What does it take to bake a cake? The RecipeRef corpus and anaphora resolution in procedural text

Biaoyan Fang[1], **Timothy Baldwin**[3,1] **and Karin Verspoor**[2,1]
[1]The University of Melbourne, Australia
[2]RMIT University, Australia
[3]MBZUAI, Abu Dhabi
biaoyanf@student.unimelb.edu.au
{tbaldwin,karin.verspoor}@unimelb.edu.au

Abstract

Procedural text contains rich anaphoric phenomena, yet has not received much attention in NLP. To fill this gap, we investigate the textual properties of two types of procedural text, recipes and chemical patents, and generalize an anaphora annotation framework developed for the chemical domain for modeling anaphoric phenomena in recipes. We apply this framework to annotate the RecipeRef corpus with both bridging and coreference relations. Through comparison to chemical patents, we show the complexity of anaphora resolution in recipes. We demonstrate empirically that transfer learning from the chemical domain improves resolution of anaphora in recipes, suggesting transferability of general procedural knowledge.

1 Introduction

Anaphora resolution is a core component in information extraction tasks (Poesio et al., 2016; Rösiger, 2019) and critical for various downstream natural language processing tasks, such as named entity recognition (Dai et al., 2019) and machine translation (Stanovsky et al., 2019). It consists of two primary anaphoric types, coreference (Ng, 2017; Clark and Manning, 2015) and bridging (Asher and Lascarides, 1998; Rösiger et al., 2018). Most anaphora corpora (Pradhan et al., 2012; Ghaddar and Langlais, 2016; Poesio et al., 2008), however, only focus on either coreference or bridging. To fill the gap in anaphora resolution, it is becoming increasingly important to have both types annotated.

Current research on anaphora resolution is mostly based on declarative text (Pradhan et al., 2012; Ghaddar and Langlais, 2016; Rösiger, 2018a; Hou et al., 2018), such as news or dialogue. Procedural text, such as chemical patents or instruction manuals, has received limited attention despite being critical for human knowledge (Yamakata et al., 2020). In turn, correct resolution of entities is the cornerstone of procedural text comprehension—resolution of anaphora in these texts is required to determine what action applies to which entity.

We focus in this work on the procedural text type of recipes. As shown in Fig. 1, recipes have rich and complex anaphora phenomena. Here, the expression *the biscuits* appears several times in text; while each occurrence relates to the same *biscuits* concept, their state and semantic meaning vary.

Our aim in this paper is to address anaphora resolution in procedural text, especially for recipes, identifying anaphoric references and determining the relationships among the entities. We first investigate the textual properties of procedural texts, i.e. chemical patents and recipes. We then adapt an existing anaphora annotation schema developed for chemical patents (Fang et al., 2021a,b) to recipes, and define four types of anaphora relationships, encompassing coreference and bridging. We further create a dataset based on this schema and achieve high inter-annotator agreement with two annotators experienced with the domain. We additionally explore the feasibility of applying transfer learning from the chemical domain to model recipe anaphora resolution. The dataset and related code are publicly available.[1]

Our contributions in this paper include: (1) adaptation of the anaphora annotation framework from chemical patents for modeling anaphoric phenomena in recipes; (2) creation of a publicly accessible recipe anaphora resolution dataset based on the annotation framework (Fang et al., 2022); (3) investigation of the textual properties of chemical patents and recipes; and (4) demonstration of the benefit of utilizing procedural knowledge from the chemical domain to enhance recipe anaphora resolution via transfer learning.

[1]Code is available at https://github.com/biaoyanf/RecipeRef, and the dataset is available at http://doi.org/10.17632/rcyskfvdv7.1.

Findings of the Association for Computational Linguistics: ACL 2022, pages 3481 - 3495
May 22-27, 2022

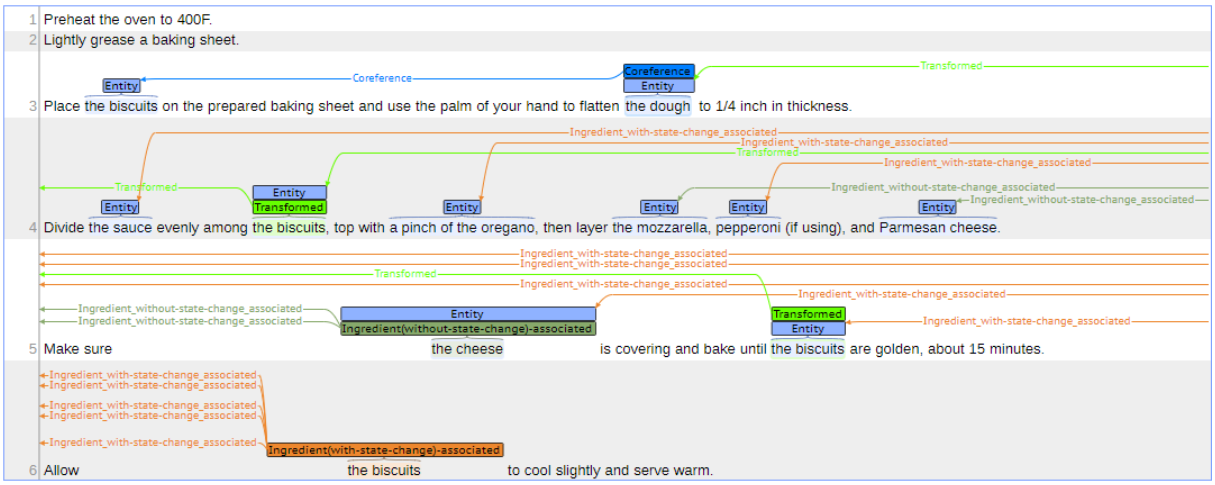

Figure 1: Excerpt of a recipe annotated for anaphora. Different color links represent different anaphora relation types. Detailed anaphora relation definitions are provided in Section 3.3.

2 Related Work

Anaphora relation subsumes two referring types: (1) coreference — expressions in the text that refer to the same entity (Clark and Manning, 2015; Ng, 2017); and (2) bridging — expressions that do not refer to the same entity, but are linked via semantic, lexical, or encyclopedic relations (Asher and Lascarides, 1998; Hou et al., 2018).

Existing anaphora corpora mostly focus on declarative text, across a range of domains (Poesio et al., 2008; Pradhan et al., 2012; Ghaddar and Langlais, 2016; Cohen et al., 2017). There have been attempts to annotate procedural text corpora for anaphora, but most focus exclusively on coreference (Mysore et al., 2019; Friedrich et al., 2020).

