



HAL
open science

Chop and change: Anaphora resolution in instructional cooking videos

Cennet Oguz, Ivana Kruijff-Korbayová, Pascal Denis, Emmanuel Vincent,
Josef van Genabith

► To cite this version:

Cennet Oguz, Ivana Kruijff-Korbayová, Pascal Denis, Emmanuel Vincent, Josef van Genabith. Chop and change: Anaphora resolution in instructional cooking videos. Findings of ACL-IJCNLP 2022 - 2nd Conference of the Asia-Pacific Chapter of the Association for Computational Linguistics - 12th International Joint Conference on Natural Language Processing, Nov 2022, Taipeh, Taiwan. hal-03807530

HAL Id: hal-03807530

<https://inria.hal.science/hal-03807530>

Submitted on 10 Oct 2022

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

Chop and Change: Anaphora Resolution in Instructional Cooking Videos

Cennet Oguz¹, Ivana Kruijff-Korbayova¹, Pascal Denis²,
Emmanuel Vincent³ and Josef van Genabith¹

¹German Research Center for Artificial Intelligence (DFKI), Saarland Informatics

²Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRISAL, F-59000 Lille, France

³Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy, France

{cennet.oguz, ivana.kruijff, josef.van_genabith}@dfki.de

{pascal.denis, emmanuel.vincent}@inria.fr

Abstract

Linguistic ambiguities arising from changes in entities in action flows are a key challenge in instructional cooking videos. In particular, temporally evolving entities present rich and to date understudied challenges for anaphora resolution. For example “oil” mixed with “salt” is later referred to as a “mixture”. In this paper we propose novel annotation guidelines to annotate recipes for the anaphora resolution task, reflecting change in entities. Moreover, we present experimental results for end-to-end multimodal anaphora resolution with the new annotation scheme and propose the use of temporal features for performance improvement.

1 Introduction

Anaphora resolution is the task of identifying the antecedent of an anaphor, i.e., find a language expression that a given entity refers to. For example, in the sentence *take a potato and wash it*, the pronoun *it* is an anaphor that refers to the antecedent *a potato*. This is a challenging NLP task which has been attracting much attention (Poesio et al., 2018; Fang et al., 2021, 2022). Different types of anaphoric relations have been identified and described in the scientific literature, e.g., identity (Poesio and Artstein, 2008), near-identity (Recasens et al., 2011; Hovy et al., 2013), and bridging (Asher and Lascarides, 1998).

Recipes provide a rich source for referring expressions (Kiddon et al., 2015) of transformed entities, and offer a challenge for anaphora resolution tasks. Fang et al. (2022) use written recipes with anaphora annotations to trace the temporal change of entities. While the ingredients undergo physical or chemical change in the action flow, they can be still referred to in the same way. For example, an *egg* before and after it is boiled can be referred to with the same noun *egg*. Compared to text recipes, instructional cooking videos raise additional challenges for anaphora resolution owing to

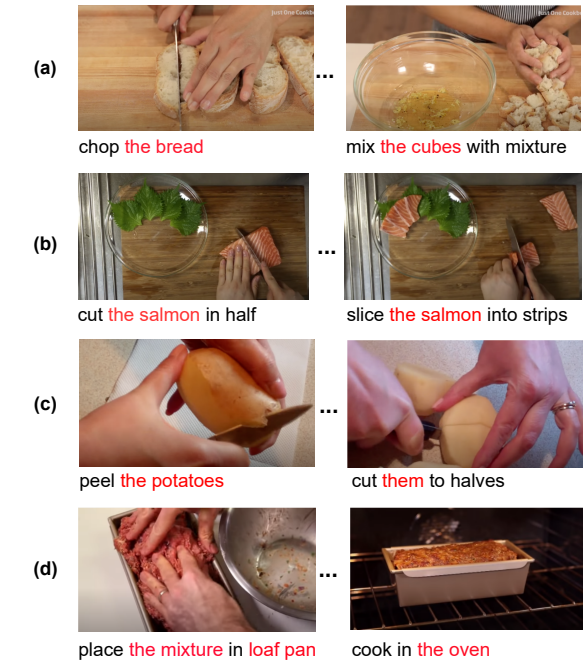


Figure 1: Examples from the YouCookII dataset showing the effect of the temporal changes on the entities and the referring expressions. Each row displays a different use of expressions and entities.

their intrinsic multimodality (Huang et al., 2016). Krishnaswamy and Pustejovsky (2019) point to various “channels of information” in the transmission of each modality. A “shared reference of entities” is introduced when two modalities refer to the same description (Krishnaswamy and Pustejovsky, 2020). As presented in cooking instructions of videos when two modalities refer to the same entity, the use of a referring expression is affected by both modalities. For example, *the cubes* is used in Figure 1a to denote the bread pieces in the text modality because the instruction *chop the bread* shaped them into cubes in the video modality. The choice of referring expressions might also differ with respect to the changes of the entities. In Figure 1b the same nominal phrase refers to a different

057 object (the whole salmon piece; and then one of the
058 halves) whereas in Figure 1c a coreferential pro-
059 noun is used although the object has changed. Fig-
060 ure 1c is in fact the most well-behaved in terms of
061 keeping the language expressions consistent across
062 instructions and with the entities being referred to.
063 Figure 1d shows the use of null arguments: the sec-
064 ond instruction *cook in the oven* does not explicitly
065 mention what to cook, whereas the image of the
066 instruction displays it.

067 The main contributions of this paper are as
068 follows: (i) We propose an anaphora annotation
069 scheme for instructional cooking videos that allows
070 us to address linguistic ambiguities in anaphora res-
071 olution. In particular, we define different types of
072 anaphoric relations to keep track of spatio-temporal
073 changes of entities. We also provide a clear defi-
074 nition of “identity of reference” and specify cate-
075 gories that make an essential change resulting in a
076 different entity. (ii) We annotate the YouCookII
077 dataset (Zhou et al., 2018b,a) according to our
078 scheme and make it publicly available.¹ (iii) Null
079 anaphors, e.g., *mix in the bowl*, are included in
080 the annotation thanks to cooking videos that offer
081 the precise visual observation of null anaphors
082 to annotators. (iv) We provide a baseline multi-
083 modal anaphora resolution model for this dataset.
084 In particular, we adapt an end-to-end (Lee et al.,
085 2017) coreference model for the anaphora resolu-
086 tion task. (v) We offer a novel method to improve
087 anaphora resolution models for instructional lan-
088 guage by leveraging temporal features capturing
089 temporal order of instructions instead of using the
090 token distance as Lee et al. (2017) and Yu and Poe-
091 sio (2020).

