# ChEMU-Ref: A Corpus for Modeling Anaphora Resolution in the Chemical Domain

**Biaoyan Fang**[1], **Christian Druckenbrodt**[2], **Saber A. Akhondi**[2], **Jiayuan He**[1,3], **Timothy Baldwin**[1] **and Karin Verspoor**[1]
[1]The University of Melbourne, Australia
[2]Elsevier
[3]RMIT University, Australia
biaoyanf@student.unimelb.edu.au
{c.druckenbrodt,s.akhondi}@elsevier.com
{estrid.he,tbaldwin,karin.verspoor}@unimelb.edu.au

## Abstract

Chemical patents contain rich coreference and bridging links, which are the target of this research. Specially, we introduce a novel annotation scheme, based on which we create the ChEMU-Ref dataset from reaction description snippets in English-language chemical patents. We propose a neural approach to anaphora resolution, which we show to achieve strong results, especially when jointly trained over coreference and bridging links.

## 1 Introduction

Chemical research has contributed greatly to human society and wellbeing, including new medicines and vaccines (Gwynne and Heabrer, 2015). Research is heavily reliant on knowledge of existing chemical processes and methods of chemical synthesis, which are documented in chemical research literature and chemical patents. Given the rapid growth of both publications and patents in chemistry, the need for automatic methods to extract semi-structured knowledge from chemical texts is becoming increasingly critical (Li et al., 2016; Akhondi et al., 2019).

Anaphora resolution is a key component of comprehensive information extraction (Rösiger, 2019; Poesio et al., 2016). In chemistry, different chemical compounds are mixed and reacted together in different ways to generate novel compounds, and to understand the precise chemical process often involves both resolving anaphoric references and understanding chemical changes/interactions a given entity is involved in. For example, as seen in Figure 1, while the final mention of *mixture* on line 3 and that on line 4 are both coreferent and chemically identical, in the case of *mixture* on line 2 and the first mention of *mixture* on line 3, the chemical composition is the same but a transformation has taken place via the *stir* and *cool* actions.

Our aim in this paper is to both identify anaphoric references in chemical patents, and determine the chemical relation between each linked pair of entities. We propose a domain-specific annotation framework based on five types of anaphora relations combining coreference and bridging. We then construct a dataset following this framework, annotated by chemical experts who achieve high inter-annotator agreement. We additionally extend existing anaphora resolution methods to model anaphora in chemical text, and compare both component-wise and joint models for anaphora resolution. This dataset will be released as part of the upcoming ChEMU 2021 shared task[1] (He et al., to appear).

Our contributions in this paper are as follows: (1) we propose a novel annotation scheme for anaphora resolution in chemical patents; (2) we develop a novel anaphora-resolution dataset based on chemical patents; and (3) we extend a general-purpose coreference resolution method, and achieve strong results via joint training over coreference and bridging with domain-specific fine-tuning.

## 2 Related Work

Anaphora occurs in two basic forms: coreference and bridging. Coreference occurs when different expressions in a text refer to the same entity in the real world (Ng, 2017; Clark and Manning, 2015), while bridging occurs between discrete entities that are linked via lexical semantic, frame-based, or encyclopedic relations (Asher and Lascarides, 1998; Hou et al., 2018).

Most existing anaphora datasets focus only on coreference, predominantly using generic relations (Pradhan et al., 2012; Ghaddar and Langlais, 2016a), but also using domain-specific relations for knowledge-rich corpora such as biomedical litera-

[1]http://chemu.eng.unimelb.edu.au/



*Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics*, pages 1362–1375
April 19 - 23, 2021. 

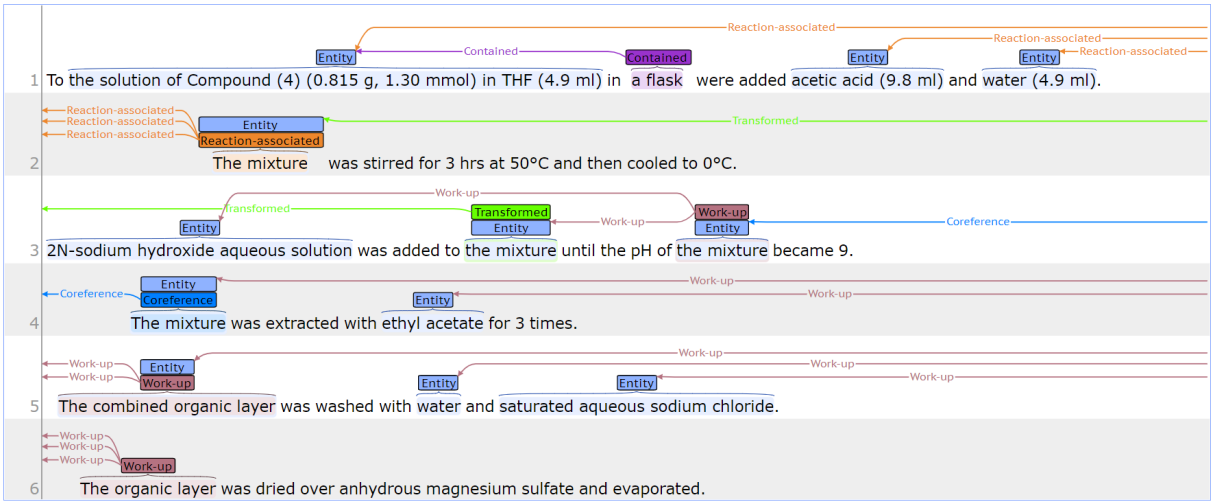


Figure 1: Annotated snippet of anaphora resolution in the chemical patents. Different color of links represent different anaphora relation types. Detailed anaphora relation definition can be seen Section 3.3.

ture (Nguyen et al., 2011; Cohen et al., 2017).

The CoNLL-2012 dataset (Pradhan et al., 2012) is a general corpus consisting of texts from three languages (English, Chinese, and Arabic). It is annotated based on OntoNotes v5.0 (Weischedel et al., 2013) and includes two types of coreference relations: IDENTITY, a symmetrical and transitive relation; and APPOSITIVE, two noun phrases that are adjacent and not linked by a copula. Coreference resolution is modelled as a clustering task. The wikicoref corpus was constructed with the same relations, over Wikipedia documents (Ghaddar and Langlais, 2016a).