Pradhan et al. (2012) developed the CoNLL 2012 corpus for generic coreference resolution. It consists of declarative texts including news and magazine articles, across three languages — English, Chinese, and Arabic. This corpus adopted the OntoNotes 5.0 (Weischedel et al., 2013) annotation scheme, modeling coreference in terms of two subtypes: (1) identity, where the anaphoric references and referents are identical; and (2) appositive, where a noun phrase is modified by an intermediately-adjacent noun phrase. It models coreference as a clustering task, ignoring the direction of relations. Following largely the same annotation framework, the WikiCoref corpus (Ghaddar and Langlais, 2016) targeted Wikipedia texts. The InScript corpus (Modi et al., 2016) consists of 1,000 stories from 10 different scenarios corresponding to a "script", i.e. a standardised sequence of events. The corpus includes coreference annotations for noun phrases.

BioNLP-ST 2011 (Nguyen et al., 2011) is a gene-related coreference corpus based on abstracts from biomedical publications. It consists of four types of coreference: RELAT (relative pronouns or relative adjectives, e.g. *that*), PRON (pronouns, e.g. *it*), DNP (definite NPs or demonstrative NPs, e.g. NPs that begin with *the*) and APPOS (coreferences in apposition). As it only focuses on gene-related annotation, coreference is limited. CRAFT-ST 2019 (Cohen et al., 2017) annotates 97 full biomedical articles for coreference resolution, based on a slightly-modified version of the OntoNotes 5.0 annotation scheme. Compared to the BioNLP 2011 corpus, it contains a wider range of relation types, and is not limited to only abstracts. SCIERC (Luan et al., 2018) contains 500 abstracts from scientific articles, and coreference annotation.

Due to the complexities of defining bridging (Zeldes, 2017; Hou et al., 2018), different corpora have adopted different definitions of bridging. According to Rösiger et al. (2018), bridging can be divided into: (1) *referential*, where the anaphoric references rely on the referent to be interpretable (e.g. *a new town hall – the door*, *the old oak tree – leaves*, etc.); and (2) *lexical*, encompassing lexical-semantic relations, such as meronymy or hyponymy (e.g. *Europe* and *Spain* are in a whole-part relation). The ARRAU corpus (Poesio et al., 2008) consists of three types of declarative text: news, dialogue and narrative text. The bridging annotations are mostly lexical, with a much smaller number of referential references. The ISNotes corpus (Hou et al., 2018) is based on 50 Wall Street

Journal (WSJ) texts from the OntoNotes corpus, and contains both coreference and referential bridging. Similar to ISNotes, BASHI (Rösiger, 2018a) is based on another 50 WSJ texts from OntoNotes with referential bridging. With the same annotation scheme as BASHI, SciCorp (Rösiger, 2016) focuses on scientific text and referential bridging.

A small number of domain-specific anaphora corpora have been developed for procedural text. The ChEMU-ref corpus (Fang et al., 2021a) contains 1,500 chemical patent excerpts describing chemical reactions. Based on generic and chemical knowledge, the corpus contains five types of anaphora relationships, i.e. Coreference, Transfers, Reaction-associated, Work-up, and Contained. Friedrich et al. (2020) developed the SOFC-Exp corpus based on 45 material sciences articles, for the purposes of information extraction. The corpus is primarily targeted at named entity recognition and relation extraction, with coreference as a secondary annotation task, based on coindexation between a common noun or pronoun and a more specific mention earlier in the text. Also in the context of material sciences, Mysore et al. (2019) annotated 230 synthesis procedures for coreference, largely based on text in parentheses and coreferent abbreviations.

Recent work in recipe comprehension includes visual instructions (Huang et al., 2017; Nishimura et al., 2020) and linguistic texts (Agarwal and Miller, 2011; Kiddon et al., 2015; Jiang et al., 2020) across Japanese (Harashima and Hiramatsu, 2020; Harashima et al., 2016) and English (Batra et al., 2020; Marin et al., 2019). Most research analyzes the text of recipes as a workflow graph based on actions (Kiddon et al., 2015; Mori et al., 2014; Yamakata et al., 2020), where the vertices represent name entities (e.g. action, food, etc.) and edges represent relational structure (e.g. action complement, food complement, etc.). Although interactions among ingredients can be derived via action nodes, this approach doesn't sufficiently capture anaphora phenomena, i.e. coreference and bridging. The RISeC corpus (Jiang et al., 2020) identifies candidate expressions for zero anaphora verbs in English recipes. However, they do not capture generic anaphoric phenomena.

In terms of modeling, most research has handled coreference and bridging separately due to limited data availability (and a lack of annotated datasets containing both coreference and bridging). For coreference resolution, span ranking models (Lee et al., 2017, 2018) have become the benchmark method, supplanting mention ranking models (Clark and Manning, 2015, 2016a,b; Wiseman et al., 2015, 2016). Various span ranking variants have been proposed (Zhang et al., 2018; Grobol, 2019; Kantor and Globerson, 2019), and achieved strong results. With the increasing number of coreference corpora, transfer learning (Brack et al., 2021; Xia and Van Durme, 2021) involving pretraining on a source domain and fine-tuning on a target domain has shown great potential at improving coreference resolution. Bridging methods can be categorised into: (1) rule-based methods (Hou et al., 2014; Rösiger et al., 2018; Rösiger, 2018b); and (2) machine learning methods (Hou, 2018a,b, 2020; Yu and Poesio, 2020). Hou (2020) modeled bridging resolution as a question answering task, and fine-tuned the question answering model from generic question answering corpora. By utilizing transfer learning, they achieved a stronger performance on the bridging task. Yu and Poesio (2020) proposed a joint training framework for bridging and coreference resolution based on an end-to-end coreference model (Lee et al., 2017). Similar to coreference, they modeled bridging as a clustering task. Through joint training, they achieved substantial improvements for bridging, but the impact on coreference was less clear. Fang et al. (2021a) adopted the same end-to-end framework for joint training, modeling bridging as a mention pair classification task, and achieved improvements on both subtasks.

3 Annotation Scheme

In this section, we describe our adapted annotation scheme for recipe anaphora annotation. The complete annotation guideline is available at Fang et al. (2022).

3.1 Corpus Selection

We create our RecipeRef dataset by random sampling texts from RecipeDB (Batra et al., 2020), a large, diverse recipe database containing 118,171 English recipes with 268 processes and more than 20,262 ingredients. It consists of ingredient lists and instruction sections. We select the instruction section of each recipe, which details the steps for preparing the dish.

3.2 Mention Types

As our goal is to capture anaphora in recipes, we focus on ingredient-related expressions. In line with previous work (Pradhan et al., 2012; Cohen et al., 2017; Fang et al., 2021a; Ghaddar and Langlais, 2016), we leave out singleton mentions, i.e. mentions that are not involved in anaphora relations (as defined in Section 3.3) are not annotated. Mention types that are considered for anaphora relations are listed below.

Ingredient Terms: In recipes, ingredient terms are essential as they indicate what ingredients are used, in the form of individual words or phrases, such as *butter*, *endive heads*, *red peppers*, or *garlic powder*.