092 2 Related Work

093 **Reference Resolution** The reference resolution
094 task addresses the linguistic ambiguities in state
095 changes of entity mentions by linking the entities
096 to their corresponding instructions (Kiddon et al.,
097 2015; Huang et al., 2016, 2018), e.g., *the mashed*
098 *potato* and *the fork* refer to the instruction *mash the*
099 *potatoes with a fork*. We depart from this type of
100 approaches, as they rely on unsound ontological as-
101 sumptions (actions/events and entities are different
102 objects) and they introduce unnecessary semantic
103 ambiguities (by linking different entity mentions to
104 the same instruction).

¹[https://github.com/OguzCennet/
Recipe-Anaphora-Resolution](https://github.com/OguzCennet/Recipe-Anaphora-Resolution)

**Anaphoric Relations: identity, near-identity, as-
sociation.** Anaphoras mainly come in two forms:
coreference and *bridging*. Coreference is defined
as language expressions referring to the same entity
(Weischedel et al., 2012), whereas bridging is an
anaphoric phenomenon based on a non-identical
associated antecedent via lexical-semantic, frame-
based, or encyclopedic relations (Asher and Las-
carides, 1998). A coreferring anaphor and its an-
tecedent in a text refer to the same entity (identity
relation), e.g., *a black Mercedes* and *the car*, while
in bridging, an anaphor and its antecedent refer to
different entities (non-identity relation), e.g., *the*
car and *the engine* in the utterance *I saw [a black*
Mercedes] parked outside the restaurant. [The car]
belonged to Bill. [The engine] was still running.
(Poesio and Artstein, 2008).

As Rösiger et al. (2018) point out, bridging studies
so far employ various methods to describe bridging
dissimilar to the coreference definition. Neverthe-
less, both the concept of sameness in the corefer-
ence definition and the bridging associations neglect
the changes referents may undergo. Therefore, the
concept of *near-identity* was introduced by Recasens
et al. (2010, 2012) as a middle ground between
coreference and bridging. It addresses spatio-temporal
changes of entities, e.g., the entity *Postville* in the
text: *On homecoming night [Postville] feels like*
Hometown, . . . it’s become a miniature Ellis Island
. . . For those who prefer [the old Postville], Mayor
John Hyman has a simple . . . This sample exem-
plifies the referential ambiguity, arising from two
language expressions referring to “almost” the same
entity, i.e., *Postville* and *the old Postville* (Recasens
et al., 2010). Rösiger et al. (2018) and Poesio et al.
(2018) claim that the introduction of the additional
near-identity category in between coreference and
bridging introduces more uncertainty. Nevertheless,
we consider the *near-identity* relationship suitable
because spatio-temporal changes are essential in
recipes and the information they convey describes the
visual content.

Coreference and Bridging Annotations. Corefer-
ence is a well studied and clearly defined concept
with some noticeable exceptions. In recent years
several annotated corpora with different corefer-
ence guidelines have been released. OntoNotes
v5.0 (Weischedel et al., 2012) exclusively focus
on coreference using a schema similar to CoNLL-
2012 (Pradhan et al., 2012) and WikiCoref (Ghad-

dar and Langlais, 2016) with two different relations: one is identity, a symmetrical and transitive relation, and the other appositive for adjacent noun phrases. The extraction of the mentions and the use of prepositions in mentions are crucial questions for coreference annotation (Rösiger et al., 2018; Poesio et al., 2018). There are many extant hypotheses explaining how bridging relations function with different annotation schemes for bridging (Hou et al., 2018). The ARRAU corpus (Poesio et al., 2018) consists of general language annotated with bridging relations of noun phrases (such as *set membership*, *subset*, *possession* and *unrestricted*.) Markert et al. (2012) present ISnotes derived from OntoNotes with unrestricted bridging relations in addition to OntoNotes coreferences. The BASHI corpus (Rösiger, 2018) is based on OntoNotes content and the bridging relations in the BASHI corpus restrict the bridging anaphors to be truly anaphoric, i.e., not interpretable without an antecedent.

All aforementioned annotation studies focus solely on the anaphoric relation between two discourse entities and neglect the change of entities over time. Instructional language raises a novel question in anaphora resolution: the definition of anaphoric relations based on the change of language with entities that undergo change. Therefore, RecipeRef (Fang et al., 2022) considers the state changes for preparing the annotation guideline for recipe text based on the ChEMU-Ref (Fang et al., 2021) anaphora annotation on chemistry patent documents. RecipeRef annotation was applied to the RecipeDB data (Batra et al., 2020) that was aggregated from recipe websites and each recipe was divided into two parts, the ingredients section, and the cooking instructions. The cooking instructions of RecipeDB contains only textual instructions without any visual content. The state changes are addressed in RecipeRef as a subtype of bridging relation, even though bridging is clearly defined as an associative relation in the literature (Clark, 1975; Asher and Lascarides, 1998; Poesio and Artstein, 2008; Poesio et al., 2018). Besides, null anaphors are not included in the annotation of RecipeRef, despite their frequent use in recipes.

Several important questions remain open regarding anaphora resolution, and RecipeRef annotation, including: (1) interpretation of the state changes of entities over time; (2) addressing the referring expression in anaphora resolution with data that has different modalities; (3) obtaining the sequence

	Train	Test
Coreference	891	330
Hyponymy	47	10
Near-Identity	699	217
Bridging	602	217
Produce	507	182
Reduce	40	22
Set-member	44	9
Part-of	11	4
Instruction	2,829	984
Token	8,754	2,966
Recipe	264	89
Entity	5,669	1,927
Null Entity	465	168
Pronoun Entity	206	61

Table 1: Statistics of annotated data with the number of annotated samples with anaphoric relations.

of state changes by annotating the null entities in recipes; (4) the judgement of anaphoric relations of state changes and different semantic relations such as identity, non-identity, near-identity, and association.

3 Corpus

We use the YouCookII dataset (Zhou et al., 2018a) that includes manually provided descriptions (i.e., instructions) of actions in the cooking videos. The dataset contains 2,000 unconstrained instructional videos from 89 cooking recipes. The videos provide a visual input of the corresponding objects to observe the changes clearly. To obtain a variety of ingredients and their state changes, we choose at least three random samples for each the 89 cooking recipes for the training set and one sample for the test set. There is no intersection between training and test recipe samples. In total, we have 264 training documents and 89 test documents as shown in Table 1.