BioNLP-ST 2011 (Nguyen et al., 2011) is a domain-specific coreference corpus over abstracts from biomedical literature, focusing mainly on gene–protein coreference, considering four relations: RELAT (relative pronouns or adjectives, e.g. *which*), PRON (pronouns, e.g. *they*), DNP (definite or demonstrative noun phrases marked with *the*, *this*, etc.), and APPOS (apposition). Instead of modelling coreference resolution as a clustering task, here the direction of coreference links is preserved. As this corpus focuses on gene–protein coreference, the range of coreference phenomena is limited. The CRAFT-CR corpus (Cohen et al., 2017) adds coreference relations to the Colorado Richly Annotated Full Text (CRAFT) corpus (Bada et al., 2012), following OntoNotes v5.0 with minor adaptions, and including discontinuous expressions, domain-specific proper nouns, and a broad range of mention types.

The definition of bridging is somewhat imprecise (Zeldes, 2017; Hou et al., 2018), and different corpora have adopted different definitions. Based on Rösiger et al. (2018), there are two types of bridging: *referential bridging*, which can be treated as a context-based relation; and *lexical bridging*, which describes lexical-semantic relations such as holonymy and meronymy. Poesio et al. (2008) introduced the ARRAU corpus of general language texts for bridging, which consists of news, dialogue, and narrative text. In the corpus, entities are limited to noun phrases, and most bridging pairs are lexical relations, with only a small number of instances of referential bridging. ISnotes (Hou et al., 2018) includes 50 Wall Street Journal (WSJ) articles from the OntoNotes corpus, and has both coreference and bridging annotations, with most of the bridging pairs being referential. BASHI (Rösiger, 2018a) has both coreference and bridging annotations over 50 WSJ articles based on the OntoNotes v5.0 guidelines, with most bridging links once again being referential. Rösiger (2016) developed a corpus called SciCorp based on English scientific papers, following the same annotation scheme as BASHI.

Due to limited dataset availability, most research has modelled coreference resolution and bridging separately. There are two basic approaches to coreference resolution. First is mention ranking methods, which aim to score the coreferent probability of mention pairs (Clark and Manning, 2015, 2016a,b; Wiseman et al., 2015, 2016), and make the assumption that mentions have been pre-identified, meaning they are heavily reliant on upstream mention detection methods. Second is span ranking methods, which combine mention detection with coreference prediction (Lee et al., 2017,

2018; Zhang et al., 2018; Grobol, 2019; Kantor and Globerson, 2019), and tend to perform better. Bridging methods can be grouped into: (1) rule-based methods (Hou et al., 2014; Rösiger, 2018b; Rösiger et al., 2018); and (2) machine learning methods (Hou, 2018a,b, 2020; Yu and Poesio, 2020). Rule-based methods have been shown to achieve competitive results on domain-specific corpora, but equally to be domain brittle. Yu and Poesio (2020) jointly trained a model for coreference resolution and bridging by adapting a span ranking method for coreference (Lee et al., 2018; Kantor and Globerson, 2019), achieving good performance over various bridging corpora. However, they evaluated their model only on bridging.

## 3 Annotation Scheme

In this section, we introduce our annotation guidelines for anaphora resolution in chemical patents. The complete annotation guidelines are made available at Fang et al. (2021).

### 3.1 Corpus Selection

We build on the ChEMU corpus (Verspoor et al., 2020) developed for the ChEMU 2020 shared task (He et al., 2020). This corpus 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.

### 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 (Pradhan et al., 2012; Cohen et al., 2017; Ghaddar and Langlais, 2016b), 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.

**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 don't 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.

**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.

**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:

- Pronouns: In chemical patents, pronouns (e.g. *they* or *it*) usually refer to a previously-mentioned chemical compounds.
- Definite NPs: Commonly used to refer to chemical compounds, e.g. *the solvent*, *the title compound*, *the mixture*.

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

- 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*.
- NPs with prepositions: Chemical NPs connected with prepositions (e.g. *in*, *with*, *of*) should be considered as a single mention. For example, *the appropriate amino derivative in dry THF* is a single mention.

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, mentions 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.

**Relationship to ChEMU 2020 entities:** Since this dataset is built on the ChEMU 2020 corpus (He et al., 2020), 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 co-refer 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 (Ng, 2017; Clark and Manning, 2015). In chemistry, identifying whether two mentions refer to the same entity needs to consider various 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:

- **Single Antecedents:** the anaphor refers to a single antecedent.
- **Multiple Antecedents:** the anaphor refers to multiple antecedents, e.g. in cases where multiple antecedents are combined to form a single *mixture*.

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

#### 3.3.2 Bridging

As stated in Section 3.3.1, 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 1, 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

| | Train | Dev | Test |
|---|---|---|---|
| Snippets | 148 | 27 | 45 |
| Sentences | 763 | 164 | 274 |
| Tokens/Sentences | 27.5 | 24.7 | 25.8 |
| Mentions | 2,284 | 430 | 736 |
| Dis. Mentions | 88 | 10 | 17 |
| Coref. | 421 | 88 | 124 |
| Bridging | 1,731 | 323 | 577 |
| TR | 85 | 17 | 29 |
| RA | 515 | 105 | 167 |
| WU | 1,063 | 172 | 364 |
| CT | 68 | 29 | 17 |

Table 1: Corpus annotation statistics. "Dis. Mentions" means discontinuous mentions. "Coref.", "TR", "RA", "WU", and "CT" denote COREFERENCE, TRANSFORMED, REACTION-ASSOCIATED, WORK-UP and CONTAINED, respectively. "Bridging" is the total across all bridging relations.

one-to-many relation, from the anaphor to the compounds (antecedents) that are used for the work-up process. As demonstrated in Figure 1, *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 some 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.

## 4 Task definition

Anaphora resolution can be decomposed into a two-step task: (1) mention detection; and (2) anaphora relation detection.

For the evaluation of mention and relation detection, we use precision, recall and F1. One issue here is that, for coreference resolution, anaphors can link to multiple antecedents. Many coreference evaluation metrics (Moosavi and Strube, 2016; Recasens and Hovy, 2011; Luo, 2005) cannot deal with this since they model coreference resolution as a clustering task, where all related antecedents and anaphors occur in one cluster, and assume a given mention occurs in a unique cluster. Hence we adopt the approach to evaluation of Kim et al. (2012), scoring coreference from two perspectives: (1) *surface coreference*; and (2) *atom coreference*. Surface coreference considers whether the anaphor refers to the closest previous antecedent(s). Atom coreference considers whether the anaphor refers to the correct antecedent(s). Atom coreference links take the coreferent transitivity into consideration and can be generated from surface coreference links, which we use by default.