Referring Expressions: We consider referring expressions to be pronouns (e.g. *it* or *they*) and generic phrases (e.g. *soup*, or *the pastry mixture*) used to represent ingredients that were previously introduced in the recipe text.

We adopt several criteria in annotating mentions:

- **Premodifiers**: One of the key challenges in procedural text is to track state changes in entities. It is critical to include premodifiers, as they play an important role in identifying an entity's state. We consider ingredients with premodifiers to be atomic mentions, e.g. *chopped chicken*, *roasted red peppers*, and *four sandwiches*.[2]

- **Numbers**: In some cases, standalone numeric expressions can be used to reference to ingredients, and in such cases are considered to be mentions. Examples of this are *1* in *Beat eggs, 1 at a time*, and *three* in *Combine together to make a sandwich. Repeat to make three*.

3.3 Relation Types

A core challenge in procedural text comprehension is tracking the state of each entity (Dalvi et al., 2018; Tandon et al., 2018). Recipes contain rich information about changes in the state of ingredients. As shown in Fig. 1, to obtain *the biscuits* in line 6, *the biscuits* in line 1 has gone through several processes, involving physical (e.g. *flatten*) and chemical change (e.g. *bake*). Capturing labeled interactions between ingredients provides a richer understanding of ingredients and their interactions (i.e. where is the ingredient from).

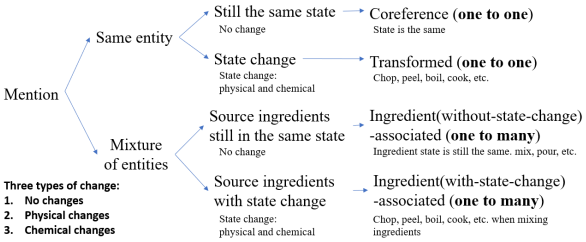

Figure 2: Overall schema of anaphora relations for recipes.

There are two basic types of anaphora: coreference and bridging. In recipes, we define bridging according to three subtypes of referring relations, based on the state of entities (with coreference making up the fourth subtype). The overall schema of anaphora relations for recipes is shown in Fig. 2.

In anaphora resolution, an *antecedent* is a linguistic expression that anchors the interpretation of a second expression, the *anaphor*, which cannot be interpreted in isolation or has little meaning on its own. *Anaphors* are linked to *antecedents* via anaphora relations. Consistent with previous work, we limit *anaphors* to link to *antecedents* appearing earlier in the text (i.e. we do not annotate instances of cataphora, which we found to occur very rarely in recipe texts), and the direction of links is preserved.

3.3.1 Coreference

In general applications, coreference focuses on expressions that refer to the same entity in the real-world (Clark and Manning, 2015; Ng, 2017). In procedural text, the state of an entity can be changed by an action applied to that entity. To capture state changes, we add an extra constraint on coreference in requiring that the two mentions refer to the same entity *in the same state*.

To eliminate ambiguity in linking coreferent mentions, the closet antecedent is linked for a given anaphor.

3.3.2 Bridging

As discussed in Section 3.3.1, we consider the state of entities to interface with anaphora in procedural text. As such, we define three subtypes of bridging relations, based on the state of the entities involved.

TRANSFORMED A one-to-one anaphoric link for an ingredient that is meaning-wise the same

[2]We use the term "premodifier" somewhat loosely, in that, strictly speaking, expressions such as *four* in our example are specifiers rather than premodifiers.

Combination Process	Chemical Patents	...5-Isopropylisoxazol-3-carboxylic acid (1.00 g, 6.45 mmol) was dissolved in methanol (20 mL), and thionyl chloride (1.51 g, 12.9 mmol) was slowly added at 0°C. The reaction solution was slowly warmed to 25°C and stirred for 12 hour...
	Recipes	... mix 2 tablespoons of the olive oil, chili powder, allspice, salt, and pepper in a small bowl and brush the turkey all over with the spice mixture...
Removal Process	Chemical Patents	...the mixture was extracted three times with ethyl acetate (50 mL). The combined ethyl acetate layer was washed with saturated brine (50 mL) and dried over anhydrous sodium sulfate...
	Recipes	...add chicken thighs to the broth and simmer until cooked through, about 10 minutes. remove chicken with slotted spoon and set aside; when cool enough to handle, slice thinly. continue to simmer broth, return to pot...

Table 1: Examples of processes in chemical patents and recipes.

but has undergone physical/chemical change (e.g. *peeling*, *baking*, or *boiling*). For example, in Fig. 1, *the biscuits* in line 4 and 5 are annotated as TRANSFORMED because of the *bake* action that changes the state of *the biscuits* in line 4.

INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED A one-to-many relationship between a processed food mention and its source ingredients, where the source ingredients have not undergone a state change (i.e. physical/chemical change). As shown in Fig. 1, *the cheese* in line 5 refers to its source ingredients *the mozzarella* and *Parmesan cheese* in line 4 and there is no state change. Thus, they are annotated as INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED.

INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED A one-to-many relationship between a processed food mention and its source ingredients, involving a state change. As an example, *the biscuits* in Fig. 1 line 6 is a combination of previously-mentioned source ingredients (i.e. *the sauce*, *a pinch of the oregano*, *pepperoni*, *the cheese*, and *the biscuits*) involving a state change through baking. They are thus annotated as INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED.

3.4 Comparison with Chemical Patents

As shown in Table 1, chemical patents and recipes have many commonalities. They use similar language to describe the application of processes (e.g. combination or removal) to source entities to obtain new entities, making it feasible to adapt the anaphora annotation scheme from chemical patents (Fang et al., 2021a,b) to recipes.

However, there are some key differences in the annotation schemes.

- **Domain Differences**: Some relation types defined for chemical patents are domain-specific, e.g. the WORK-UP relation is specific to chemistry and cannot be directly applied to recipes.

- **Determining State Change**: In both chemical patents and recipes, anaphora resolution aims to capture anaphoric relations between mentions involving possible state changes. In the chemical domain, we are most concerned with chemical changes (e.g. *oxidation* or *acidification*). However, in the recipe domain, we are also interested in physical changes (e.g. *chop* or *slice*).

- **Rich Semantic Meaning in Recipes**: Ingredient terms in recipes may represent a combination of ingredients. As shown in Fig. 1, *the biscuits* in line 6 represent a combination of previously-mentioned ingredients and not just the biscuit ingredient itself. However, in chemical patents, chemical names have specific meanings and cannot be semantically extended. This is a key challenge in resolving anaphora in recipes.

- **Variability in Instruction Descriptions**: Although chemical patents and recipes have similar structure, instruction descriptions in recipes are structurally more variable. In chemical patents, processed entities are mostly directly used in the immediately-proceeding process. However, processed entities in recipes can be mentioned far later in the text (esp. in "modular" recipes, e.g. where a cake, cake filling, and cake icing are separately prepared, and only combined in a final step).

