details presented in chemical patents. We provided a formal specification of this task, built an evaluation corpus, and adapted a neural co-reference architecture to this task, as well as introducing several other baseline methods. Our neural method achieved promising results for detecting (child) reaction descriptions containing references, but identifying the precise (parent) reference of child spans is challenging for the model. Key issues for future work are more effective learning of reference patterns in reaction descriptions, and better handling of long text spans, as well as long-distance relations.

Some aspects of reaction reference are not well explored in this paper. For example, we did not observe many examples involving multiple references and references to general conditions in our silver-standard dataset. This is partly because the database used to construct our corpus is not designed for this use. Very recently, ChEMU [7], a shared task of chemical information extraction was held and the reaction reference resolution task was included as one of the key tasks. As a part of the shared task, a dataset with gold annotation of reaction reference was made publicly available. [5] We will evaluate the performance of our baselines on the gold-standard dataset. In addition, although we have cast the problem as a text processing task, a more complete description of the reaction would involve the combination of the textual body, figures, and tables in the patent. Thus, a multi-modal version of the reaction reference task is a possible direction to explore in the future. Another interesting extension of this taks would be considering cross-document references of chemical reactions.

ACKNOWLEDGMENTS

REFERENCES

[1] C. S. Ai, Paul E. Blower, and R. H. Ledwith. 1990. Extraction of Chemical Reaction Information from Primary Journal Text. *Journal of Chemical Information and Computer Sciences* 30, 2 (1990), 163–169. https://doi.org/10.1021/ci00066a012 arXiv:https://pubs.acs.org/doi/pdf/10.1021/ci00066a012

[2] Saber A. Akhondi, Hinnerk Rey, Markus Schwörer, Michael Maier, John Toomey, Heike Nau, Gabriele Ilchmann, Mark Sheehan, Matthias Irmer, Claudia Bobach, Marius Doornenbal, Michelle Gregory, and Jan A. Kors. 2019. Automatic Identification of Relevant Chemical Compounds from Patents. *Database* 2019 (2019). https://doi.org/10.1093/database/baz001 arXiv:https://academic.oup.com/database/article-pdf/doi/10.1093/database/baz001/27636778/baz001.pdf

[5] http://chemu2021.eng.unimelb.edu.au/

[3] Cecilia Arighi, Lynette Hirschman, Thomas Lemberger, Samuel Bayer, Robin Liechti, Donald Comeau, and Cathy Wu. 2017. Bio-ID Track Overview. In *Proceedings of the BioCreative VI Workshop*. 14–19.

[4] Chandra Bhagavatula, Sergey Feldman, Russell Power, and Waleed Ammar. 2018. Content-Based Citation Recommendation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 238–251. https://doi.org/10.18653/v1/N18-1022

[5] James E. Blake and Robert C. Dana. 1990. CASREACT: More than a Million Reactions. *Journal of Chemical Information and Modeling* 30, 4 (1990), 394–399. https://doi.org/10.1021/ci00068a008

[6] Peter Corbett, Colin Batchelor, and Simone Teufel. 2007. Annotation of Chemical Named Entities. In *Biological, Translational, and Clinical Language Processing*. Association for Computational Linguistics, 57–64.

[7] Jiayuan He, Biaoyan Fang, Hiyori Yoshikawa, Yuan Li, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Zubair Afzal, Zenan Zhai, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU 2021: Reaction Reference Resolution and Anaphora Resolution in Chemical Patents. In *Advances in Information Retrieval*, Djoerd Hiemstra, Marie-Francine Moens, Josiane Mothe, Raffaele Perego, Martin Potthast, and Fabrizio Sebastiani (Eds.). Springer International Publishing, 608–615.

[8] Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Jingqi Wang, Yuankai Ren, Zhi Zhang, Yaoyun Zhang, Mai Hoang Dao, Pedro Ruas, Andre Lamurias, Francisco M. Couto, Jenny Copara, Nona Naderi, Julien Knafou, Patrick Ruch, Douglas Teodoro, Daniel Lowe, John Mayfield, Abdullatif Köksal, Hilal Dönmez, Elif Özkirimli, Arzucan Özgür, Darshini Mahendran, Gabrielle Gurdin, Nastassja Lewinski, Christina Tang, Bridget T. McInness, C.S. Malarkodi, Pattabhi Rk Rao, Sobha Lalitha Devi, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2020. An Extended Overview of the CLEF 2020 ChEMU Lab : Information Extraction of Chemical Reactions from Patents. *Proceedings of the CLEF 2020 conference* (2020).