For the corpus annotation, we use the BRAT text annotation tool.[2] To date, 220 snippets have been annotated by two chemical experts, a PhD candidate and a final year bachelor student in Chemistry. 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 compared their annotations; annotation guidelines were then refined based on discussion. After several rounds of training, we achieved a high inner-annotator agreement of Krippendorff's $\alpha = 0.92$ (Krippendorff, 2004) at the mention level,[3] and $\alpha = 0.84$ for relations. In total, 1,500 snippets will be annotated in the final dataset that will be used in the ChEMU 2021 shared task.

The statistics of the current corpus, and train/dev/test set splits that form the basis of our experiments in this paper, are shown in Table 1. The dev and test partitions were both double annotated by the two expert annotators, with any disagreements merged by an adjudicator.

## 5 Methodology

We propose a joint neural model for anaphora resolution.[4] Similar to Yu and Poesio (2020), our model adopts an end-to-end neural conference resolution (Lee et al., 2017, 2018), as outlined in Figure 2.

Assume the snippet has $T$ tokens represented as vector $X = \{x_1, ..., x_T\}$, consisting of fixed pretrained word and character embeddings learned from a convolution neural network (CNN).

For mention candidate detection, we follow the assumption of Lee et al. (2018), considering continuous tokens as a potential span and computing the span score ($s_m$) for each possible span. Specifically, span representation $s_i$ is obtained by the concatenation of output token representations ($x_i^*$) from a bidirectional LSTM (BiLSTM) (Hochreiter and Schmidhuber, 1997), the syntactic head representation ($h_i$) obtained from an attention mecha-

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

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

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

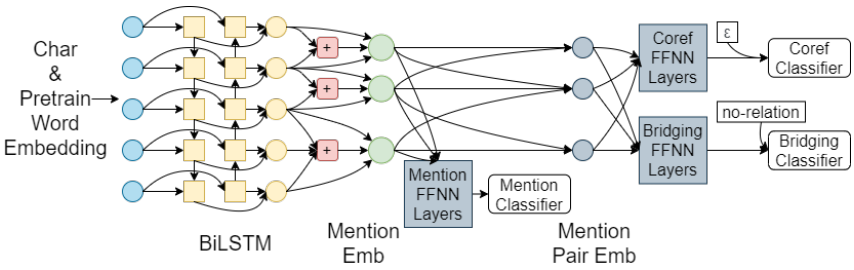


Figure 2: Joint training architecture.

nism (Bahdanau et al., 2015), and a feature vector of the mention ($\phi(i)$):

$$X^* = \text{BiLSTM}(X)$$
$$\alpha_t = w_\alpha \cdot \text{FFNN}_\alpha(x_t^*)$$
$$a_{i,t} = \frac{\exp(\alpha_t)}{\sum_{k=\text{START}(i)}^{\text{END}(i)} \exp(\alpha_k)}$$
$$h_i = \sum_{t=\text{START}(i)}^{\text{END}(i)} a_{i,t} \cdot x_t$$
$$s_i = [x^*_{\text{START}(i)}, x^*_{\text{END}(i)}, h_i, \phi(i)]$$

and the span score $s_m(i)$ is computed as:

$$s_m(i) = w_s \cdot \text{FFNN}_s(s_i)$$

where FFNN denotes a feed-forward neural network, and START(i) and END(i) represent the starting and ending token index for span $i$, respectively. To reduce the number of spans considered, we use a beam of $\lambda T$ candidate mention spans.

Inspired by Zhang et al. (2018), the mention loss is defined as:

$$L_{mention} = -\sum_{i=1}^{\lambda T} m_i * \log(\text{sigmoid}(s_m(i))) + (1 - m_i) * \log(1 - \text{sigmoid}(s_m(i)))$$

where:

$$m_i = \begin{cases} 0 & \text{span } i \notin \text{GOLD}_m \\ 1 & \text{span } i \in \text{GOLD}_m \end{cases}$$

$\text{GOLD}_m$ is the set of gold mentions that are involved in anaphora relations.

For anaphoric relation detection, a span pair embedding is obtained by the concatenation of each span embedding $(s_m(i), s_m(j))$ and the element-wise multiplication of the span embeddings $(s_m(i) \circ s_m(j))$ and a feature vector $(\phi(i,j))$ for span pair $i$ and $j$:

$$s_{i,j} = [s_m(i), s_m(j), s_m(i) \circ s_m(j), \phi(i,j)]$$

As coreference and bridging are different, we consider them separately.

For coreference resolution, we follow Lee et al. (2018) in optimizing the marginal log-likelihood of all correct antecedents for a given anaphor:

$$L_{coref} = \log \prod_{i=1}^{N} \sum_{\hat{y} \in Y(i) \bigcap \text{GOLD}_c(i)} P(\hat{y})$$

where $N$ is the number of candidate mentions; and $Y(i) = \{\epsilon, 1, ..., i-1\}$ is the set of possible assignments for each $y_i$, which $\epsilon$ represents a dummy antecedent and the numbers represent the proceeding spans. $\text{GOLD}_c(i)$ is the gold coreferent antecedents that span $i$ refers to. If span $i$ doesn't have a coreferent antecedent, $\text{GOLD}_c(i) = \epsilon$. $P(y_i)$ is obtained via softmax over the antecedent scores $s_c$ for the corresponding anaphor:

$$s_c(i,j) = \begin{cases} 0 & j = \epsilon \\ w_c \cdot \text{FFNN}_c(s_{i,j}) & j \neq \epsilon \end{cases}$$

For bridging resolution, as we have four relations, we model it as a multiclass classification task for each span pair. We represent the bridging relation as a one-hot representation and introduce a new relation type NO-RELATION for span pairs that do not have a bridging relation. The loss for bridging is:

$$y_b(i,j) = \text{softmax}(w_b \cdot \text{FFNN}_b(s_{i,j}))$$
$$L_{bridging} = -\sum_{c=1}^{K_c} \sum_{i=1}^{N} \sum_{j=1}^{i} b_{i,j,c} \log(y_b(i,j,c))$$

where $K_c$ represents the number of bridging categories, $y_b(i,j,c)$ denotes the prediction of $y_b(i,j)$ under category $c$, and:

$$b_{i,j,c} = \begin{cases} 0 & \text{span pair}(i,j) \notin \text{GOLD}_b(c) \\ 1 & \text{span pair}(i,j) \in \text{GOLD}_b(c) \end{cases}$$

where $\text{GOLD}_b(c)$ is the gold bridging relation under category $c$.