Recipe A recipe is text containing a list of cooking instructions with a list of ingredients, see Figure 2. Here, we use the YouCookII annotation, all instructions for each video are manually annotated with temporal boundaries and described by imperative English sentences. Since the video inputs show the entities and actions clearly, the use of referring expressions and null entities is very common contrary to textual recipes.

Instruction. Each video recipe contains 3 to 15 instructions. Each instruction is a temporally-aligned imperative sentence that is described according to the corresponding action on the video by human annotators. The instructions are not uttered by the instructor of the video but annotated by the human annotator from a third-person viewpoint while watching the video. Each instruction defines an action, i.e., a predicate, applied to a set of objects, i.e., entities. Video segments provide the visual status of the spatio-temporal changes for the mentioned entities for each instruction. Unlike other common types of texts, cooking instructions focus on processes and entities undergoing change during the process. So, the corresponding videos in the YouCookII dataset enable us to comprehend the use of referring expressions of entities for each change.

4 Annotation Categories and Guidelines

In this section, we explain our strategy of mention selection and the use of our annotation schema on the YouCookII data.

4.1 Mention Selection

In our work, we segment multiple-action instructions, e.g., *put the chickpeas into the processor and blend all the ingredients*, into single-action instructions *put the chickpeas into the processor* and *blend all the ingredients* while preserving the order of actions. Each recipe instruction contains one predicate and 0 to 8 entities. Null arguments and ellipses are extremely common in recipes (Kiddon et al., 2015; Huang et al., 2016), since some objects are not verbally expressed, but deduced from the context of the remaining elements or videos. For example *stir for 5 minutes* does not explicitly mention the entity to be stirred. Nominal phrases with (in)definite noun phrases and pronouns are also used to mention the objects of recipes as in the following instruction: *coat the pork in the marinade* and *place it in the oven*. Therefore, we consider null arguments (i.e., null anaphors) and nominal phrases to define mentions. Contrary to ONTONOTES (Weischedel et al., 2012), we include expressions that do not refer to any other mention as singletons in the annotation.

4.2 Anaphoric Relations and Entity Change

In this section, we explain how we define anaphoric relations occurring in the recipes with state changes

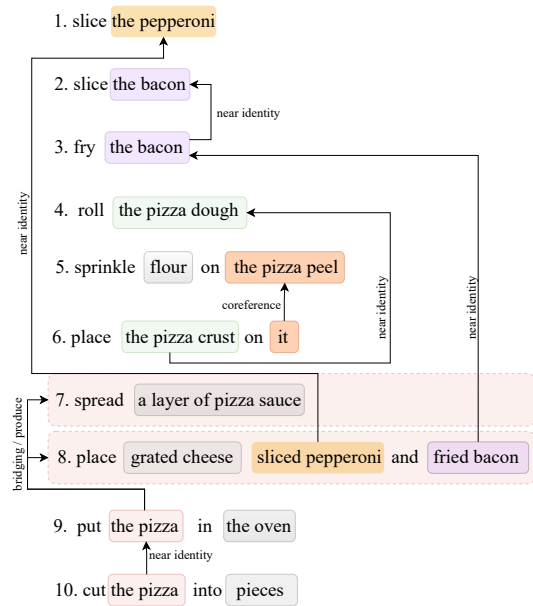


Figure 2: Example of annotation of a recipe from the YouCookII dataset named “stone baked pizza”. The start point of each arrow denotes the anaphor and the end point the corresponding antecedent. The antecedent and anaphor pairs are highlighted in the same color. Grey boxes represent new entities (e.g., singletons) without antecedent.

of entities, see Figure 2. It is worth noting that the recipe videos are exploited to judge the “sameness” of entities after an action (e.g., wash, cut, etc.) was applied. Thus, the visual features from cooking videos clarify the state change of entities in the instructions and our annotation does not rely only on the mental image of entities based on text only settings as in other coreference datasets (Weischedel et al., 2012; Pradhan et al., 2012) and anaphora datasets (Roesiger, 2016; Poesio and Artstein, 2008; Fang et al., 2021, 2022).

4.2.1 Coreference

The anaphor and the antecedent are identical and point to the same entity. Some actions such as washing or transferring the result to another container preserve the properties of the entity involved. For example, a tomato is the same tomato after washing, or a piece of meat is the same amount of meat after putting it in a pan.

4.2.2 Hyponymy

The hyponymy relation was considered as bridging by Poesio and Vieira (1998), however Baumann and Riester (2012) use the term not as context-dependent but as “lexical accessibility” to define the hyponymy relation between words as corefer-

ence, as Rösiger et al. (2018). For example *the herb* refers to the entities *mint and parsley* in the instruction *Wash mint and parsley*. Here again the anaphor may refer to a group of entities as the corresponding antecedent.

4.2.3 Near-Identity

Some actions alter either the physical or chemical properties of the entities involved. For instance, boiling a potato or an egg changes their chemical properties whereas cutting a potato or an egg changes their physical properties. Here, anaphor and antecedent entities are neither identical nor associated, they are partially the same entity sharing many crucial commonalities, but differing in at least one crucial dimension. For this type of anaphoric relation, Recasens et al. (2010) propose the near-identity relation to describe the spatio-temporal changes of the entities as a middle ground between coreference and bridging. Even though Rösiger et al. (2018) claim that additional categories between coreference and bridging introduce further uncertainty which makes the annotation process more arduous, we consider the near identity relationship more suitable because spatio-temporal changes are essential in recipes and the information they convey describes the visual content. Therefore, if they are not the same entity, the antecedent is not reduced to its parts for the anaphor, and the antecedent is not mixed with other entities to produce a new entity for the anaphor, then we define such entities as near-identical. For example, an egg or a potato are accepted as near-identical entities before and after boiling.

4.2.4 Bridging

In bridging, the antecedent is related and not identical; in contrast to coreference the anaphor is also not interchangeable with the given antecedent. As mentioned in Section 2, various phenomena are identified as bridging, resulting in diverse guidelines for bridging annotations. In accordance with the variety of associations, we assign different anaphora relations in our annotation schema.

