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

Anonymous ACL submission

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 modelling 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. The corpus is made available at `withheld_for_review`.

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, 2016a; 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, 2016b; Rösiger, 2018a; Hou et al., 2018), such as news or dialogue. Procedural text, such as patents describing chemical synthesis or instruction manuals, has received more limited attention although it is 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.

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

Our contributions in this paper include: (1) generalisation of the anaphora annotation framework from chemical patents for modelling anaphoric phenomena in recipes; (2) creation of a publicly accessible recipe anaphora resolution dataset based on the annotation framework; (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 solve recipe anaphora resolution via transfer learning.

[1] [link withhold for anonymous submission]

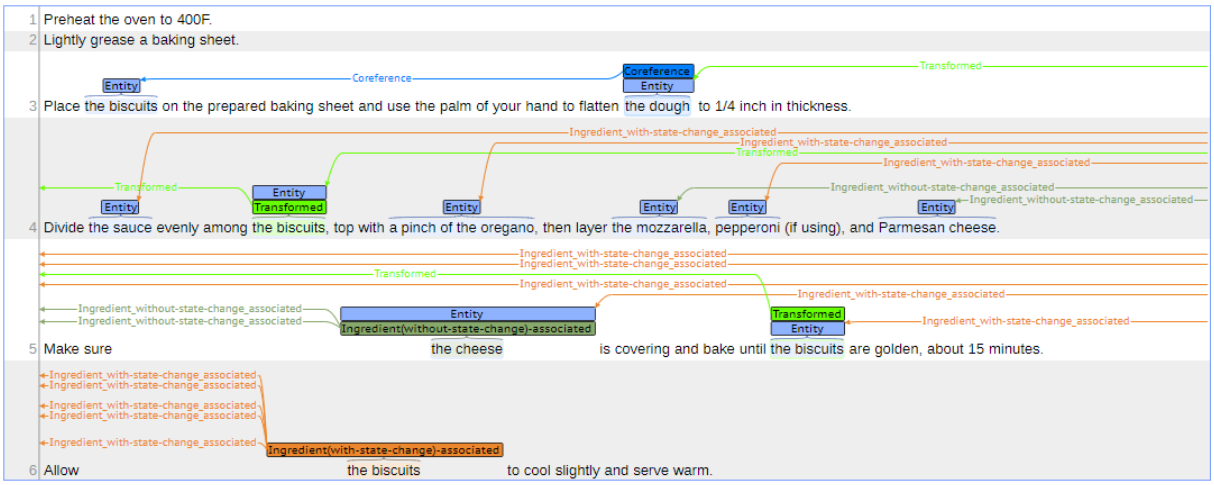

Figure 1: Annotated excerpt of anaphora resolution in the recipes. Different color of links represent different anaphora relation types. Detailed anaphora relation definition can be seen Section 3.3.

2 Related Work

Anaphora relation subsumes two referring types, coreference — expressions in the text that refer to the same entity (Clark and Manning, 2015; Ng, 2017), and bridging — expressions that 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 various domains (Poesio et al., 2008; Pradhan et al., 2012; Ghaddar and Langlais, 2016b; Cohen et al., 2017). A few procedural corpora are annotated for anaphora resolution but most only have coreference annotated (Mysore et al., 2019; Friedrich et al., 2020).

Pradhan et al. (2012) propose the CoNLL 2012 corpus for generic coreference resolution. It consists of three languages, English, Chinese and Arabic, in declarative texts including news and magazine articles. This corpus follows the OntoNotes 5.0 (Weischedel et al., 2013) annotation, modelling coreference in terms of two subtypes: Identity, where the anaphoric references and referents are identical, and Appositive, where a noun phrase is modified by an intermediately-adjacent noun phrase. It models coreference as a clustering task and the direction of relations is not preserved. Following the same annotation framework largely, the WikiCoref corpus (Ghaddar and Langlais, 2016b) annotates Wikipedia texts.

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, the coreference is limited. CRAFT-ST 2019 (Cohen et al., 2017) annotates 97 full biomedical articles for coreference resolution based on the OntoNotes 5.0 annotation framework with minor adaptations. Compared to the BioNLP 2011 corpus, it contains a wider range of annotations and is not limited to only abstracts. SCIERC (Luan et al., 2018) contains 500 abstracts from scientific articles. They annotate coreference of any two expressions that point to the same entity.