| Relation | Method | $P_A$ | $R_A$ | $F_A$ | $P_R$ | $R_R$ | $F_R$ |
|---|---|---|---|---|---|---|---|
| Coref. (Surface) | coreference | 84.9 | **50.0** | **62.9** | 73.7 | **41.9** | 53.4 |
| | joint_train | **89.4** | 45.8 | 60.5 | **81.7** | 40.6 | **54.2** |
| Coref. (Atom) | coreference | 84.9 | **50.0** | **62.9** | 75.6 | **42.6** | 54.4 |
| | joint_train | **89.4** | 45.8 | 60.5 | **82.3** | 40.8 | **54.5** |
| Bridging | bridging | 88.4 | 80.9 | 84.5 | 76.0 | 65.4 | 70.3 |
| | joint_train | **89.5** | **81.8** | **85.5** | **77.0** | **66.1** | **71.1** |
| TR | bridging | **77.5** | 63.8 | 69.7 | **76.2** | 63.8 | 69.1 |
| | joint_train | 76.9 | **69.0** | **72.7** | 75.9 | **69.0** | **72.3** |
| RA | bridging | 82.7 | 83.3 | 83.0 | 66.0 | 57.5 | 61.4 |
| | joint_train | **89.0** | **85.0** | **86.9** | **70.8** | **60.5** | **65.1** |
| WU | bridging | **92.0** | 82.5 | **87.0** | **81.1** | **68.5** | **74.3** |
| | joint_train | 91.6 | **82.7** | 86.9 | 79.4 | 67.9 | 73.1 |
| CT | bridging | **100.0** | **88.9** | **94.1** | 72.1 | **79.4** | 75.4 |
| | joint_train | 95.8 | 85.2 | 90.2 | **89.4** | 78.4 | **83.4** |
| Overall | joint_train | 89.5 | 70.6 | 78.9 | 77.5 | 61.6 | 68.6 |

Table 2: Anaphora resolution results over the test dataset (%). Models are trained for "coreference", "bridging" or "joint_train" (both tasks jointly). Models were trained over 10,000 epochs, and averaged over 3 runs with different random seeds. "$F_A$" and "$F_R$" denote the F1 score for anaphor and relation prediction, respectively.

The total loss is $L = L_{mention} + L_{ref}$, where:

$$L_{ref} = \begin{cases} L_{coref} & \text{for coreference} \\ L_{bridging} & \text{for bridging} \\ L_{coref} + L_{bridging} & \text{for joint training} \end{cases}$$

## 6 Experiments

In this section, we detail our experiments. We use similar hyperparameters to Lee et al. (2018). Specifically, we use GloVe embeddings (Pennington et al., 2014) with window size=2 for head word embeddings. For BiLSTM, GloVe embeddings with window size=10 and contextualized ELMo word representations (Peters et al., 2018) are used. Character embeddings are learned from a charactor CNN with windows of 3, 4, and 5 characters, each with 50 filters. For bridging prediction, the feed-forward neural networks are composed of two hidden layers with 150 dimensions and rectified linear units (Nair and Hinton, 2010).

We separate the gold mentions into those for coreference and bridging. For joint training, the gold mentions are combined.

Table 2 presents the results. For coreference evaluation, given that the results in Table 2 indicate that the surface and atom coreference results are not substantially different, we use surface coreference as our primary evaluation metric in the remainder of this paper. For bridging evaluation, we consider the overall bridging result as our primary analysis.

Overall, the joint training configuration achieves 54.2% $F_1$ score for coreference resolution and 71.1% $F_1$ score for bridging, representing +0.8% and +1.2% $F_1$ score absolute improvement over the component-wise models. This indicates that joint training improves the performance of both tasks. Compared to bridging, the performance of anaphor detection in coreference resolution is lower, particularly in terms of recall, possibly because the data is sparser.

| Relation | Method | $F_A$ | $F_R$ |
|---|---|---|---|
| Coref. | coreference | 62.9 | 53.4 |
| | - w/ oracle mentions | 81.7 | 79.2 |
| | joint_train | 60.5 | 54.2 |
| | - w/ oracle mentions | 79.5 | 74.9 |
| Bridging | bridging | 84.5 | 70.3 |
| | - w/ oracle mentions | 91.9 | 83.3 |
| | joint_train | 85.5 | 71.1 |
| | - w/ oracle mentions | 91.8 | 83.5 |
| Overall | joint_train | 78.9 | 68.6 |
| | - w/ oracle mentions | 88.3 | 82.1 |

Table 3: Comparisons with providing oracle mention during training; results on test dataset, using surface scoring for coreference. "$F_A$"= F1 for anaphor prediction; "$F_R$"= F1 for relation prediction.

To investigate the contribution of each step (mention detection vs. relation detection), we experiment with providing oracle mentions during the training process. Table 3 shows that the performance of both tasks improves substantially with gold mentions. We achieve 82.1% $F_1$ score for relation prediction result under joint training, with +13.5% $F_1$ absolute score improvement. That is, further improvement at mention detection will improve resolution results.

| Relation | Method | $F_A$ | $F_R$ |
|---|---|---|---|
| Coref. | coreference | 62.9 | 53.4 |
| | - w/ ChELMo | 65.3 | 56.9 |
| | joint_train | 60.5 | 54.2 |
| | - w/ ChELMo | 64.4 | 58.3 |
| Bridging | bridging | 84.5 | 70.3 |
| | - w/ ChELMo | 87.8 | 74.8 |
| | joint_train | 85.5 | 71.1 |
| | - w/ ChELMo | 88.7 | 75.6 |
| Overall | joint_train | 78.9 | 68.6 |
| | - w/ ChELMo | 82.2 | 73.1 |

Table 4: Comparison of different pretrained embeddings; results over test dataset, using surface scoring for coreference. "$F_A$"= F1 for anaphor prediction; "$F_R$"= F1 for relation prediction.