PRODUCED: We define PRODUCED as the relationship when the anaphor refers to an antecedent producing the anaphor. The antecedent is always an instruction with predicates and given ingredients. Here, the anaphor may refer to a group of instructions as the corresponding antecedent. For example, *the dough* is produced by the instruction

mix water and flour or *dressing* is produced by the instruction *mix yogurt and pepper*.

REDUCED: We define REDUCED as the bridging relation linking an entity. The anaphor might be a number expression (e.g., *to the whole entity*), an indefinite pronoun (*some*), or an indefinite noun phrase (e.g., *one piece*). We use REDUCED in cases when the anaphor means a part of the corresponding antecedent, provided no mereological relation exists. For example *one slice* is reduced from a bread by the instruction *slice the bread into pieces*.

SET-MEMBER: In a recipe, SET-MEMBER refers to a relation between a group of entities and its definite subset. In other words, this relation defines a bridge from a subset or element to the whole collection. For example, *cucumber, tomato, and lettuce* is an antecedent of the anaphor *ingredients* in *cut the ingredients*.

PART-OF: The antecedent may associate in a mereological relationship with the anaphor, and cannot be captured well by pre-defined lexical relations. For example, the antecedent *lemon* in the instruction *cut the lemon* relates to the anaphor *seeds* in *take the seeds out*.

4.3 Inter-annotator Agreement

50 randomly selected recipes have been annotated by two Computational Linguists, a PhD candidate and a final year Master student in Computational Linguistics. Five rounds of annotation training were completed prior to beginning the official annotation. In each round, the two annotators individually annotated the same 5 recipes (different across each round of annotation), and compared their annotations; annotation guidelines were then refined based on discussion. Finally, We achieved a high inner-annotator agreement of Krippendorff’s $\alpha = 0.99$ for the creation of a new entity and reference, $\alpha = 0.95$ for the selection of the antecedent and $\alpha = 0.93$ for selection of anaphoric relations.

5 Method

In this section, we present our end-to-end multimodal anaphora resolution model. Figure 3 shows our joint neural model similar to Yu and Poesio (2020) and Fang et al. (2021), adapted from Lee et al. (2017). We extend the model with novel temporal features, see Section 5.3.

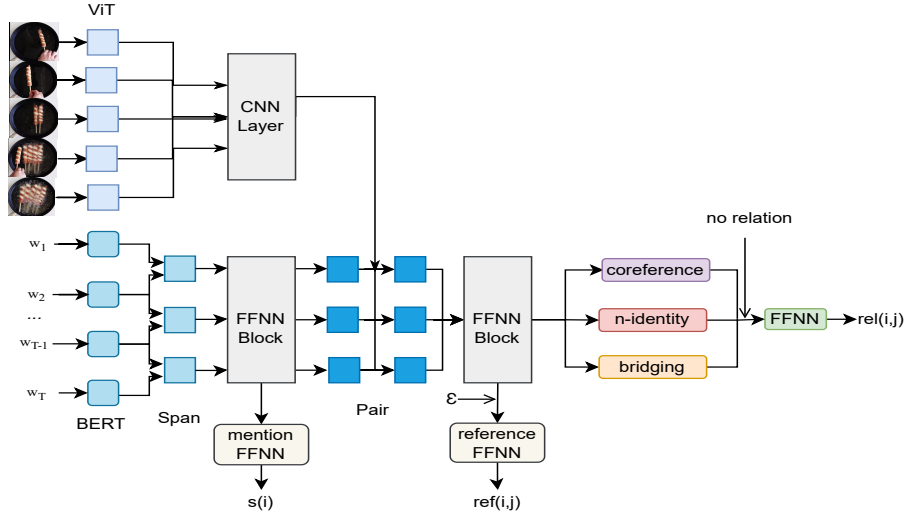


Figure 3: Proposed anaphora resolution architecture. The CNN Layer is a convolutional layer with five input channels (one per frame). The FFNN Block refers to a layer block with FFNN+ReLU+Dropout, w_t indicates the t -th word of Recipe R . ViT is a Transformer-based model to represent the features of the video inputs.

5.1 Task

In linguistics, the term Anaphora Resolution refers to the method of identifying the antecedent for an anaphor. To achieve anaphora resolution on cooking instructions, we propose two different sub-tasks: recognizing mentions, and finding the anaphor-antecedent pairs. Additionally, relation classification is used to find the relation between each anaphor and its antecedent.

We adopt the following notations. Each recipe R consists of T tokens w_1, \dots, w_T and $n \geq 1$ instructions a_i such that $R = a_1, \dots, a_n$. Each instruction $a_i = (p_i, e_\ell)$, e.g., *pour olive oil on the Italian bread cubes*, contains one action predicate p_i and an entity list e_ℓ . The entity list consists of zero or more entities $e_\ell = \emptyset$ or $e_\ell = \{e_1, \dots, e_m\}$ where \emptyset denotes null entities which are extremely common in recipe instructions (Kiddon et al., 2015; Huang et al., 2017) and e_i indicates entities such as *the Italian bread cubes*.

We define three sub-tasks. The first task is mention detection: it extracts all mentions e_ℓ from a_i . The second task is anaphora resolution: it assigns each e_i to an antecedent $y_i \in \{\epsilon, a_1, \dots, a_{i-1}, e_{1,\ell}, \dots, e_{i-1,\ell}\}$, if any. The third task is relation classification: it assigns one of the relation classes {NO-RELATION, COREFERENCE, NEAR-IDENTITY, BRIDGING} to each pair (e_i, y_i) . The selection of ϵ as the antecedent collapses two different situations: (1) the span is not an entity, or (2) the span is an entity but it is not referent (Lee et al., 2017). Likewise, if the relation is NO-

RELATION for relation classification, this points to two scenarios: (1) the span is not an entity, or (2) the span is an entity but it is not referent and so does not have an anaphoric relation to other entities.

5.2 Baseline

5.2.1 Visual Features

Each video consists of n segments, v_1, \dots, v_n , each corresponding to one instruction. Following Zhou et al. (2018a), we evenly divide each segment into five clips and randomly sample one frame from each clip to capture the temporal features of that segment. Each frame f_i is encoded using the Vision Transformer (ViT) model (Dosovitskiy et al., 2021). The instruction’s visual feature vector is obtained by concatenating the frame-level feature vectors: $v_i = \text{CNN}([\text{ViT}(f_1), \dots, \text{ViT}(f_5)])$.