- **Hierarchical Structure in Recipe Relation Types**: Anaphora relation types in recipes are defined hierarchically (as shown in Fig. 2), such that a simplified version of the recipe anaphora resolution task, without considering state change, can be easily derived. In chemical patents, there is no clear way of simplify-

	RecipeRef	ChEMU-ref
Documents	80	1,125
Sentences	999	5,768
Tokens per sentence	12.6	27.6
Mentions	1,408	17,023
Mentions per doc	17.6	15.1
COREF	229 / 415	3,243
COREF per doc	2.9 / 5.2	2.9
Bridging*	1,104 / 918	12,796
Bridging* per doc	13.8 / 11.5	11.4
TR	186 / —	—
IWOA	91 / 918	—
IWA	827 / —	—

Table 2: Corpus statistics. For ChEMU-ref, we include the training and development set. "COREF", "TR", "IWOA" and "IWA" denote the COREFERENCE, TRANSFORMED, INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED and INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED relations, respectively. "/" shows the number of relations with and without consideration of state change. "Bridging*" is the total number of bridging relations across all subtypes.

ing the scheme while preserving the anaphoric relations.

4 Task Definition

Following the approach of Fang et al. (2021a), anaphora resolution is modeled as a two-step task: (1) mention detection; and (2) anaphora relation detection.

As anaphora relation types in recipes are defined hierarchically, we can derive a simplified version of the recipe anaphora resolution task by removing state changes. That is, COREFERENCE and TRANSFORMED can be merged when we remove consideration of state changes, and INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED and INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED can similarly be merged. As such, we evaluate recipe anaphora resolution both with state change (4-way), and without state change (2-way).

As our corpus includes one-to-many anaphoric relations for bridging, standard coreference evaluation metrics (Luo, 2005; Recasens and Hovy, 2011; Moosavi and Strube, 2016), which assume a given mention only occurs in a unique cluster, are not suitable for this task. Although coreferences involving one-to-one relations in this task could be evaluated with these metrics, to maintain a unified evaluation for bridging and coreference, we utilize precision, recall and F1 as our core metrics.[3] Specifically, we follow the evaluation of the ChEMU-ref corpus, scoring coreference from two perspectives: (1) surface coreference, where a coreferent anaphor links to its closest antecedent; and (2) atom coreference, where a coreferent anaphor links to a correct antecedent (Kim et al., 2012).

For manual annotation, we use the Brat rapid annotation tool.[4] In an attempt to achieve high quality, we went through 8 rounds of annotation training and refinement of the anaphora annotation with two annotators experienced with the recipe domain. In each round of training, the annotators independently annotated 10 recipes (different for each round of annotation) and met afterwards to compare annotation results. Further refinements of the annotation guidelines were made based on the discussion.

After training, we reached a high inter-annotator agreement (IAA) of Krippendorff's α = 0.85, mention-level F1 = 0.88, and relation-level F1 = 0.67. As a point of comparison, the respective values after the first round of annotator training were 0.45, 0.51 and 0.29, respectively.

We use 80 double-annotated recipes with harmonized annotations as our corpus. The statistics of this corpus in comparison with the ChEMU-ref corpus (Fang et al., 2021a) are shown in Table 2.

5 Methodology

To investigate the benefit of transfer learning from the chemical domain, we follow the configuration of Fang et al. (2021a), modeling bridging as a classification task and adopting the benchmark end-to-end neural coreference model of Lee et al. (2017, 2018) for joint training of the two anaphora resolution types.

For each span x_i, the model learns: (1) a mention score s_{m_i} for mention detection:

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

and (2) a distribution $P(\cdot)$ over possible antecedent spans $Y(i)$ for coreference resolution:

$$P(y) = \frac{\exp(s_c(i,y))}{\sum_{y' \in Y} \exp(s_c(i,y'))}$$

[3] We additionally include results based on standard coreference metrics for coreference only (but not bridging, due to the many-to-one relations) in Appendix A.

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

where $s_c(i, y)$ is the output of a feed-forward neural network with span pair embedding $s_{i,y}$, and (3) a pair-wise score $s_b(i, y)$ of each possible antecedent span y for bridging resolution:

$$s_b(i, y) = \text{softmax}(w_b \cdot \text{FFNN}_b(s_{i,y}))$$

A span representation s_i is 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 mechanism (Bahdanau et al., 2015), and a feature vector of the mention ($\phi(i)$):

$$s_i = [x^*_{\text{START}(i)}, x^*_{\text{END}(i)}, h_i, \phi(i)]$$

where START(i) and END(i) represent the starting and ending token index for span i, respectively.

A span pair embedding $s_{i,y}$ is obtained by the concatenation of each span embedding ($s(i), s(y)$) and the element-wise multiplication of the span embeddings ($s(i) \circ s(y)$) and a feature vector ($\phi(i, y)$) for span pair i and y:

$$s_{i,y} = [s(i), s(y), s(i) \circ s(y), \phi(i, y)]$$

For mention loss, we use cross-entropy loss:

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

where:

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

and GOLD_m is the set of gold mentions that are involved in anaphora relations.

For coreference resolution, we compute the loss as follows, where $\text{GOLD}_c(i)$ is the gold coreferent antecedents that span i refers to:

$$L_c = \log \prod_{i=1}^{\lambda T} \sum_{\hat{y} \in Y(i) \bigcap \text{GOLD}_c(i)} P(\hat{y})$$

For bridging resolution, the loss is obtained by multiclass cross-entropy:

$$L_b = -\sum_{c=1}^{K_c} \sum_{i=1}^{\lambda T} \sum_{y} b_{i,j,c} \log(s_b(i, y, c))$$

where K_c represents the number of bridging categories, $s_b(i, j, c)$ denotes the prediction of $s_b(i, j)$ 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.

We compute the total loss as $L = L_m + L_{ref}$, where:

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

6 Experiments

In this section, we present experimental results both with and without state change for recipe anaphora resolution. We use a similar configuration to Lee et al. (2018). Specifically, we use the concatenation of 300-dimensional GloVe embeddings (Pennington et al., 2014), 1024-dimensional ELMo word representations (Peters et al., 2018), and 8-dimensional character embeddings that are learned from a character CNN with windows of 3, 4, and 5 characters as the pretrained token embeddings. Each feed-forward neural network consists of two hidden layers with 150 dimensions and rectified linear units (Nair and Hinton, 2010). The gold mentions are separated in coreference and bridging. For joint training, the gold mentions are combined.

We use 10-fold cross-validation to evaluate our model on recipe anaphora resolution. Since end-to-end model performance varies due to random initialization (Lee et al., 2017), we randomly shuffle the dataset 5 times and run cross-validation 3 times for each shuffle. Averaged results are reported.

Table 3 shows our primary results, without state change. For coreference resolution, we provide experimental results on both surface and atom coreference metrics. For bridging resolution, we focus on overall bridging results. Since surface and atom coreference metrics show the same trends in performance, we use surface coreference and overall bridging to compute overall results.