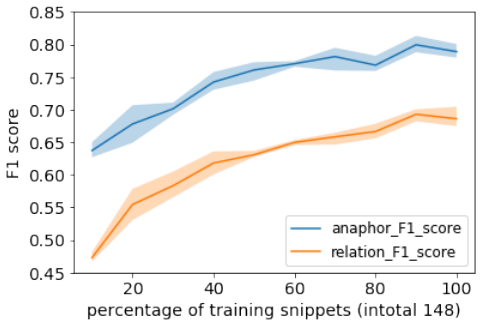


Figure 3: Joint training configuration performance on test dataset over different % of training dataset.

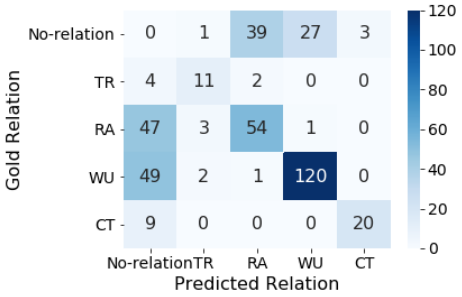


Figure 4: Confusion matrix of bridging relation detection on dev dataset with joint training configuration

To determine the importance of domain fine-tuning, we also experiment with an ELMo model pretrained on a 1 billion word chemical patent corpus (Zhai et al., 2019), referred to as ChELMo. The experimental results are provided in Table 4. With ChELMo, the performance of anaphor detection and relation detection improve by +3.3% and +4.5% absolute $F_1$ score, respectively.

We also plot model performance with increasing amounts of training data in Figure 3. While the model performance is starting to plateau, potential gains could be attained with more annotated data. The strong correlation between anaphor detection and relation detection is also self-evident in the graph.

To perform error analysis, we analysed the model errors on the dev dataset. As detailed in Table 1, the corpus contains discontinuous mentions. However, our proposed model only considers continuous spans, accounting for some of the low recall.

For coreference resolution, errors can be attributed to three primary phenomena:

1. **Long-distance relations**: as illustrated in Table 5 Ex 1, *the title compound (360 mg, 1.05 mmol, 32%)* refers to a compound at the beginning of the snippet; the model generally fails to capture such long-distance relations.

2. **Multiple antecedents**: as discussed in Section 3.3.1, an anaphor may have multiple antecedents, however the models predict a single antecedent for each anaphor.

3. **Imbalance of coreference and bridging relations**: bridging is more prevalent than coreference, meaning the model has more difficulty with coreference.

For bridging, as shown in Table 2, the performance suffers from low recall in anaphor detection. Furthermore, the confusion matrix of fine-grained bridging relations in Figure 4 shows that the model achieves poor performance for Reaction-associated and work-up relation prediction, both in terms of precision and recall.

We further investigated the over-prediction problem in bridging. As shown in Table 5 Ex 2, *the reaction mixture* in line 3 has a Relation-associated link with *The reaction mixture* in line 2 and *sodium borohydride (10 mg, 0.27 mmol)*. The model overpredicts additional links to the two additional compounds that are linked to the previous mention of *The reaction mixture* in line 2. The work-up relation in Ex 5 is similar: the second-mentioned *the organic layer* links to the first-mentioned *the organic layer* and *magnesium sulfate*. *The filtered material*, *chloroform* and *water* should be linked with the first-mentioned *the organic layer*, but are linked to the second. Such errors result from individual span-pair predictions, making it hard to capture interactions between anaphors. Evaluating the antecedents simultane-

| | |
|---|---|
| 1 | Step D: Ethyl 7-chloro-6-(difluoromethyl)-2-(trifluoromethyl)pyrazolo[1,5 -a]pyridine-3-carboxylate. ... Purification (FCC, SiO2, eluting with n-hexane:dichloromethane (2:1)) afforded the title compound (360 mg, 1.05 mmol, 32%). |
| 2 | ... to a suspension of methyl 5-bromo-6-methoxypicolinate (20 mg, 0.079 mmol) (Ark Pharm) in ethanol (0.25 mL) was added sodium borohydride (9.6 mg, 0.25 mmol). The reaction mixture was then heated at 50°C. for 2.5 h. An additional portion of sodium borohydride (10 mg, 0.27 mmol) was added, and the reaction mixture was heated at 50°C. for an additional 2 h... |
| 3 | ... after 55.8 mg of 6-chloro-7-deazapurine and 191 mg of potassium carbonate were sequentially added into the reaction mixture, the reaction mixture was refluxed for about 36 hours and then cooled down at room temperature... |
| 4 | ... In the same manner as in Synthesis Example 8 except for using 2.11 g of the intermediate 6 in place of the intermediate 21 and using 1.00 g of 4-bromobiphenyl in place of bromobenzene, 1.49 g (yield: 56%) of a white solid was obtained.,. |
| 5 | ... The filtered material was extracted with chloroform and water, and then the organic layer was dried by using magnesium sulfate. Thereafter, the organic layer was distilled under reduced pressure... |
| 6 | ... acetonitrile (150 mL) was added under ultrasonic to get a large amount of with precipitate. After suck filtration, the filter cake was washed with acetonitrile (20 mLx3), dried in vacuum to obtain the title compound (1.52 g, 86.9 %). |

Table 5: Examples of anaphora phenomena from the dev dataset.

ously may address this.

There is room for improvement in our model's ability to model context. In Table 5 Ex 3, due to the expression *add into*, the first-mentioned *the reaction mixture* does not include the chemicals mentioned prior, unlike the first mention of the phrase in Ex 2.

There are several causes of false negatives:

1. **Reaction description variation:** Chemical reactions are usually described step by step, and our model performs well in this structure. However, only part of a reaction may be described. Table 5 Ex 4 illustrates chemical compounds that are listed without a process.

2. **Abstract expressions:** In Table 5 Ex 6, *precipitate* should have a WORK-UP relation with *acetonitrile (150 mL)*, and *the title compound ...* with *the filter cake*; these are missed due to inadequate modelling of domain terminology.

## 7 Conclusion

We propose a novel annotation scheme for anaphora resolution in chemical patents. For our annotation, we incorporate generic and domain-specific knowledge to define coreference and bridging specific to the chemical domain, based on which we created the novel ChEMU-Ref dataset. Our corpus analysis and inner-annotator agreement show the complexity of the task, as well as the high quality of annotation. We model anaphora resolution as two sub-tasks, mention detection and anaphora relation detection, and also propose a joint training model, which outperforms the separately-trained models. By incorporating embeddings pretrained on the chemical domain, we found that domain knowledge boosts performance. With detailed error analysis, we also identified directions to further enhance performance.