5.2.2 Mention Detection

For mention detection, following Lee et al. (2017), we consider all continuous tokens with up to L words as a potential span and compute the corresponding span score. BERT (Devlin et al., 2019) is used to extract the contextualised word embeddings $x_t^* = \text{BERT}(w_1, \dots, w_T)$ where x_t^* refers to the vector representation of the token at time t of R . The vector representation g_i of a given span is obtained by concatenating the word vectors of its boundary tokens and its width feature:

$$g_i = [x_{\text{START}(i)}^*, x_{\text{END}(i)}^*, \phi(i)]$$

$$\phi(i) = \text{WIDTH}(\text{END}(i) - \text{START}(i)).$$

START(i) and END(i) represent the starting and ending token indexes for g_i , respectively. $\phi(i)$ is the width feature of the span where WIDTH(.) is the embedding function of the predefined bins of [1, 2, 3, 4, 8, 16] as defined by Clark and Manning (2016).

The use of head attention (Lee et al., 2017; Yu and Poesio, 2020; Fang et al., 2021) is very common in coreference/anaphora resolution models. However, we disregard the head representation of spans for two reasons: (1) the common use of null anaphors in our data: instead the instruction a_i of the null anaphor is used for extracting the vector representation, (2) the self-attention mechanism (Vaswani et al., 2017) of the BERT model implicitly captures the mention head word.

The mention score $\text{softmax}(\text{FFNN}(g_i))$ is computed for each span, and the mention model is trained using the cross-entropy loss.

5.2.3 Anaphora Resolution

For anaphora resolution, the representation of span pair g_{ij} is obtained by concatenating the two span embeddings $[g_i, g_j]$ and their element-wise multiplication, $g_i \cdot g_j$, among others:

$$g_{ij} = [g_i, g_j, g_i \cdot g_j, v_i \cdot v_j, \phi_{dist}(i, j)]$$

$$\phi_{dist}(i, j) = \text{DISTANCE}(\text{START}(j) - \text{START}(i))$$

where the feature vector $\phi_{dist}(i, j)$ is the distance between the index of span i and span j . DISTANCE(.) is an embedding function of the predefined bins of [1, 2, 3..., 30] as defined by Clark and Manning (2016).

For anaphora resolution, we minimize the cross entropy loss for candidate span pairs with $\text{sigmoid}(\text{FFNN}(g_{ij}))$.

5.2.4 Relation Classification

As shown in Table 1, the number of observed hyponym, reduce, set-member, and part-of instance relations is low. Therefore, we define the anaphoric relations in term of the three main categories: coreference, near-identity, and bridging.

To learn the vectors for each relation of feature vector g_{ij} , we apply an FFNN layer:

$$\text{coreference}_{ij} = \text{FFNN}(g_{ij})$$

$$\text{n-identity}_{ij} = \text{FFNN}(g_{ij})$$

$$\text{bridging}_{ij} = \text{FFNN}(g_{ij}).$$

Then, we concatenate coreference_{ij} , n-identity_{ij} ,

and bridging_{ij} into the relation vector rel_{ij} :

$$\text{rel}_{ij} = [\text{coreference}_{ij}, \text{n-identity}_{ij}, \text{bridging}_{ij}].$$

To classify the anaphoric relation for each input pair, we then compute $\text{softmax}(\text{FFNN}([g_{ij}, \text{rel}_{ij}]))$.

5.3 Temporal Features

Recipe instructions are written with an implied temporal order (Jermurawong and Habash, 2015), and the entities involved go through this temporal order until the cooking is complete. We propose to select the number of instructions (see Figure 2) as the temporal marker of entities instead of token distance $\phi_{dist}(i, j)$ to avoid issues with different instruction and entity lengths. We design our experiments to explain how the temporal stage of entities in action flows influences the pair representation of mentions in cooperating with the anaphora resolution model. Thus, we formulate our temporal features as

$$\phi_{temp}(i, j) = \text{TEMPORAL}(\#a_j - \#a_i)$$

where TEMPORAL(.) is an embedding function that uses the list of bins [1,2,3...,30]. $\#a_i$ refers to the instruction index of span i and $\#a_j$ to the instruction index of span j . We concatenate $\phi_{temp}(i, j)$ in place of $\phi_{dist}(i, j)$ to obtain the vector representation of a span pair:

$$g_{ij} = [g_i, g_j, g_i \cdot g_j, v_i \cdot v_j, \phi_{temp}(i, j)].$$

Token distance varies depending on the use of token numbers in instructions and entities. For example, the instruction *mix red chili cinnamon stick cloves cumin seeds mustard seeds pepper garlic vinegar sugar and wine* might also be written *mix red chili cinnamon stick cloves cumin seeds mustard seeds* followed by *add pepper garlic vinegar in the bowl* and *mix with sugar and wine*. Therefore, temporal features are not captured well by token distance in instructional language.

6 Experimental Setup

6.1 Input

Cooking Instructions. To encode the recipes we use BERT (Devlin et al., 2019), a bidirectional transformer model trained on a masked language modeling task. First, we fine-tune BERT-large-uncased by using the YouCookII dataset (Zhou et al., 2018a) after removing our test recipes. Because of sub word embeddings, there are different

	Candidate Spans			Gold Spans		
	Precision	Recall	F1-score	Precision	Recall	F1-score
w/o Temporal						
Anaphora Resolution	48.1	34.1	39.9	48.9	46.7	47.8
Coreference	34.2	43.4	38.2	40.1	47.5	43.5
Near-identity	66.8	37.0	47.7	78.5	38.8	51.9
Bridging	12.0	37.5	18.2	16.7	45.0	24.3
Overall Relation	21.6	44.6	29.2	28.4	50.3	36.3
w Temporal						
Anaphora Resolution	48.7	34.2	40.0	51.2	50.0	50.6
Coreference	29.1	45.8	35.6	46.1	50.6	48.3
Near-identity	57.0	33.8	42.4	90.1	44.7	59.7
Bridging	14.7	41.9	21.7	24.4	43.7	31.3
Overall Relation	22.6	46.2	30.4	32.6	54.3	40.8

Table 2: Average evaluation results over 3 runs of the proposed anaphora resolution model on our annotated test data for 200 epochs. **w Temporal** and **w/o Temporal** refer to the results with or without temporal features, respectively. Candidate Spans refers to all the possible spans of continuous tokens extracted from the recipes whereas Gold Spans refers the mentions with nominal phrases, null anaphors, and instructions.

choices of presenting words. We use the first sub-token for representing the word as proposed by Devlin et al. (2019). Additionally, due to the structure of multiple successive layers, the last hidden layer is used to represent the words in recipes.

