Bridging and Anaphorcity

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

German Research Center for Artificial Intelligence

December 7, 2023

Table of Contents

Coreference

I saw [**a Mercedes**] outside the restaurant. [**It**] belongs to Bill.

Bridging

I saw [**a Mercedes**] outside the restaurant. [**The engine**] was still running.

Introduction
Relationships of Referring Expression

Type	Example
1. Direct reference	
Identity	(1) I met *a man* yesterday. **The man** told me a story.
Pronominalization	(2) I met *a man* yesterday. **He** told me a story.
Epithets	(3) I met *a man* yesterday. **The bastard** stole all my money.
Set membership	(4) I met *two people* yesterday. **The woman** told me a story.
2. Indirect reference by association	
Necessary parts	(5) I looked into *the room*. **The ceiling** was very high.
Probable parts	(6) I walked into *the room*. **The windows** looked out to the bay.
Inducible parts	(7) I walked into *the room*. **The chandeliers** sparkled brightly.
3. Indirect reference by characterization	
Necessary roles	(8) John was *murdered* yesterday. **The murderer** got away.
Optional roles	(9) John was *murdered* yesterday. **The knife** lay nearby.
4. Reasons, causes, consequences and concurrences	
Reasons	(10) John fell, what he wanted to do was scare Mary.
Causes	(11) John fell. What he did was trip on a rock.
Consequences	(12) John fell. What he did was break his arm.
Concurrences	(13) John is a Republican. Mary is slightly daft too.

What do we mean by sameness and relatedness?

I saw [**a Mercedes**] outside the restaurant. [**It**] belongs to Bill.

I saw [**a Mercedes**] outside the restaurant. [**The engine**] was still running.

I bought [**a Mercedes**] last year. [**It**] is crushed in an accident yesterday.

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 . . . .

What do we mean by sameness and relatedness?
Question: How to define anaphoric relation for state changes of entities?

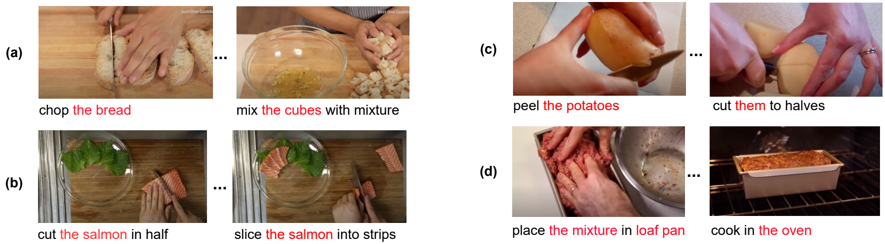
What do we mean by sameness and relatedness?

Question: How to define anaphoric relation for state changes of entities?

Scenario: Cooking recipes provide various changes for various entities ...

Introduction

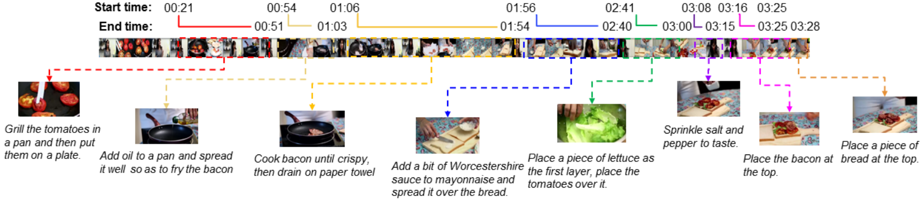
Referring Expression in Cooking Videos



The gap in the literature with the data, annotation schema, and resolution model for anaphoric relations to keep track of spatio-temporal changes of entities

Instructions of cooking videos provide visual and textual appearance for experimenting spatio-temporal changes

Data and Annotation
YouCookII (Zhou et al., 2018)