## Acknowledgements

Funding for the ChEMU project is provided by an Australian Research Council Linkage Project, project number LP160101469, and Elsevier. A graduate research scholarship is provided by Melbourne School of Engineering to Biaoyan Fang. We would also like to thank Dr. Meladel Mistica and our two chemical expert annotators Colleen Yeow Hui Shiuan and Sacha Novakovic for their contributions to refining the annotation guidelines.

## References

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.

Nicholas Asher and Alex Lascarides. 1998. Bridging. *Journal of Semantics*, 15(1):83–113.

Michael Bada, Miriam Eckert, Donald Evans, Kristin Garcia, Krista Shipley, Dmitry Sitnikov, William A Baumgartner, K Bretonnel Cohen, Karin Verspoor, Judith A Blake, and Lawrence E Hunter. 2012. Concept annotation in the CRAFT corpus. *BMC Bioinformatics*, 13(1):161.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *3rd International Conference on Learning Representations (ICLR 2015)*, San Diego, USA.

Kevin Clark and Christopher D Manning. 2015. Entity-centric coreference resolution with model stacking. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language*

*Processing (Volume 1: Long Papers)*, pages 1405–1415, Beijing, China.

Kevin Clark and Christopher D. Manning. 2016a. Deep reinforcement learning for mention-ranking coreference models. In *Proc. of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 2256–2262, Austin, USA.

Kevin Clark and Christopher D. Manning. 2016b. Improving coreference resolution by learning entity-level distributed representations. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 643–653, Berlin, Germany.

K Bretonnel Cohen, Arrick Lanfranchi, Miji Joo-young Choi, Michael Bada, William A Baumgartner, Natalya Panteleyeva, Karin Verspoor, Martha Palmer, and Lawrence E Hunter. 2017. Coreference annotation and resolution in the colorado richly annotated full text (CRAFT) corpus of biomedical journal articles. *BMC Bioinformatics*, 18(1):372.

Biaoyan Fang, Christian Druckenbrodt, Colleen Yeow Hui Shiuan, Sacha Novakovic, Ralph Hössel, Saber A. Akhondi, Jiayuan He, Meladel Mistica, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU-Ref dataset for modeling anaphora resolution in the chemical domain. Mendeley Data.

Abbas Ghaddar and Philippe Langlais. 2016a. Wikicoref: An English coreference-annotated corpus of Wikipedia articles. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 136–142, Portorož, Slovenia.

Abbas Ghaddar and Phillippe Langlais. 2016b. WikiCoref: An English coreference-annotated corpus of Wikipedia articles. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 136–142, Portorož, Slovenia.

Loïc Grobol. 2019. Neural coreference resolution with limited lexical context and explicit mention detection for oral French. In *Proc. of the Second Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 8–14, Minneapolis, USA.

P Gwynne and G Heabrer. 2015. Recent developments in drug discovery: Improvements in efficiency. *Science*.

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. to appear. ChEMU 2021: Reaction reference resolution and anaphora resolution in chemical patents. In *Proc. of the 43rd European Conference on Information Retrieval*, online.

Jiayuan He, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Zubair Afzal, Zenan Zhai, Biaoyan Fang, Hiyori Yoshikawa, Ameer Albahem, Lawrence Cavedon, Trevor Cohn, Timothy Baldwin, and Karin Verspoor. 2020. Overview of ChEMU 2020: Named entity recognition and event extraction of chemical reactions from patents. In *Proc. of the Eleventh International Conference of the CLEF Association (CLEF 2020)*, online.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural Computation*, 9(8):1735–1780.

Yufang Hou. 2018a. A deterministic algorithm for bridging anaphora resolution. In *Proc. of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 1938–1948, Brussels, Belgium.

Yufang Hou. 2018b. Enhanced word representations for bridging anaphora resolution. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 1–7, New Orleans, USA.

Yufang Hou. 2020. Bridging anaphora resolution as question answering. In *Proc. of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1428–1438, Online.

Yufang Hou, Katja Markert, and Michael Strube. 2014. A rule-based system for unrestricted bridging resolution: Recognizing bridging anaphora and finding links to antecedents. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP 2014)*, pages 2082–2093, Doha, Qatar.

Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Ben Kantor and Amir Globerson. 2019. Coreference resolution with entity equalization. In *Proc. of the 57th Annual Meeting of the Association for Computational Linguistics (ACL 2019)*, pages 673–677, Florence, Italy.

Jin-Dong Kim, Ngan Nguyen, Yue Wang, Jun'ichi Tsujii, Toshihisa Takagi, and Akinori Yonezawa. 2012. The Genia event and protein coreference tasks of the BioNLP shared task 2011. *BMC Bioinformatics*, 13(11):S1.

Klaus Krippendorff. 2004. Measuring the reliability of qualitative text analysis data. *Quality and quantity*, 38:787–800.

Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution. In *Proc. of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 188–197, Copenhagen, Denmark.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 2 (Short Papers)*, pages 687–692, New Orleans, USA.

Jiao Li, Yueping Sun, Robin J Johnson, Daniela Sciaky, Chih-Hsuan Wei, Robert Leaman, Allan Peter Davis, Carolyn J Mattingly, Thomas C Wiegers, and Zhiyong Lu. 2016. BioCreative V CDR task corpus: a resource for chemical disease relation extraction. *Database*, 2016.

Xiaoqiang Luo. 2005. On coreference resolution performance metrics. In *Proc. of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing (EMNLP 2005)*, pages 25–32, Vancouver, Canada.

Nafise Sadat Moosavi and Michael Strube. 2016. Which coreference evaluation metric do you trust? a proposal for a link-based entity aware metric. In *Proc. of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 632–642, Berlin, Germany.

Vinod Nair and Geoffrey E Hinton. 2010. Rectified linear units improve restricted Boltzmann machines. In *Proc. of the 33rd International Conference on Machine Learning (ICML 2016)*, New York, USA.

Vincent Ng. 2017. Machine learning for entity coreference resolution: A retrospective look at two decades of research. In *Proc. of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI'17)*, pages 4877–4884, San Francisco, USA.