Video Frames. To encode each video frame, ViT (Dosovitskiy et al., 2021) is pre-trained on ImageNet (Russakovsky et al., 2015) and fine-tuned on Food-101 (Bossard et al., 2014) images. In the end, each instruction (i.e., segment) is represented by a 3,840-dimensional vector v_i .

6.2 Experiments

Candidate Spans Without any pruning, we consider all continuous tokens (Clark and Manning, 2016; Lee et al., 2017) as a potential spans for the training and testing phases.

Gold Spans In order to investigate the performance of anaphora resolution and relation classification models without mention detection noise, we also consider gold spans for the training and testing phases.

6.3 Evaluation

Following Hou et al. (2018) and Yu and Poesio (2020), we analyze the performance of our end-to-end anaphora resolution model with its subtasks. For mention detection, anaphora resolution and relation classification we report F1-scores.

To evaluate mention detection, precision is computed as the fraction of correctly detected mentions among all detected mentions whereas recall is the fraction of correctly detected mentions among all

gold mentions. The F1-score for anaphora resolution is computed where precision is the result of dividing the number of correctly predicted pairs by the total number of predicted pairs and recall is computed by dividing the number of correctly predicted pairs by the total number of gold pairs. To evaluate relation classification we compute the F1-score where precision is computed by dividing the number of correctly predicted relations by the total number of predicted relations and recall is computed by dividing the number of correctly predicted relations by the total number of gold relations.

6.4 Results and Discussion

6.4.1 Overview

We investigate the anaphora resolution and relation classification results of gold and candidate spans comparing the F1-scores with the distance and temporal features. Overall, our results in Table 2 demonstrate that replacing token distance with our temporal features improves anaphora resolution and relation classification for both candidate and gold spans.

The performance of each task is propagated to subsequent tasks due to the sequential structure of the end-to-end system (see Section 5). The difference between the results of candidate and gold spans demonstrates that the mention detection model propagates errors to anaphora resolution and relation classification. For example, temporal features are not predictive features for anaphoric relations, but they are valuable for finding the antecedent of an anaphor, i.e., anaphora resolution. Our observations show that improvements in re-

616 lation classification are propagated from the pre-
617 ceding anaphora resolution task in the end-to-end
618 system for gold spans.

619 Additionally, binary mention detection results
620 show a precision of 0.92, a recall of 0.88, and an F1-
621 score of 0.90. However, the differences between
622 the scores in anaphora resolution and relation clas-
623 sification results for the candidate and gold spans
624 (see Table 2) reveal issues in transferring the men-
625 tion features. We observe the main problem of
626 mention detection in distinguishing the singletons.

627 6.4.2 Anaphora Resolution

628 We detect a significant improvement in anaphora
629 resolution with temporal features, since temporal
630 features often conspire to reduce unwelcome lexi-
631 cal similarity. For example, *potato* → *it* → *potato*,
632 the first *potato* is the antecedent of *it*, and *it* is the
633 antecedent of the second *potato*. Temporal features
634 prevent predicting the first *potato* as an antecedent
635 for the second *potato* and designate the anaphora
636 link from the second *potato* to *it*, because *it* is in
637 the instruction closer in the temporal line. The
638 improvements with temporal features reveal the
639 issues of contextualized embeddings. While we
640 use contextualized embeddings, the bias of lexical
641 similarity induces complexity to link the anaphor
642 with a correct antecedent; as recurrent in the *bacon*
643 → *bacon* → *fried bacon* sample in Figure 2. The
644 sliced bacon is predicted as the antecedent of the
645 bacon of instruction 3, and it is also the antecedent
646 of fried bacon of instruction 8. This issue occurs
647 for rare entities and predicates. When we compare
648 the false positives in accordance with temporality,
649 the improvement due to temporal features mainly
650 affects pronoun resolution. Hence, we observe that
651 the antecedents of pronouns are closer to the pro-
652 nouns. Some anomalies can be observed in the
653 results of anaphora resolution with candidate spans
654 due to the propagated error from mention detec-
655 tion. For example, we have the candidate spans
656 *the pizza*, *pizza dough*, and *the pizza dough* for the
657 mention *the pizza dough* of instruction 4 with the
658 same temporal features.

659 6.4.3 Relation Classification

660 Table 2 shows that temporal features significantly
661 improve anaphora resolution results for gold spans.
662 Especially for bridging pairs, a noteworthy benefit
663 of temporal features can also be observed in gold
664 and candidate spans. However, the mistakes can
665 also be observed in the results of near-identity and

666 coreference classification for candidate spans.

667 Overall, the end-to-end model suffers from mis-
668 takes in detecting and resolving null anaphors. Ex-
669 pecting that all instructions contain a null anaphor
670 increases the input noise for candidate spans. Re-
671 lation classification follows anaphora resolution
672 and mention detection. Therefore, some problems
673 in relation classification originate from mention
674 detection and anaphora resolution errors.

675 False positive bridging relations are due to sin-
676 gleton spans (non-referents) whereas false positive
677 coreference and near-identical relations are due to
678 the preference for surface words with/without state
679 changes. For instance, in the example *wash the egg*
680 $\xrightarrow{\text{coreference}}$ *boil the egg* $\xrightarrow{\text{near-identity}}$ *crack the egg*,
681 the use of the same words for changing entities
682 introduces an immense modelling challenge.

683 7 Conclusion and Future Work

684 We introduce a novel anaphora annotation scheme
685 including the state changes of entities and near-
686 identical relations. This fresh approach relies on
687 video inputs for visual observation for anaphora an-
688 notation. Likewise, we provide baseline anaphora
689 resolution results with novel temporal features on
690 the annotated data. In future work, the mention
691 detection model will be designed to perform with
692 null entities and singleton mentions to improve the
693 performance of the end-to-end model. Addition-
694 ally, different visual feature extraction methods for
695 single frames, e.g., CLIP (Radford et al., 2021)
696 or for videos, e.g., S3D (Xie et al., 2018) will be
697 investigated to find the best way of learning from
698 cooking videos for anaphora resolution.