- YouCookII is a task-oriented, instructional video dataset for cooking recipes
- 2000 long untrimmed videos from 89 cooking recipes
- The instructional steps for each video are annotated with temporal boundaries and described by imperative English sentences

Data and Annotation

Anaphoric Relations and Entity Change

wash the tomatoes well

...

take the tomatoes aside

mix yogurt oil milk spices chives

put dressing on the salad

- **Coreference**: The anaphor and the antecedent are identical and point to the same entity.
  - wash [**the tomatoes**], take [**the tomatoes**] aside

Data and Annotation

Anaphoric Relations and Entity Change

wash the tomatoes well

...

take the tomatoes aside

mix yogurt oil milk spices chives

put dressing on the salad

- **Coreference**: The anaphor and the antecedent are identical and point to the same entity.
  - wash [**the tomatoes**], take [**the tomatoes**] aside
- **Bridging**: The antecedent is related and not identical to the anaphor.
  - Produced: [**mix yogurt oil milk spices chives**], put [**dressing**] on the salad
  - Reduced: slice [**the bread**], put cheese on [**one piece**]
  - Set-Member: wash [**cucumber, tomato, and lettuce**], cut [**the ingredients**]
  - Part-Of: cut [**the lemon**], take [**the seeds**] out

Data and Annotation

Anaphoric Relations and Entity Change

Near-Identity: Changes of physical or chemical properties (Recasens et al., 2011)

(a)

chop the bread ... mix the cubes with mixture

The cubes denotes the pieces after the bread is changed by cutting

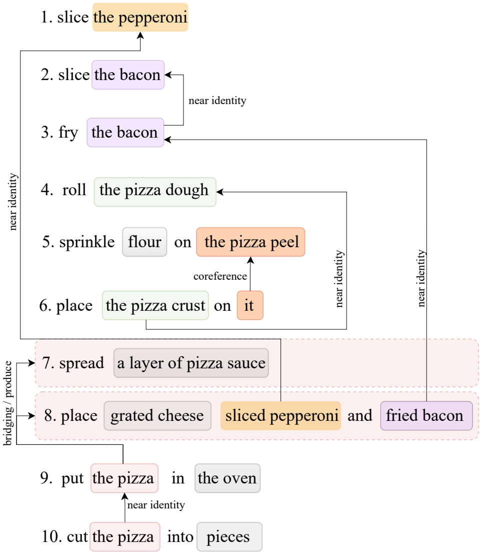
(c)

peel the potatoes ... cut them to halves

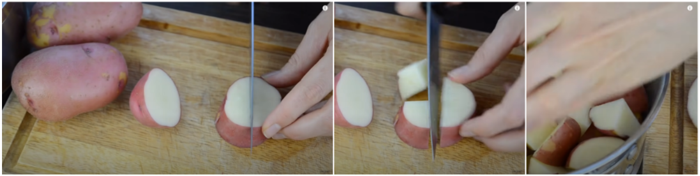
They denotes the potatoes after changed by peeled

Data and Annotation

Anaphoric Relations and Entity Change



An example of annotation schema with a recipe

...

...

... Cut the potatoes ...

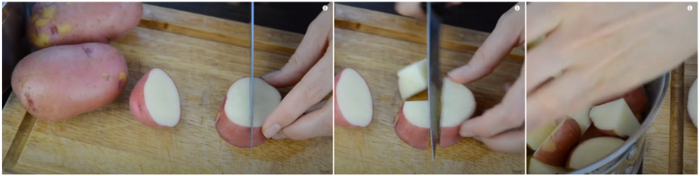
Visual

Divide each segment into five clips
Sample one frame from each clip
Encode frames with Vision Transformer (ViT) (Dosovitskiy et al., 2021)

Method

Input

...



...

... Cut the potatoes ...