Ngan Nguyen, Jin-Dong Kim, and Jun'ichi Tsujii. 2011. Overview of BioNLP 2011 protein coreference shared task. In *Proc. of BioNLP Shared Task 2011 Workshop*, pages 74–82, Portland, USA.

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. GloVe: Global vectors for word representation. In *Proc. of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1532–1543, Doha, Qatar.

Matthew Peters, Mark Neumann, Mohit Iyyer, Matt Gardner, Christopher Clark, Kenton Lee, and Luke Zettlemoyer. 2018. Deep contextualized word representations. In *Proc. of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, pages 2227–2237, New Orleans, USA.

Massimo Poesio, Ron Artstein, et al. 2008. Anaphoric annotation in the ARRAU corpus. In *Proc. of the Sixth International Conference on Language Resources and Evaluation (LREC 2008)*, Marrakech, Morocco.

Massimo Poesio, Roland Stuckardt, and Yannick Versley. 2016. *Anaphora Resolution*. Springer.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. CoNLL-2012 shared task: Modeling multilingual unrestricted coreference in OntoNotes. In *Proc. of EMNLP-CoNLL 2012: Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pages 1–40, Jeju, Korea.

Marta Recasens and Eduard Hovy. 2011. BLANC: Implementing the Rand index for coreference evaluation. *Natural Language Engineering*, 17(4):485–510.

Ina Rösiger. 2016. Scicorp: A corpus of English scientific articles annotated for information status analysis. In *Proc. of the Tenth International Conference on Language Resources and Evaluation (LREC 2016)*, pages 1743–1749, Portorož, Slovenia.

Ina Rösiger. 2018a. BASHI: A corpus of Wall Street Journal articles annotated with bridging links. In *Proc. of the Eleventh International Conference on Language Resources and Evaluation (LREC-2018)*.

Ina Rösiger. 2018b. Rule- and learning-based methods for bridging resolution in the ARRAU corpus. In *Proc. of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 23–33, New Orleans, USA.

Ina Rösiger. 2019. *Computational modelling of coreference and bridging resolution*. Ph.D. thesis, Stuttgart University.

Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018. Bridging resolution: Task definition, corpus resources and rule-based experiments. In *Proc. of the 27th International Conference on Computational Linguistics*, pages 3516–3528, Santa Fe, USA.

Karin Verspoor, Dat Quoc Nguyen, Saber A. Akhondi, Christian Druckenbrodt, Camilo Thorne, Ralph Hoessel, Jiayuan He, and Zenan Zhai. 2020. ChEMU dataset for information extraction from chemical patents. Mendeley Data.

Ralph Weischedel, Martha Palmer, Mitchell Marcus, Eduard Hovy, Sameer Pradhan, Lance Ramshaw, Nianwen Xue, Ann Taylor, Jeff Kaufman, Michelle Franchini, Mohammed El-Bachouti, Robert Belvin, and Ann Houston. 2013. OntoNotes release 5.0. Linguistic Data Consortium Catalog No. LDC2013T19.

Sam Wiseman, Alexander M. Rush, Stuart Shieber, and Jason Weston. 2015. Learning anaphoricity and antecedent ranking features for coreference resolution. In *Proc. of the 53rd Annual Meeting of the Association for Computational Linguistics and the 7th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, pages 1416–1426, Beijing, China.

Sam Wiseman, Alexander M. Rush, and Stuart M. Shieber. 2016. Learning global features for coreference resolution. In *Proc. of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 994–1004, San Diego, USA.

Juntao Yu and Massimo Poesio. 2020. Multi-task learning based neural bridging reference resolution. *arXiv preprint arXiv:2003.03666*.

Amir Zeldes. 2017. The GUM corpus: Creating multilayer resources in the classroom. *Language Resources and Evaluation*, 51(3):581–612.

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 *Proc. of the 18th BioNLP Workshop and Shared Task*, pages 328–338, Florence, Italy.

Rui Zhang, Cícero Nogueira dos Santos, Michihiro Yasunaga, Bing Xiang, and Dragomir Radev. 2018. Neural coreference resolution with deep biaffine attention by joint mention detection and mention clustering. In *Proc. of the 56th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 102–107, Melbourne, Australia.

## A Additional Experimental Results

In the following tables, we provide detailed experiment results described in the main paper.

Table 6 provides a full comparison of training with gold-standard oracle mentions per anaphora relation on the test dataset.

Table 7 provides a full comparison of training with different pretrained embeddings per anaphora relation on the test dataset.

| Relation | | Method | $P_A$ | $R_A$ | $F_A$ | $P_R$ | $R_R$ | $F_R$ |
|---|---|---|---|---|---|---|---|---|
| Coref. (Surface) | | coreference | 84.9 | 50.0 | 62.9 | 73.7 | 41.9 | 53.4 |
| | | - w/ oracle mentions | 86.0 | 78.1 | 81.7 | 84.8 | 74.5 | 79.2 |
| | | joint_train | 89.4 | 45.8 | 60.5 | 81.7 | 40.6 | 54.2 |
| | | - w/ oracle mentions | 90.5 | 70.8 | 79.5 | 87.0 | 65.9 | 74.9 |
| Coref. (Atom) | | coreference | 84.9 | 50.0 | 62.9 | 75.6 | 42.6 | 54.4 |
| | | - w/ oracle mentions | 86.0 | 78.1 | 81.7 | 85.1 | 74.7 | 79.5 |
| | | joint_train | 89.4 | 45.8 | 60.5 | 82.3 | 40.8 | 54.5 |
| | | - w/ oracle mentions | 90.5 | 70.8 | 79.5 | 88.7 | 66.3 | 75.9 |
| Bridging | | bridging | 88.4 | 80.9 | 84.5 | 76.0 | 65.4 | 70.3 |
| | | - w/ oracle mentions | 91.1 | 92.8 | 91.9 | 83.8 | 82.8 | 83.3 |
| | | joint_train | 89.5 | 81.8 | 85.5 | 77.0 | 66.1 | 71.1 |
| | | - w/ oracle mentions | 91.3 | 92.4 | 91.8 | 82.8 | 84.3 | 83.5 |
| | TR | bridging | 77.5 | 63.8 | 69.7 | 76.2 | 63.8 | 69.1 |
| | | - w/ oracle mentions | 90.2 | 90.8 | 90.3 | 90.2 | 90.8 | 90.3 |
| | | joint_train | 76.9 | 69.0 | 72.7 | 75.9 | 69.0 | 72.3 |
| | | - w/ oracle mentions | 91.5 | 86.2 | 88.6 | 90.6 | 86.2 | 88.1 |
| | RA | bridging | 82.7 | 83.3 | 83.0 | 66.0 | 57.5 | 61.4 |
| | | - w/ oracle mentions | 88.0 | 88.3 | 88.1 | 83.4 | 71.1 | 76.7 |
| | | joint_train | 89.0 | 85.0 | 86.9 | 70.8 | 60.5 | 65.1 |
| | | - w/ oracle mentions | 85.4 | 93.9 | 89.4 | 78.0 | 76.6 | 77.3 |
| | WU | bridging | 92.0 | 82.5 | 87.0 | 81.1 | 68.5 | 74.3 |
| | | - w/ oracle mentions | 92.4 | 94.4 | 93.4 | 83.7 | 86.7 | 85.2 |
| | | joint_train | 91.6 | 82.7 | 86.9 | 79.4 | 67.9 | 73.1 |
| | | - w/ oracle mentions | 93.7 | 92.6 | 93.1 | 85.2 | 86.9 | 86.0 |
| | CT | bridging | 100.0 | 88.9 | 94.1 | 72.1 | 79.4 | 75.4 |
| | | - w/ oracle mentions | 93.3 | 100.0 | 96.5 | 79.5 | 100.0 | 88.3 |
| | | joint_train | 95.8 | 85.2 | 90.2 | 89.4 | 78.4 | 83.4 |
| | | - w/ oracle mentions | 90.6 | 100.0 | 94.9 | 71.9 | 100.0 | 83.3 |
| Overall | | joint_train | 89.5 | 70.6 | 78.9 | 77.5 | 61.6 | 68.6 |
| | | - w/ oracle mentions | 91.1 | 85.7 | 88.3 | 83.4 | 81.0 | 82.1 |