Overall, joint training achieves 26.2% F_1 score for surface coreference and 26.9% F_1 score for bridging, with +1.4% and +0.9% F_1 score absolute improvement over the component-wise models. As such, joint training improves the performance of both tasks. Compared to precision, recall in

Relation	Method	P_A	R_A	F_A	P_R	R_R	F_R
COREF (Surface)	coreference	62.0 ± 1.0	$\mathbf{37.8 \pm 0.8}$	46.1 ± 0.8	33.6 ± 0.9	20.4 ± 0.6	24.8 ± 0.7
	joint_train	$\mathbf{65.2 \pm 0.9}$	37.5 ± 0.9	$\mathbf{46.7 \pm 0.8}$	$\mathbf{36.8 \pm 0.9}$	$\mathbf{21.0 \pm 0.6}$	$\mathbf{26.2 \pm 0.7}$
COREF (Atom)	coreference	62.0 ± 1.0	$\mathbf{37.8 \pm 0.8}$	46.1 ± 0.8	46.8 ± 1.1	26.1 ± 0.7	32.9 ± 0.7
	joint_train	$\mathbf{65.2 \pm 0.9}$	37.5 ± 0.9	$\mathbf{46.7 \pm 0.8}$	$\mathbf{50.4 \pm 1.1}$	$\mathbf{26.7 \pm 0.7}$	$\mathbf{34.4 \pm 0.8}$
Bridging	bridging	56.1 ± 1.2	35.1 ± 0.9	41.7 ± 0.8	36.3 ± 0.9	21.5 ± 0.8	26.0 ± 0.7
	joint_train	$\mathbf{57.7 \pm 1.3}$	$\mathbf{35.5 \pm 0.9}$	$\mathbf{42.7 \pm 0.8}$	$\mathbf{38.0 \pm 0.8}$	$\mathbf{21.9 \pm 0.7}$	$\mathbf{26.9 \pm 0.7}$
Overall	joint_train	62.1 ± 0.7	37.0 ± 0.5	46.0 ± 0.5	37.4 ± 0.7	21.8 ± 0.5	27.1 ± 0.5

Table 3: Anaphora resolution results based on 10-fold cross validation without considering state change. Models were trained over 10,000 epochs, and averaged over 3 runs with 5 different random seeds (a total of 5×3×10 runs). Models are trained for "coreference", "bridging" or "joint_train" (both tasks jointly). "F_A" denotes the F1 score for anaphor prediction, and "F_R" for relation prediction.

anaphor and relation detection is lower, indicating the complexity in anaphoric forms in recipes.

We also experimented with joint coreference resolution and change-of-state classification, and observed similar trends in the results, at reduced performance levels due to the difficulty in additionally predicting state changes (as shown in Appendix A).

Relation	Method	F_A	F_R
COREF (Surface)	coreference	46.1 ± 0.8	24.8 ± 0.7
	- w/ transfer	46.7 ± 0.8	25.3 ± 0.7
	joint_train	46.7 ± 0.8	26.2 ± 0.7
	- w/ transfer	45.3 ± 0.9	26.9 ± 0.7
COREF (Atom)	coreference	46.1 ± 0.8	32.9 ± 0.7
	- w/ transfer	46.7 ± 0.8	33.5 ± 0.8
	joint_train	46.7 ± 0.8	34.4 ± 0.8
	- w/ transfer	45.3 ± 0.9	33.9 ± 0.8
Bridging	bridging	41.7 ± 0.8	26.0 ± 0.7
	- w/ transfer	40.6 ± 0.9	26.7 ± 0.7
	joint_train	42.7 ± 0.8	26.9 ± 0.7
	- w/ transfer	43.4 ± 0.8	27.9 ± 0.7
Overall	joint_train	46.0 ± 0.5	27.1 ± 0.5
	- w/ transfer	45.2 ± 0.6	27.9 ± 0.5

Table 4: Experiments with transfer learning, without considering state change. "F_A" denotes the F1 score for anaphor prediction, and "F_R" for relation prediction.

As discussed in Section 3.4, chemical patents and recipes have similar text structure. Based on the hypothesis that this structural similarity can lead to successful domain transfer, we experiment with transfer learning from the chemical domain to recipes. Specifically, we pretrain the anaphora resolution model on the ChEMU-ref corpus (Fang et al., 2021a,b) with 10,000 epochs, and fine-tune it over the recipe corpus.

Table 4 shows the results with transfer learning, demonstrating consistent improvements over coreference and bridging resolution. Overall, we achieve 27.9% F_1 score for relation prediction under joint training and transfer learning, obtaining +0.8% F_1 score absolute improvement. Incorporating procedural knowledge also improves component-wise models by +0.5% and +0.7% F_1 score (absolute) for surface coreference and bridging, respectively.

We performance error analysis on 5 randomly-selected batches from 10-fold cross-validation based on joint training. There are two primary causes of error. First, the model struggles to capture the semantics of ingredient terms as they are combined with other ingredients. As discussed in Section 3.4, ingredient terms can semantically represent a mixture. E.g. *the biscuits* in Fig. 1 line 6 and *the yellowtail* in Table 5 Ex 1 both represent a mixture of previous ingredients which includes the key ingredient of *biscuits* and *yellowtail*, respectively. The model fails to capture the fact that these mentions incorporate multiple antecedents, and incorrectly analyzes them as COREFERENCE. The second cause of error is in failing to detect state change, mostly in falsely analyzing TRANSFORMED as COREFERENCE, and INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED as INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED.

Errors in coreference resolution occur due to two primary factors: (1) imbalance of coreference and bridging; and (2) entities with different surface expressions. As shown in Table 2, coreference relations are not common in recipes, making it hard for models to capture coreference links. Models also fail to capture the coreference relationship of entities in the face of lexical variation.

In bridging resolution, models also tend to predict anaphoric links as INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED due to its predominance in the annotated data. Furthermore, given that it is a many-to-one relation, models

1	Season the yellowtail fillets with salt and pepper, then dust 1 side only with flour, shaking off any excess. in a medium sized saute pan, heat the olive oil until just nearly smoking and add the yellowtail, flour side down...
2	In a bowl, mash the corned beef as much as you can. Add the tinned tomatoes, onions and curry powder. Mix well until the mixture becomes free of any lump of corned beef. Transfer to a frying pan on a medium heat, cook the mixture for about 10 – 15 minutes until the mixture is heated through...
3	In a ceramic or glass bowl, combine chiles, orange juice, lemon juice, and orange peel. Add the fish and refrigerate for 4 to 6 hours, stirring occasionally until the fish loses all translucency. You may leave in the refrigerator overnight to marinate, if desired. Remove the fish, reserving the juice.
4	...Add the white wine and passion fruit. Over medium heat, reduce by 3/ the liquid in the pan will begin to look thick and bubbly. Remove the pan from the heat and slowly whisk in the butter a little bit at a time, making sure all butter is whisked in before adding more...