[9] Matthias Irmer, Lutz Weber, Timo Böhme, Anett Püschel, Claudia Bobach, and Ulf Laube. 2015. OCMiner for Patents. Extracting Chemical Information from Patent Texts. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*. 119–123.

[10] David M. Jessop, Sam E. Adams, and Peter Murray-Rust. 2011. Mining Chemical Information from Open Patents. *Journal of Cheminformatics* 3, 1 (2011), 40. https://doi.org/10.1186/1758-2946-3-40

[11] David M. Jessop, Sam E. Adams, Egon L. Willighagen, Lezan Hawizy, and Peter Murray-Rust. 2011. OSCAR4: A Flexible Architecture for Chemical Text-Mining. *Journal of Cheminformatics* 3, 1 (2011), 41. https://doi.org/10.1186/1758-2946-3-41

[12] Jyun-Yu Jiang, Mingyang Zhang, Cheng Li, Michael Bendersky, Nadav Golbandi, and Marc Najork. 2019. Semantic Text Matching for Long-Form Documents. In *The World Wide Web Conference (WWW '19)*. Association for Computing Machinery, 795–806. https://doi.org/10.1145/3308558.3313707

[13] J.-D. Kim, T. Ohta, Y. Tateisi, and J. Tsujii. 2003. GENIA Corpus— a Semantically Annotated Corpus for Bio-Textmining. *Bioinformatics* 19, suppl_1 (2003), i180–i182. https://doi.org/10.1093/bioinformatics/btg1023

[14] Diederik P. Kingma and Jimmy Ba. 2015. Adam: A Method for Stochastic Optimization. In *3rd International Conference on Learning Representations, ICLR 2015, San Diego, CA, USA, May 7-9, 2015, Conference Track Proceedings*, Yoshua Bengio and Yann LeCun (Eds.).

[15] Martin Krallinger, Obdulia Rabal, Saber A Akhondi, et al. 2017. Overview of the BioCreative VI Chemical-Protein Interaction Track. In *Proceedings of the Sixth BioCreative Challenge Evaluation Workshop*, Vol. 1. 141–146.

[16] Martin Krallinger, Obdulia Rabal, Florian Leitner, Miguel Vazquez, David Salgado, Zhiyong Lu, Robert Leaman, Yanan Lu, Donghong Ji, Daniel M. Lowe, Roger A. Sayle, Riza Theresa Batista-Navarro, Rafal Rak, Torsten Huber, Tim Rocktäschel, Sérgio Matos, David Campos, Buzhou Tang, Hua Xu, Tsendsuren Munkhdalai, Keun Ho Ryu, SV Ramanan, Senthil Nathan, Slavko Žitnik, Marko Bajec, Lutz Weber, Matthias Irmer, Saber A. Akhondi, Jan A. Kors, Shuo Xu, Xin An, Utpal Kumar Sikdar, Asif Ekbal, Masaharu Yoshioka, Thaer M. Dieb, Miji Choi, Karin Verspoor, Madian Khabsa, C. Lee Giles, Hongfang Liu, Komandur Elayavilli Ravikumar, Andre Lamurias, Francisco M. Couto, Hong-Jie Dai, Richard Tzong-Han Tsai, Caglar Ata, Tolga Can, Anabel Usié, Rui Alves, Isabel Segura-Bedmar, Paloma Martínez, Julen Oyarzabal, and Alfonso Valencia. 2015. The CHEMDNER Corpus of Chemicals and Drugs and Its Annotation Principles. *Journal of Cheminformatics* 7, 1 (2015), S2.

[17] Martin Krallinger, Obdulia Rabal, Anália Lourenço, Julen Oyarzabal, and Alfonso Valencia. 2017. Information Retrieval and Text Mining Technologies for Chemistry. *Chemical Reviews* 117, 12 (2017), 7673–7761. https://doi.org/10.1021/acs.chemrev.6b00851

[18] Lev Lafayette, Greg Sauter, Linh Vu, and Bernard Meade. 2017. Spartan HPC-Cloud Hybrid: Delivering Performance and Flexibility. https://doi.org/10.4225/49/58ead90dceaaa

[19] Alexander Johnston Lawson, Stefan Roller, Helmut Grotz, Janusz L Wisniewski, and Libuse Goebels. 2011. Method and Software for Extracting Chemical Data. German patent no. DE102005020083A1.