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: *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 *lexical*, describing 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 few referential. 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.

There are a few domain-specific anaphora corpora 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, they model five types of anaphora relationships, i.e. Coreference, Transfers, Reaction-associated, Work-up and Contained. Friedrich et al. (2020) propose the SOFC-Exp corpus based on 45 material sciences articles for the information extraction task. As this corpus mainly focuses on named entity extraction and relation extraction, coreference is presented as a supplemented annotation based on the notion of coindexation between a common noun or a pronoun and a more specific mention appears earlier in the text. Mysore et al. (2019) work on 230 synthesis procedures and capture coreference within text in parenthesis, coreferent abbreviation, etc. The InScript corpus (Modi et al., 2016) consists of 1,000 stories from 10 different scenarios and annotates coreference for noun phrases.

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 models linguistic 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 processing information (e.g. action complement, food complement, etc.). Although interactions among ingredients can be derived via action nodes, this approach doesn't sufficiently capture anaphoric 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.

Most research handles coreference and bridging separately due to limited data availability. For coreference resolution, span ranking models (Lee et al., 2017, 2018) have become the benchmark method over mention ranking models (Clark and Manning, 2015, 2016a,b; Wiseman et al., 2015, 2016). Various span ranking variants have proposed (Zhang et al., 2018; Grobol, 2019; Kantor and Globerson, 2019) and achieved strong performance. With the increasing amount 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 to improve 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) modelled 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 the end-to-end coreference model (Lee et al., 2017). Similar to coreference, they modelled bridging as a clustering task. They achieved great improvement over the bridging task. However, the impact on the coreference task is not clear. Fang et al. (2021a) adopted the same end-to-end framework for joint training anaphora resolution. They modelled bridging as a mention pair classification task and showed improvement on both subtasks.

3 Annotation Scheme

In this section, we describe our adopted annotation scheme for recipe anaphora annotation. The complete annotation guideline is available at [Link withhold for anonymous submission].

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 recipes for the corpus, detailing the steps for preparing the recipe.

3.2 Mention Types

As our goal is to capture anaphoric phenomena in recipes, we focus on ingredient-related expressions. Verbs (e.g. *bake*, *chop*, etc.) are not annotated. In line with previous work (Pradhan et al., 2012; Cohen et al., 2017; Fang et al., 2021a; Ghaddar and Langlais, 2016b), we leave out singleton mentions, i.e. mentions that are not involved in anaphora rela-

tions (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*, *garlic powder*, etc.

Referring Expressions We consider referring expressions to be pronouns (e.g. *it*, *they*, etc.) and generic phrases (e.g. *soup*, *the pastry mixture*, etc.) used to represent ingredients previously introduced.

We adopt several assumptions for mentions:

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

- **Numbers**: In some cases, individual number expressions can be used to imply the ingredients and are considered as mentions. For example, *1* in "Beat eggs, 1 at a time", *three* in "Combine together to make a sandwich. Repeat to make three".

3.3 Relation Types

One of the core components in procedural comprehension is understanding entities state (Dalvi et al., 2018; Tandon et al., 2018). Recipes contain rich information about the change of ingredients state. 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 interaction relations among ingredients benefits in understanding ingredients (i.e. where is the ingredient from) and detailing the relation types with states gives a deeper understanding of recipes (i.e. how to get the ingredient).

There are two basic types of anaphora: coreference and bridging. In recipes, we define bridging as three subtypes of referring relations based on the state of entities. The overall schema of anaphora relations in recipes is shown in Fig 2.

In anaphora resolution, an *antecedent* is a linguistic expression that provides the interpretation

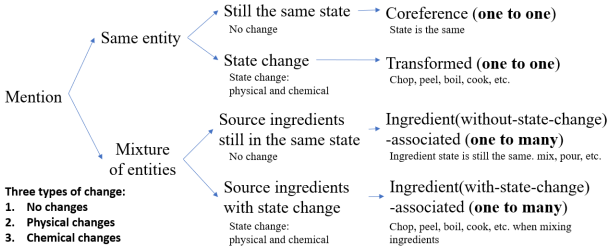

Figure 2: Overall schema for anaphora relations in recipes.

for a second expression, *anaphor*, which cannot be interpreted in isolation or only 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, and the direction of links is preserved.