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

tend to over-predict INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED relations to mentions which are not associated with the given anaphor. A natural explanation for this is that span-pair predictions are made independent of one another, and there is no way for the model to capture interactions between anaphors. Simultaneously evaluating candidate antecedents might address this issue.

By incorporating procedural knowledge via transfer learning, models achieve better performance. The improvement occurs in two main forms. First, mention detection improves. For example in Table 5 Ex 3, *the juice* and its related anaphoric relations are predicted by models with transfer learning, yet not captured by standard joint training models. Second, detection of lexically-varied coreferent mentions improves. With Ex 4, standard joint training models fails to capture the the COREFERENCE relation between *the butter* and *all butter* due to variation in expression, but this relation is correctly captured by models with transfer learning.

Directions for future work include: (1) joint learning with COREFERENCE and TRANSFORMED relations, which differ only in whether there is a state change or not, such that considering them together may be effective; (2) incorporation of external knowledge, including knowledge about ingredient entities, which may further improve transfer learning; and (3) utilization of transformer based models (Joshi et al., 2020; Xia and Van Durme, 2021), which have been shown to perform well in general-domain coreference settings.

7 Conclusion

In this paper, we have extended earlier work on anaphora resolution over chemical patents to the domain of recipes. We adapted the annotation schema and guidelines for chemical patents, and created a labeled anaphora resolution corpus for recipes. We further defined two tasks for modeling anaphora phenomena in recipes, with and without consideration of state change. Our experiments show the benefit of joint training, and also transfer learning from the chemical domain.

Acknowledgements

This work was done in the framework of the ChEMU project, supported by Australian Research Council Linkage Project project number LP160101469 and Elsevier. A graduate research scholarship was provided by the University of Melbourne Faculty of Engineering and IT to Biaoyan Fang. We would also like to thank Dr. Christian Druckenbrodt, Dr. Saber A. Akhondi, and Dr. Camilo Thorne from Elsevier, as well as our two expert recipe annotators Kate Baldwin and Ayah Tayeh, for their contributions in refining the annotation guidelines.

References

Rahul Agarwal and Kevin Miller. 2011. Information extraction from recipes. *Department of Computer Science, Stanford University-2008*.

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

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.

Devansh Batra, Nirav Diwan, Utkarsh Upadhyay, Jushaan Singh Kalra, Tript Sharma, Aman Kumar Sharma, Dheeraj Khanna, Jaspreet Singh Marwah, Srilakshmi Kalathil, Navjot Singh, Rudraksh Tuwani, and Ganesh Bagler. 2020. RecipeDB: A resource for exploring recipes. *Database*, 2020.

Arthur Brack, Daniel Uwe Müller, Anett Hoppe, and Ralph Ewerth. 2021. Coreference resolution in research papers from multiple domains. In *Proc. of the 43rd European Conference on Information Retrieval*, online.

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.

Zeyu Dai, Hongliang Fei, and Ping Li. 2019. Coreference aware representation learning for neural named entity recognition. In *IJCAI*, pages 4946–4953.

Bhavana Dalvi, Lifu Huang, Niket Tandon, Wen tau Yih, and Peter Clark. 2018. Tracking state changes in procedural text: A challenge dataset and models for process paragraph comprehension. In *NAACL*.

Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021a. ChEMU-ref: A corpus for modeling anaphora resolution in the chemical domain. In *Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume*, pages 1362–1375, Online. Association for Computational Linguistics.

Biaoyan Fang, Christian Druckenbrodt, Saber A. Akhondi, Camilo Thorne, Timothy Baldwin, and Karin Verspoor. 2022. RecipeRef corpus for modeling anaphora resolution from the procedural text of recipes. Mendeley Data.

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

Annemarie Friedrich, Heike Adel, Federico Tomazic, Johannes Hingerl, Renou Benteau, Anika Marusczyk, and Lukas Lange. 2020. The SOFC-exp corpus and neural approaches to information extraction in the materials science domain. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 1255–1268, Online. Association for Computational Linguistics.

Abbas Ghaddar and Phillippe Langlais. 2016. 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.

Jun Harashima, Michiaki Ariga, Kenta Murata, and Masayuki Ioki. 2016. A large-scale recipe and meal data collection as infrastructure for food research. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 2455–2459, Portorož, Slovenia. European Language Resources Association (ELRA).

Jun Harashima and Makoto Hiramatsu. 2020. Cookpad parsed corpus: Linguistic annotations of Japanese recipes. In *Proceedings of the 14th Linguistic Annotation Workshop*, pages 87–92, Barcelona, Spain. Association for Computational Linguistics.

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.

De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Niebles. 2017. Unsupervised visual-linguistic reference resolution in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 2183–2192.

Yiwei Jiang, Klim Zaporojets, Johannes Deleu, Thomas Demeester, and Chris Develder. 2020. Recipe instruction semantics corpus (RISeC): Resolving semantic structure and zero anaphora in recipes. In *Proceedings of the 1st Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics and the 10th International Joint Conference on Natural Language Processing*, pages 821–826, Suzhou, China. Association for Computational Linguistics.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. SpanBERT: Improving pre-training by representing and predicting spans. *Transactions of the Association for Computational Linguistics*, 8:64–77.

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.

Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 982–992, Lisbon, Portugal. Association for Computational Linguistics.

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.

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.

Yi Luan, Luheng He, Mari Ostendorf, and Hannaneh Hajishirzi. 2018. Multi-task identification of entities, relations, and coreference for scientific knowledge graph construction. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3219–3232, Brussels, Belgium. Association for Computational Linguistics.

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.

Javier Marin, Aritro Biswas, Ferda Ofli, Nicholas Hynes, Amaia Salvador, Yusuf Aytar, Ingmar Weber, and Antonio Torralba. 2019. Recipe1m+: A dataset for learning cross-modal embeddings for cooking recipes and food images. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 43(1):187–203.

Ashutosh Modi, Tatjana Anikina, Simon Ostermann, and Manfred Pinkal. 2016. InScript: Narrative texts annotated with script information. In *Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC'16)*, pages 3485–3493, Portorož, Slovenia. European Language Resources Association (ELRA).

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.

Shinsuke Mori, Hirokuni Maeta, Yoko Yamakata, and Tetsuro Sasada. 2014. Flow graph corpus from recipe texts. In *Proceedings of the Ninth International Conference on Language Resources and Evaluation (LREC'14)*, pages 2370–2377, Reykjavik, Iceland. European Language Resources Association (ELRA).