[20] Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-End Neural Coreference Resolution. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 188–197. https://doi.org/10.18653/v1/D17-1018

[21] Daniel M. Lowe. 2012. *Extraction of Chemical Structures and Reactions from the Literature*. Ph.D. Dissertation. University of Cambridge.

[22] Daniel Magilavy and Polly Pine. 2018. 2,4 Substituted Pyrimidinediamines for Use in Discoid Lupus. European patent no. EP2683385B1.

[23] Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed Representations of Words and Phrases and Their Compositionality. In *Advances in Neural Information Processing Systems 26*, C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger (Eds.). 3111–3119.

[24] Dat Quoc Nguyen, Zenan Zhai, Hiyori Yoshikawa, Biaoyan Fang, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Saber A Akhondi, Trevor Cohn, Timothy Baldwin, et al. 2020. ChEMU: Named Entity Recognition and Event Extraction of Chemical Reactions from Patents. In *European Conference on Information Retrieval*. Springer, 572–579.

[25] Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep Contextualized Word Representations. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*. Association for Computational Linguistics, 2227–2237. https://doi.org/10.18653/v1/N18-1202

[26] Stanislaw Rachwal and Yufen Hu. 2018. Highly Photo-Stable Bis-Triazole Fluorophores. US patent no. US20180002337A1.

[27] Pedro Ruas, Andre Lamurias, and Francisco M. Couto. 2020. Linking Chemical and Disease Entities to Ontologies by Integrating PageRank with Extracted Relations from Literature. *Journal of Cheminformatics* 12, 1 (2020), 57. https://doi.org/10.1186/s13321-020-00461-4

[28] Nadine Schneider, Daniel M Lowe, Roger A Sayle, Michael A Tarselli, and Gregory A Landrum. 2016. Big Data from Pharmaceutical Patents: A Computational Analysis of Medicinal Chemists' Bread and Butter. *Journal of medicinal chemistry* 59, 9 (2016), 4385–4402.

[29] Ingemar Starke, Mikael Ulf Johan Dahlstrom, David Blomberg, Suzanne Alenfalk, Tore Skjaret, and Malin Lemurell. 2017. Benzothiazepine and Benzothiadiazepine Derivatives with Ileal Bile Acid Transport (Ibat) Inhibitory Activity for the Treatment Hyperlipidaemia. European patent no. EP1427423B9.

[30] Jorge A. Vanegas, Sérgio Matos, Fabio González, and José L. Oliveira. 2015. An Overview of Biomolecular Event Extraction from Scientific Documents. *Computational and Mathematical Methods in Medicine* 2015 (2015), 571381. https://doi.org/10.1155/2015/571381

[31] Chih-Hsuan Wei, Yifan Peng, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Jiao Li, Thomas C Wiegers, and Zhiyong Lu. 2015. Overview of the BioCreative V Chemical Disease Relation (CDR) Task. In *Proceedings of the Fifth BioCreative Challenge Evaluation Workshop*, Vol. 14.

[32] Hiyori Yoshikawa, Dat Quoc Nguyen, Zenan Zhai, Christian Druckenbrodt, Camilo Thorne, Saber A. Akhondi, Timothy Baldwin, and Karin Verspoor. 2019. Detecting Chemical Reactions in Patents. In *Proceedings of the The 17th Annual Workshop of the Australasian Language Technology Association*. Australasian Language Technology Association, 100–110.

[33] Zenan Zhai, Dat Quoc Nguyen, Saber Akhondi, Camilo Thorne, Christian Druckenbrodt, Trevor Cohn, Michelle Gregory, and Karin Verspoor. 2019. 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, 328–338.

[34] Xuhui Zhou, Nikolaos Pappas, and Noah A. Smith. 2020. Multilevel Text Alignment with Cross-Document Attention. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, 5012–5025. https://doi.org/10.18653/v1/2020.emnlp-main.407