3.3.1 Coreference

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 the action applied to the entity. To distinguish this subtle information, we consider mentions are coreferent when they point to the same entity and there is no state change, such as a physical or chemical change.

Also, the entity can be repeated in text. To eliminate ambiguity in linking coreferent antecedents, the closet antecedent is linked for a given anaphor.

3.3.2 Bridging

As discussed in Section 3.3.1, we aim to preserve the state change information of entities in procedural text. In the case of recipes, we define three types of bridging relation based on the entity state.

TRANSFORMED A one-to-one anaphoric link for a set of ingredients that is meaning-wise the same but has undergone physical/chemical change (e.g. *peeling*, *baking*, *boiling*, etc.). For example, in Fig 1, *the biscuits* in line 4 and 5 are linked 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 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*

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.

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 and its source ingredients which have undergone a state change. As an example, *the biscuits* in Fig 1 line 6 is a combination of previous source ingredients (i.e. *the sauce*, *a pinch of the oregano*, *pepperoni*, *the cheese* and *the biscuits*) via baking. They are linked as INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED as *bake* changes the state of the previous ingredients.

3.4 Comparison with Chemical Patents

As shown in Table 1, chemical patents and recipes have commonalities. They use similar language to describe the application of processes (e.g. combination, removal, etc.) to source entities to obtain new entities, making it feasible to generalize 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. Such relation types cannot be directly applied to the general domain.

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

- **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 previous 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.

- **Variance in Instruction Descriptions**: Although chemical patents and recipes have similar structures, instruction descriptions in recipes are more variable. In chemical patents, processed entities are mostly directly used in the immediately following process after a mention. However, processed entities in recipes can be mentioned far later in text.

- **Hierarchical Structure in Recipe Relation Types**: Anaphora relation types in recipes are defined in a hierarchy (as shown in Fig 2). A simplified version of recipe anaphora resolution task, i.e. without considering state change, can be easily derived. In chemical patents, there is no clear way simplifying the scheme while presenting anaphoric relations.

4 Task definition

Following the definition in Fang et al. (2021a), anaphora resolution is modelled as a two-step task, mention detection and anaphora relation detection.

As anaphora relation types in recipes are defined in a hierarchy, we can derive a simplified version of recipe anaphora resolution task by removing state changes. As such, COREFERENCE and TRANSFORMED can be merged without considering state changes and similarly for INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED and INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED relationships. We evaluate recipe anaphora resolution both with and without state change.

	RecipeRef	ChEMU-ref
Excerpts	80	1,125
Sentences	999	5,768
Tokens/Sentences	12.6	27.6
Mentions	1,408	17,023
Mentions/Excerpts	17.6	15.1
Coref.	229 / 415	3,243
Coref./Excerpts	2.9 / 5.2	2.9
Bridging*	1,104 / 918	12,796
Bridging*/Excerpts	13.8 / 11.5	11.4
TR	186 / -	-
IwoA	91 / 918	-
IwA	827 / -	-

Table 2: Corpus annotation statistics. For ChEMU-ref corpus, we include its training and development set. "Coref.", "TR", "IwoA" and "IwA" denote COREFERENCE, TRANSFORMED, INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED and INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED, respectively. "/" in relation categories shows the separation in with and without state change scenarios. "Bridging*" is the total number of bridging relations across all subtypes. Bridging subtypes are different in ChEMU-ref corpus, hence we calculate the total number of bridging relations.

For evaluation, we use precision, recall and F1. Although our recipe corpus models coreference as a one-to-one relation and it is transitive, we follow the coreference evaluation of the ChEMU-ref corpus and do not use traditional coreference evaluation metrics (Luo, 2005; Recasens and Hovy, 2011; Moosavi and Strube, 2016). Surface coreference, where a coreferent anaphor links to the closest antecedent, and atom coreference, where a coreferent anaphor links to a correct antecedent, are applied to evaluate coreference resolution (Kim et al., 2012).