Sheshera Mysore, Zachary Jensen, Edward Kim, Kevin Huang, Haw-Shiuan Chang, Emma Strubell, Jeffrey Flanigan, Andrew McCallum, and Elsa Olivetti. 2019. The materials science procedural text corpus: Annotating materials synthesis procedures with shallow semantic structures. In *Proceedings of the 13th Linguistic Annotation Workshop*, pages 56–64, Florence, Italy. Association for Computational Linguistics.

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.

Taichi Nishimura, Suzushi Tomori, Hayato Hashimoto, Atsushi Hashimoto, Yoko Yamakata, Jun Harashima, Yoshitaka Ushiku, and Shinsuke Mori. 2020. Visual grounding annotation of recipe flow graph. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 4275–4284, Marseille, France. European Language Resources Association.

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.

Gabriel Stanovsky, Noah A. Smith, and Luke Zettlemoyer. 2019. Evaluating gender bias in machine translation. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 1679–1684, Florence, Italy. Association for Computational Linguistics.

Niket Tandon, Bhavana Dalvi Mishra, Joel Grus, Wen tau Yih, Antoine Bosselut, and Peter Clark. 2018. Reasoning about actions and state changes by injecting commonsense knowledge. In *EMNLP*.

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.

Patrick Xia and Benjamin Van Durme. 2021. Moving on from OntoNotes: Coreference resolution model transfer. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5241–5256, Online and Punta Cana, Dominican Republic. Association for Computational Linguistics.

Yoko Yamakata, Shinsuke Mori, and John Carroll. 2020. English recipe flow graph corpus. In *Proceedings of the 12th Language Resources and Evaluation Conference*, pages 5187–5194, Marseille, France. European Language Resources Association.

Juntao Yu and Massimo Poesio. 2020. Multitask learning-based neural bridging reference resolution. In *Proceedings of the 28th International Conference on Computational Linguistics*, pages 3534–3546, Barcelona, Spain (Online). International Committee on Computational Linguistics.

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

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 experimental results.

Table 6 provides anaphora resolution results with state changes based on 10-fold cross validation.

Table 7 provides a full comparison of transfer learning per anaphora relation *with* state change based on 10-fold cross validation.

Table 8 provides a full comparison of transfer learning per anaphora relation *without* state change based on 10-fold cross validation.

Table 9 provides a full comparison of transfer learning for coreference resolution based on 10-fold cross validation, under standard coreference evaluation metrics, i.e. MUC, BCUBED, and CRAFE. Specifically, models are trained with the same parameters (e.g. data partitions, training epochs, etc.) discussed in Section 6 but with a change of coreference evaluation metric, i.e. standard coreference evaluation metrics. We consider the "Ave. F" as the main evaluation metric, computed by averaging F1 scores of MUC, BCUBED, and CRAFE.

Relation	Method	P_A	R_A	F_A	P_R	R_R	F_R
COREF (Surface)	coreference	46.5 ± 2.2	13.3 ± 0.7	19.7 ± 0.9	22.7 ± 2.0	6.2 ± 0.5	9.2 ± 0.7
	joint_train	**48.6 ± 1.9**	**15.3 ± 0.7**	**22.0 ± 0.9**	**28.7 ± 1.7**	**8.6 ± 0.5**	**12.5 ± 0.7**
COREF (Atom)	coreference	46.5 ± 2.2	13.3 ± 0.7	19.7 ± 0.9	27.9 ± 2.1	7.5 ± 0.5	11.2 ± 0.8
	joint_train	**48.6 ± 1.9**	**15.3 ± 0.7**	**22.0 ± 0.9**	**33.5 ± 1.8**	**9.8 ± 0.5**	**14.4 ± 0.7**
Bridging	bridging	51.7 ± 1.0	**25.3 ± 0.6**	**33.2 ± 0.6**	36.3 ± 0.8	**19.4 ± 0.6**	24.5 ± 0.6
	joint_train	**52.6 ± 1.0**	24.6 ± 0.6	32.7 ± 0.7	**37.7 ± 0.8**	19.1 ± 0.6	**24.7 ± 0.6**
TR	bridging	47.0 ± 2.3	**16.6 ± 0.9**	**23.0 ± 1.2**	32.9 ± 1.9	13.2 ± 0.8	17.3 ± 0.9
	joint_train	**52.0 ± 2.3**	16.0 ± 0.9	22.9 ± 1.1	**37.5 ± 2.2**	13.2 ± 0.8	**17.9 ± 1.0**
IWOA	bridging	**5.9 ± 1.6**	**3.3 ± 1.1**	**3.7 ± 1.1**	**3.1 ± 1.1**	**2.3 ± 1.1**	**2.3 ± 1.0**
	joint_train	4.3 ± 1.3	2.4 ± 0.7	2.7 ± 0.7	2.5 ± 1.0	0.9 ± 0.4	1.1 ± 0.4
IWA	bridging	55.2 ± 1.2	**36.8 ± 1.0**	**42.9 ± 0.9**	37.9 ± 0.9	**22.7 ± 0.8**	27.3 ± 0.7
	joint_train	**55.6 ± 1.2**	35.8 ± 1.0	42.3 ± 0.9	**39.4 ± 1.0**	22.4 ± 0.8	**27.5 ± 0.7**
Overall	joint_train	51.6 ± 0.8	21.5 ± 0.4	29.9 ± 0.5	36.3 ± 0.7	17.3 ± 0.5	23.0 ± 0.5

Table 6: Anaphora resolution results based on 10-fold cross validation *with* state change. Models were trained over 10,000 epochs, and averaged over 3 runs with 5 different random seeds (a total of 5×3×10 runs). Models are trained for "coreference", "bridging" or "joint_train" (both tasks jointly). "F_A" denotes the F1 score for anaphor prediction, and "F_R" for relation prediction.