Table 6: Test results with gold-standard mentions during training. Models trained for "coreference", "bridging" or "joint_train" (both tasks jointly). Models trained over 10,000 epochs; averaged over 3 runs with different random seeds. "$F_A$" and "$F_R$" denote the F1 score for anaphor and relation prediction, respectively.

| Relation | | Method | $P_A$ | $R_A$ | $F_A$ | $P_R$ | $R_R$ | $F_R$ |
|---|---|---|---|---|---|---|---|---|
| Coref. (Surface) | | coreference | 84.9 | 50.0 | 62.9 | 73.7 | 41.9 | 53.4 |
| | | - w/ CHELMO | 86.5 | 52.5 | 65.3 | 76.9 | 45.2 | 56.9 |
| | | joint_train | 89.4 | 45.8 | 60.5 | 81.7 | 40.6 | 54.2 |
| | | - w/ CHELMO | 87.2 | 51.1 | 64.4 | 80.6 | 45.7 | 58.3 |
| Coref. (Atom) | | coreference | 84.9 | 50.0 | 62.9 | 75.6 | 42.6 | 54.4 |
| | | - w/ CHELMO | 86.5 | 52.5 | 65.3 | 81.1 | 46.5 | 59.0 |
| | | joint_train | 89.4 | 45.8 | 60.5 | 82.3 | 40.8 | 54.5 |
| | | - w/ CHELMO | 87.2 | 51.1 | 64.4 | 81.5 | 46.0 | 58.8 |
| Bridging | | bridging | 88.4 | 80.9 | 84.5 | 76.0 | 65.4 | 70.3 |
| | | - w/ CHELMO | 88.4 | 87.2 | 87.8 | 75.9 | 73.7 | 74.8 |
| | | joint_train | 89.5 | 81.8 | 85.5 | 77.0 | 66.1 | 71.1 |
| | | - w/ CHELMO | 91.3 | 86.4 | 88.7 | 78.1 | 73.4 | 75.6 |
| | TR | bridging | 77.5 | 63.8 | 69.7 | 76.2 | 63.8 | 69.1 |
| | | - w/ CHELMO | 82.5 | 65.5 | 73.0 | 81.4 | 65.5 | 72.5 |
| | | joint_train | 76.9 | 69.0 | 72.7 | 75.9 | 69.0 | 72.3 |
| | | - w/ CHELMO | 81.4 | 64.4 | 71.8 | 79.3 | 64.4 | 70.9 |
| | RA | bridging | 82.7 | 83.3 | 83.0 | 66.0 | 57.5 | 61.4 |
| | | - w/ CHELMO | 89.5 | 83.9 | 86.5 | 74.1 | 62.9 | 68.0 |
| | | joint_train | 89.0 | 85.0 | 86.9 | 70.8 | 60.5 | 65.1 |
| | | - w/ CHELMO | 91.8 | 82.2 | 86.6 | 76.8 | 63.7 | 69.5 |
| | WU | bridging | 92.0 | 82.5 | 87.0 | 81.1 | 68.5 | 74.3 |
| | | - w/ CHELMO | 88.3 | 92.4 | 90.3 | 76.2 | 79.0 | 77.5 |
| | | joint_train | 91.6 | 82.7 | 86.9 | 79.4 | 67.9 | 73.1 |
| | | - w/ CHELMO | 92.2 | 92.0 | 92.1 | 78.5 | 78.2 | 78.3 |
| | CT | bridging | 100.0 | 88.9 | 94.1 | 72.1 | 79.4 | 75.4 |
| | | - w/ CHELMO | 100.0 | 85.2 | 91.7 | 78.5 | 78.4 | 78.4 |
| | | joint_train | 95.8 | 85.2 | 90.2 | 89.4 | 78.4 | 83.4 |
| | | - w/ CHELMO | 100.0 | 81.5 | 89.4 | 80.8 | 80.4 | 80.3 |
| Overall | | joint_train | 89.5 | 70.6 | 78.9 | 77.5 | 61.6 | 68.6 |
| | | - w/ CHELMO | 90.4 | 75.3 | 82.2 | 78.4 | 68.5 | 73.1 |

Table 7: Results with different pretrained embeddings. "coreference", "bridging" and "joint_training" represent models that are trained on the coreference resolution task, bridging task, and both tasks jointly, respectively. We train the models over 10,000 epochs, and averages over 3 runs with different random seeds. "$F_A$" and "$F_R$" denote the F1 score for anaphor and relation prediction, respectively.