For manual annotation, we use the Brat rapid annotation tool.[2] To achieve high quality, we went through 8 rounds of annotation training and refinement of the anaphora annotation with two annotators experienced with the domain. In each round of training, they independently annotated 10 recipes (different for each round of annotation) and met afterwards to compare annotation results. Further refinement of annotation guidelines were made based on the discussion.

After annotation training, we reached a high inner annotator agreement (IAA) between annotators. Krippendorff's α score, F1 score at mention level and relation level are 0.85, 0.88 and 0.67, respectively. As a comparison, it was 0.45, 0.51 and 0.29 at the beginning, respectively. The individual annotation is in progress.

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

5 Methodology

To investigate the benefit of utilizing transfer learning from chemical domain, we follow the configuration of Fang et al. (2021a), modelling bridging as a classification task and adopting the benchmark end-to-end neural coreference model (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:

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

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

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

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 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:

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

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

For mention loss, we unitize 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:

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

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 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 pretrianed 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 show our primary results without state change. For coreference resolution, we show 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 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 predicting state changes (as shown in Appendix A).

As discussed in Section 3.4, chemical patents and recipes share similar text structures. We argue that the structure information can be beneficial for the anaphora resolution task. We hence experiment with utilizing transfer learning from 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 with the recipe corpus.

Table 4 shows results with transfer learning, demonstrating consistent improvement 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 se-

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 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 (total 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	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.

lected batches from 10-fold cross-validation based on joint training models. Overall, models suffer from (1) semantic understanding of ingredient terms. As we discussed in section 3.4, ingredient terms can semantically represent a mixture, e.g. *the biscuits* in Fig 1 line 6 represents a mixture of previous ingredients. Models cannot tell the subtle differences and incorrectly link those ingredient terms as COREFERENCE. (2) detection of state change. Models fail to capture the state transition of entities, mostly falsely inferring TRANSFORMED as COREFERENCE and inferring INGREDIENT(WITHOUT-STATE-CHANGE)-ASSOCIATED as INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED.

Errors in coreference resolution occur also due to (1) imbalance of coreference and bridging and (2) entities with different expressions. As shown in Table 2, coreference relations are not common in recipe anaphora, making it harder for models to capture coreference links. Models also fail to capture the coreference relationship of entities in the face of variations in expression.

In bridging resolution, models also tend to predict anaphoric links as INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED due to its domination in recipe anaphora relations. Furthermore, within the INGREDIENT(WITH-STATE-CHANGE)-ASSOCIATED relationship, models over-predict the relations for a given anaphor. One of the possible reasons is the individual span-pair prediction, which makes it hard to capture the interactions within anaphors. Simultaneously evaluating candidate antecedents might address this issue.

By incorporating procedural knowledge via transfer learning, models achieve better performance. However, models suffer more severely from false negatives due to the difference in the annotation scheme, as discussed in Section 3.4.

Future directions include (1) Joint learning with COREFERENCE and TRANSFORMED relations. These only differ in whether or not state change is considered; considering them together may be effective. (2) Incorporation of external knowledge including world knowledge about ingredient entities; this may further improve transfer learning.

7 Conclusion

We investigate the textual properties in chemical patents and recipes and generalized the annotation guideline for chemical patents to recipes. We create a publicly available recipe anaphora resolution corpus based on the adopted annotation scheme. We further define two tasks for modelling anaphoric phenomena in recipes, with and without state change. Our experiment shows the benefit of utilizing joint training setting and transfer learning from chemical domain.

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, et al. 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. *arXiv preprint arXiv:2101.00884*.

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. *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, 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 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.

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.

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. *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. *arXiv preprint arXiv:2104.08457*.

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. 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.

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 5 provides anaphora resolution results with state changes on 10 fold cross validation.

Table 6 provides a full comparison of transfer learning per anaphora relation with state change on 10 fold cross validation.

Table 7 provides a full comparison of transfer learning per anaphora relation without state change on 10 fold cross validation.

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 5: Anaphora resolution results on 10 fold cross validation with considering state change. Models were trained over 10,000 epochs, and averaged over 3 runs with 5 different random seeds (total 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 6: Experiments with transfer learning on 10 fold cross validation with considering state change. Models were trained over 10,000 epochs, and averaged over 3 runs with 5 different random seeds (total 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 7: Experiments with transfer learning 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 (total 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.