Relation	Method	P_A	R_A	F_A	P_R	R_R	F_R
COREF (Surface)	coreference	**45.6 ± 2.3**	**13.9 ± 0.8**	**20.0 ± 1.0**	**27.9 ± 2.1**	**8.3 ± 0.6**	**11.9 ± 0.8**
	joint_train	43.4 ± 2.3	12.3 ± 0.7	18.1 ± 1.0	24.5 ± 1.9	6.5 ± 0.5	9.7 ± 0.6
COREF (Atom)	coreference	**45.6 ± 2.3**	**13.9 ± 0.8**	**20.0 ± 1.0**	**32.9 ± 2.2**	**9.4 ± 0.6**	**13.7 ± 0.8**
	joint_train	43.4 ± 2.3	12.3 ± 0.7	18.1 ± 1.0	29.1 ± 2.1	7.6 ± 0.5	11.3 ± 0.7
Bridging	bridging	53.4 ± 1.0	24.9 ± 0.5	33.3 ± 0.6	38.9 ± 0.8	**19.8 ± 0.6**	25.7 ± 0.6
	joint_train	**55.2 ± 1.0**	**25.6 ± 0.6**	**34.3 ± 0.6**	**39.6 ± 0.8**	19.7 ± 0.5	**25.8 ± 0.6**
TR	bridging	50.6 ± 2.2	**17.8 ± 0.9**	**24.3 ± 1.0**	**37.8 ± 2.1**	**14.3 ± 0.8**	**18.9 ± 0.9**
	joint_train	**53.8 ± 2.4**	16.5 ± 0.9	23.5 ± 1.2	36.3 ± 2.2	12.9 ± 0.8	17.3 ± 0.9
IWOA	bridging	4.4 ± 1.4	1.9 ± 0.6	2.3 ± 0.7	1.2 ± 0.5	0.5 ± 0.2	0.6 ± 0.2
	joint_train	**5.0 ± 1.5**	**2.9 ± 1.1**	**3.3 ± 1.1**	**2.6 ± 1.1**	**1.9 ± 1.0**	**2.0 ± 1.0**
IWA	bridging	56.9 ± 1.2	35.4 ± 1.0	42.4 ± 0.9	40.5 ± 0.9	23.1 ± 0.7	28.5 ± 0.7
	joint_train	**58.2 ± 1.2**	**37.8 ± 1.0**	**44.4 ± 0.9**	**41.5 ± 0.9**	**23.4 ± 0.7**	**29.0 ± 0.7**
Overall	joint_train	53.2 ± 0.8	21.3 ± 0.4	30.0 ± 0.5	37.9 ± 0.7	17.5 ± 0.4	23.6 ± 0.5

Table 7: Experiments with transfer learning based on 10-fold cross validation *with* state change. Models were trained over 10,000 epochs, and averaged over 3 runs with 5 different random seeds (a total of 5×3×10 runs). Models are trained for "coreference", "bridging" or "joint_train" (both tasks jointly). "F_A" denotes the F1 score for anaphor prediction, and "F_R" for relation prediction.

Relation	Method	P_A	R_A	F_A	P_R	R_R	F_R
COREF (Surface)	coreference	63.3 ± 0.9	$\mathbf{37.8 \pm 0.8}$	$\mathbf{46.7 \pm 0.8}$	34.4 ± 0.9	20.5 ± 0.6	25.3 ± 0.7
	joint_train	$\mathbf{66.4 \pm 1.0}$	35.4 ± 0.9	45.3 ± 0.9	$\mathbf{39.7 \pm 1.0}$	$\mathbf{21.0 \pm 0.6}$	$\mathbf{26.9 \pm 0.7}$
COREF (Atom)	coreference	63.3 ± 0.9	$\mathbf{37.8 \pm 0.8}$	$\mathbf{46.7 \pm 0.8}$	47.8 ± 1.1	$\mathbf{26.3 \pm 0.7}$	33.5 ± 0.8
	joint_train	$\mathbf{66.4 \pm 1.0}$	35.4 ± 0.9	45.3 ± 0.9	$\mathbf{52.2 \pm 1.2}$	25.8 ± 0.7	$\mathbf{33.9 \pm 0.8}$
Bridging	bridging	55.5 ± 1.3	33.1 ± 0.9	40.6 ± 0.9	38.0 ± 1.0	21.5 ± 0.7	26.7 ± 0.7
	joint_train	$\mathbf{58.4 \pm 1.2}$	$\mathbf{35.8 \pm 0.9}$	$\mathbf{43.4 \pm 0.8}$	$\mathbf{40.3 \pm 1.0}$	$\mathbf{22.3 \pm 0.6}$	$\mathbf{27.9 \pm 0.7}$
Overall	joint_train	63.0 ± 0.7	35.8 ± 0.6	45.2 ± 0.6	39.8 ± 0.6	22.0 ± 0.5	27.9 ± 0.5

Table 8: Experiments with transfer learning based on 10-fold cross validation *without* state change. Models were trained over 10,000 epochs, and averaged over 3 runs with 5 different random seeds (total $5 \times 3 \times 10$ runs). Models are trained for "coreference", "bridging" or "joint_train" (both tasks jointly). "F_A" denotes the F1 score for anaphor prediction, and "F_R" for relation prediction.

State	Method	MUC			BCUBED			CRAFE			Ave. F
		P	R	F	P	R	F	P	R	F	
With state	coreference	30.1 ± 2.0	8.8 ± 0.6	12.7 ± 0.8	37.9 ± 1.8	10.8 ± 0.5	15.7 ± 0.7	46.2 ± 1.7	12.1 ± 0.5	18.5 ± 0.7	15.6 ± 0.7
	- w/ transfer	35.1 ± 2.0	11.2 ± 0.6	16.0 ± 0.8	40.8 ± 1.8	12.4 ± 0.6	17.8 ± 0.7	48.3 ± 1.7	12.9 ± 0.5	19.6 ± 0.7	17.8 ± 0.7
	joint_train	30.4 ± 1.7	10.9 ± 0.7	15.3 ± 0.9	37.1 ± 1.6	12.3 ± 0.6	17.4 ± 0.8	43.0 ± 1.6	13.5 ± 0.6	19.9 ± 0.8	17.5 ± 0.8
	- w/ transfer	36.4 ± 2.2	9.5 ± 0.6	14.2 ± 0.8	41.8 ± 2.0	10.5 ± 0.5	15.7 ± 0.7	46.1 ± 1.8	11.4 ± 0.5	17.6 ± 0.7	15.8 ± 0.7
Without state	coreference	50.5 ± 1.1	32.2 ± 0.8	38.7 ± 0.8	49.3 ± 0.9	30.2 ± 0.7	36.5 ± 0.6	54.6 ± 0.8	28.1 ± 0.7	36.5 ± 0.7	37.2 ± 0.7
	- w/ transfer	51.9 ± 1.1	30.3 ± 0.8	37.7 ± 0.8	51.9 ± 1.0	28.4 ± 0.6	35.7 ± 0.6	55.4 ± 0.8	27.6 ± 0.5	36.5 ± 0.5	36.6 ± 0.6
	joint_train	53.4 ± 1.1	32.2 ± 0.9	39.5 ± 0.9	53.6 ± 1.0	30.1 ± 0.8	37.5 ± 0.7	56.2 ± 0.8	29.6 ± 0.7	38.2 ± 0.7	38.4 ± 0.7
	- w/ transfer	54.5 ± 1.1	30.2 ± 0.8	38.2 ± 0.8	55.4 ± 1.1	28.4 ± 0.6	36.6 ± 0.6	57.0 ± 0.8	29.2 ± 0.6	38.1 ± 0.6	37.6 ± 0.7

Table 9: Results based on standard coreference evaluation metrics, i.e. MUC, BCUBED, and CRAFE, based on 10-fold cross validation *without* state change. Models were trained over 10,000 epochs, and averaged over 3 runs with 5 different random seeds (a total of $5 \times 3 \times 10$ runs). Models are trained for "coreference", or "joint_train" (both tasks jointly). "Ave. F" denotes the average F1 score of MUC, BCUBED, and CRAFE.