699 8 Acknowledgements

700 We would like to thank Iuliia Zaitova for help-
701 ing with the annotation study. This research was
702 funded by the joint IMPRESS (01IS20076) project
703 between the French National Institute for Research
704 in Digital Science and Technology (Inria) and the
705 German Research Center for Artificial Intelligence
706 (DFKI).

707 References

- 708 Nicholas Asher and Alex Lascarides. 1998. Bridging.
709 *Journal of Semantics*, 15(1):83–113.
- 710 Devansh Batra, Nirav Diwan, Utkarsh Upadhyay,
711 Jushaan Singh Kalra, Tript Sharma, Aman Kumar
712 Sharma, Dheeraj Khanna, Jaspreet Singh Marwah,

713	Srilakshmi Kalathil, Navjot Singh, Rudraksh Tuwani, and Ganesh Bagler. 2020. RecipeDb: a resource for exploring recipes . <i>Database: The Journal of Biological Databases and Curation</i> , 2020.	770
714		771
715		772
716		
717	Stefan Baumann and Arndt Riester. 2012. Referential and lexical givenness: Semantic, prosodic and cognitive aspects. In <i>Prosody and meaning</i> , pages 119–162. De Gruyter Mouton.	773
718		774
719		775
720		776
721	Lukas Bossard, Matthieu Guillaumin, and Luc Van Gool. 2014. Food-101–mining discriminative components with random forests. In <i>European conference on computer vision</i> , pages 446–461. Springer.	777
722		
723		779
724		780
725	Herbert H. Clark. 1975. Bridging . In <i>Theoretical Issues in Natural Language Processing</i> .	781
726		782
727	Kevin Clark and Christopher D. Manning. 2016. Improving coreference resolution by learning entity-level distributed representations . In <i>Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)</i> , pages 643–653, Berlin, Germany. Association for Computational Linguistics.	783
728		784
729		785
730		786
731		787
732		788
733		789
734	Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. BERT: Pre-training of deep bidirectional transformers for language understanding . In <i>Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)</i> , pages 4171–4186, Minneapolis, Minnesota. Association for Computational Linguistics.	790
735		791
736		792
737		793
738		794
739		795
740		796
741		797
742		798
743	Alexey Dosovitskiy, Lucas Beyer, Alexander Kolesnikov, Dirk Weissenborn, Xiaohua Zhai, Thomas Unterthiner, Mostafa Dehghani, Matthias Minderer, Georg Heigold, Sylvain Gelly, Jakob Uszkoreit, and Neil Houlsby. 2021. An image is worth 16x16 words: Transformers for image recognition at scale. <i>ArXiv</i> , abs/2010.11929.	799
744		800
745		801
746		802
747		
748		803
749		804
750	Biaoyan Fang, Timothy Baldwin, and Karin Verspoor. 2022. What does it take to bake a cake? the RecipeRef corpus and anaphora resolution in procedural text . In <i>Findings of the Association for Computational Linguistics: ACL 2022</i> , pages 3481–3495, Dublin, Ireland. Association for Computational Linguistics.	805
751		806
752		807
753		808
754		809
755		810
756		811
757	Biaoyan Fang, Christian Druckenbrodt, Saber A Akhondi, Jiayuan He, Timothy Baldwin, and Karin Verspoor. 2021. ChEMU-ref: A corpus for modeling anaphora resolution in the chemical domain . In <i>Proceedings of the 16th Conference of the European Chapter of the Association for Computational Linguistics: Main Volume</i> , pages 1362–1375, Online. Association for Computational Linguistics.	812
758		813
759		814
760		815
761		816
762		817
763		818
764		819
765	Abbas Ghaddar and Philippe Langlais. 2016. Wiki-coref: An english coreference-annotated corpus of wikipedia articles . In <i>Proceedings of the Tenth International Conference on Language Resources and Evaluation (LREC’16)</i> , pages 136–142.	820
766		821
767		822
768		823
769		824
		825
		826
	Yufang Hou, Katja Markert, and Michael Strube. 2018. Unrestricted bridging resolution. <i>Computational Linguistics</i> , 44(2):237–284.	
	Eduard Hovy, Teruko Mitamura, Felisa Verdejo, Jun Araki, and Andrew Philpot. 2013. Events are not simple: Identity, non-identity, and quasi-identity . In <i>Workshop on Events: Definition, Detection, Coreference, and Representation</i> , pages 21–28, Atlanta, Georgia. Association for Computational Linguistics.	
	De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Nieves. 2018. Finding "it": Weakly-supervised reference-aware visual grounding in instructional videos . In <i>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition</i> , pages 5948–5957.	
	De-An Huang, Joseph J Lim, Li Fei-Fei, and Juan Carlos Nieves. 2017. Unsupervised visual-linguistic reference resolution in instructional videos. In <i>Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition</i> , pages 2183–2192.	
	Ting-Hao Huang, Francis Ferraro, Nasrin Mostafazadeh, Ishan Misra, Aishwarya Agrawal, Jacob Devlin, Ross Girshick, Xiaodong He, Pushmeet Kohli, Dhruv Batra, et al. 2016. Visual storytelling. In <i>Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies</i> , pages 1233–1239.	
	Jermisak Jermisurawong and Nizar Habash. 2015. Predicting the structure of cooking recipes . In <i>Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing</i> , pages 781–786, Lisbon, Portugal. Association for Computational Linguistics.	
	Chloé Kiddon, Ganesa Thandavam Ponnuraj, Luke Zettlemoyer, and Yejin Choi. 2015. Mise en place: Unsupervised interpretation of instructional recipes . In <i>Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing</i> , pages 982–992.	
	Nikhil Krishnaswamy and James Pustejovsky. 2019. Generating a novel dataset of multimodal referring expressions . In <i>Proceedings of the 13th International Conference on Computational Semantics - Short Papers</i> , pages 44–51, Gothenburg, Sweden. Association for Computational Linguistics.	
	Nikhil Krishnaswamy and James Pustejovsky. 2020. A formal analysis of multimodal referring strategies under common ground . In <i>Proceedings of the Twelfth Language Resources and Evaluation Conference</i> , pages 5919–5927, Marseille, France. European Language Resources Association.	
	Kenton Lee, Luheng He, Mike Lewis, and Luke Zettlemoyer. 2017. End-to-end neural coreference resolution . In <i>Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing</i> , pages 188–197, Copenhagen, Denmark. Association for Computational Linguistics.	

827	Katja Markert, Yufang Hou, and Michael Strube. 2012.	(LREC'16), pages 1743–1749, Portorož, Slovenia.	883
828	Collective classification for fine-grained information	European Language Resources Association (ELRA).	884
829	status. In <i>Proceedings of the 50th Annual Meeting of</i>		
830	<i>the Association for Computational Linguistics (Vol-</i>	Ina Rösiger. 2018. BASHI: A corpus of Wall Street	885
831	<i>ume 1: Long Papers)</i> , pages 795–804.	Journal articles annotated with bridging links . In <i>Pro-</i>	886
		<i>ceedings of the Eleventh International Conference on</i>	887
832	Massimo Poesio and Ron Artstein. 2008. Anaphoric	<i>Language Resources and Evaluation (LREC 2018)</i> ,	888
833	annotation in the ARRAU corpus . In <i>Proceedings</i>	Miyazaki, Japan. European Language Resources As-	889
834	<i>of the Sixth International Conference on Language</i>	sociation (ELRA).	890
835	<i>Resources and Evaluation (LREC'08)</i> , Marrakech,		
836	Morocco. European Language Resources Associa-	Ina Rösiger, Arndt Riester, and Jonas Kuhn. 2018.	891
837	tion (ELRA).	Bridging resolution: Task definition, corpus re-	892
		sources and rule-based experiments. In <i>Proceedings</i>	893
838	Massimo Poesio, Yulia Grishina, Varada Kolhatkar,	<i>of the 27th International Conference on Computa-</i>	894
839	Nafise Moosavi, Ina Roesiger, Adam Roussel, Fabian	<i>tional Linguistics</i> , pages 3516–3528.	895
840	Simonjetz, Alexandra Uma, Olga Uryupina, Juntao		
841	Yu, and Heike Zinsmeister. 2018. Anaphora resolu-	Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause,	896
842	tion with the ARRAU corpus . In <i>Proceedings of the</i>	Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej	897
843	<i>First Workshop on Computational Models of Refer-</i>	Karpathy, Aditya Khosla, Michael Bernstein, Alexan-	898
844	<i>ence, Anaphora and Coreference</i> , pages 11–22, New	der C. Berg, and Li Fei-Fei. 2015. ImageNet Large	899
845	Orleans, Louisiana. Association for Computational	Scale Visual Recognition Challenge . <i>International</i>	900
846	Linguistics.	<i>Journal of Computer Vision (IJCV)</i> , 115(3):211–252.	901
847	Massimo Poesio and Renata Vieira. 1998. A corpus-	Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob	902
848	based investigation of definite description use . <i>Com-</i>	Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz	903
849	<i>putational Linguistics</i> , 24(2):183–216.	Kaiser, and Illia Polosukhin. 2017. Attention is all	904
		you need . In <i>Advances in Neural Information Pro-</i>	905
850	Sameer Pradhan, Alessandro Moschitti, Nianwen Xue,	<i>cessing Systems</i> , volume 30. Curran Associates, Inc.	906
851	Olga Uryupina, and Yuchen Zhang. 2012. Conll-		
852	2012 shared task: Modeling multilingual unrestricted	R Weischedel, S Pradhan, L Ramshaw, J Kaufman,	907
853	coreference in ontonotes. In <i>Joint Conference on</i>	M Franchini, M El-Bachouti, N Xue, M Palmer,	908
854	<i>EMNLP and CoNLL-Shared Task</i> , pages 1–40.	JD Hwang, C Bonial, et al. 2012. Ontonotes release	909
		5.0. linguistic data consortium. Technical report,	910
855	Alec Radford, Jong Wook Kim, Chris Hallacy, Aditya	Philadelphia, Technical Report.	911
856	Ramesh, Gabriel Goh, Sandhini Agarwal, Girish Sas-		
857	try, Amanda Askell, Pamela Mishkin, Jack Clark,	Saining Xie, Chen Sun, Jonathan Huang, Zhuowen Tu,	912
858	et al. 2021. Learning transferable visual models	and Kevin Murphy. 2018. Rethinking spatiotemporal	913
859	from natural language supervision. In <i>International</i>	feature learning: Speed-accuracy trade-offs in video	914
860	<i>Conference on Machine Learning</i> , pages 8748–8763.	classification. In <i>Proceedings of the European con-</i>	915
861	PMLR.	<i>ference on computer vision (ECCV)</i> , pages 305–321.	916
862	Marta Recasens, Eduard Hovy, and M. Antònia Martí.	Juntao Yu and Massimo Poesio. 2020. Multitask	917
863	2010. A typology of near-identity relations for coref-	learning-based neural bridging reference resolution .	918
864	erence (NIDENT) . In <i>Proceedings of the Seventh</i>	In <i>Proceedings of the 28th International Conference</i>	919
865	<i>International Conference on Language Resources</i>	<i>on Computational Linguistics</i> , pages 3534–3546,	920
866	<i>and Evaluation (LREC'10)</i> , Valletta, Malta. Euro-	Barcelona, Spain (Online). International Committee	921
867	pean Language Resources Association (ELRA).	on Computational Linguistics.	922
868	Marta Recasens, Eduard Hovy, and M Antònia Martí.	Luowei Zhou, Nathan Louis, and Jason J. Corso. 2018a.	923
869	2011. Identity, non-identity, and near-identity: Ad-	Weakly-supervised video object grounding from text	924
870	dressing the complexity of coreference. <i>Lingua</i> ,	by loss weighting and object interaction. In <i>BMVC</i> .	925
871	121(6):1138–1152.		
		Luowei Zhou, Chenliang Xu, and Jason J. Corso. 2018b.	926
872	Marta Recasens, M. Antònia Martí, and Constantin	Towards automatic learning of procedures from web	927
873	Orasan. 2012. Annotating near-identity from coref-	instructional videos. In <i>AAAI</i> .	928
874	erence disagreements . In <i>Proceedings of the Eighth</i>		
875	<i>International Conference on Language Resources</i>		
876	<i>and Evaluation (LREC'12)</i> , pages 165–172, Istanbul,		
877	Turkey. European Language Resources Association		
878	(ELRA).		
879	Ina Roesiger. 2016. SciCorp: A corpus of English		
880	scientific articles annotated for information status		
881	analysis . In <i>Proceedings of the Tenth International</i>		
882	<i>Conference on Language Resources and Evaluation</i>		