Visual

Divide each segment into five clips
Sample one frame from each clip
Encode frames with Vision Transformer (ViT) (Dosovitskiy et al., 2021)

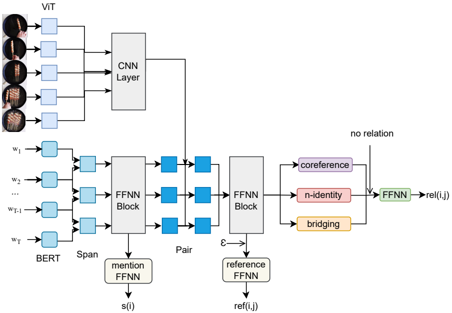
Textual

Use the full recipe to encode the words and use BERT to encode recipe
Extract the word n-grams:

- cut, the, potatoes, cut the, the potatoes, cut the potatoes

Use the vector of the boundary tokens to represent the span

We adapted end-to-end coreference resolution (Lee et al., 2017) on multitasks learning (Yu and Poesio, 2020)



Image

$$v_i = \mathrm{CNN}([\mathrm{ViT}(f_1), \ldots, \mathrm{ViT}(f_5)])$$

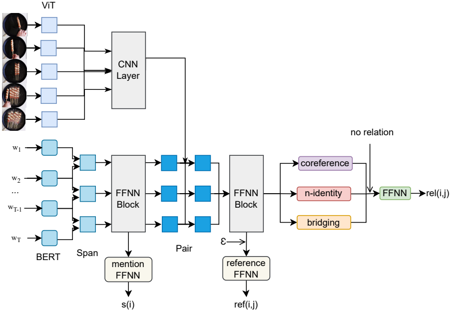
Text

$$\phi(i) = \textsc{width}(\textsc{end}(i) - \textsc{start}(i))$$

$$g_i = [x^*_{\textsc{start}(i)}, x^*_{\textsc{end}(i)}, \phi(i)]$$

Method

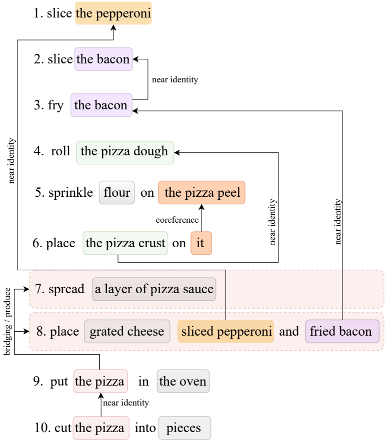
Model



$$\phi_{dist}(i,j) = \textsc{dist}(\textsc{start}(j) - \textsc{start}(i))$$
$$\mathrm{g}_{ij} = [g_i, g_j, g_i \cdot g_j, v_i \cdot v_j, \phi_{dist}(i,j)]$$
$$\mathrm{coreference}_{ij} = \mathrm{FFNN}(g_{ij})$$
$$\text{n-identity}_{ij} = \mathrm{FFNN}(g_{ij})$$

$$\mathrm{bridging}_{ij} = \mathrm{FFNN}(g_{ij})$$
$$\mathrm{rel}_{ij} = [\mathrm{coreference}_{ij}, \text{n-identity}_{ij}, \mathrm{briding}_{ij}]$$
$$\mathrm{softmax}(\mathrm{FFNN}([\mathrm{g}_{ij}, \mathrm{rel}_{ij}]))$$

Instead of using the token distance $\phi_{dist}(i,j)$, use the instruction distance



$$\phi_{temp}(i,j) = \text{TEMPORAL}(\#a_j - \#a_i)$$
$$g_{ij} = [g_i, g_j, g_i \cdot g_j, v_i \cdot v_j, \phi_{dist}(i,j)]$$
$$g_{ij} = [g_i, g_j, g_i \cdot g_j, v_i \cdot v_j, \phi_{temp}(i,j))]$$

- Compare the temporal features and the distance features
- See the effect of different features with candidate and gold spans
  - Example n-grams: cut, the, potatoes, cut the, the potatoes, cut the potatoes
  - Candidate spans: cut, the, potatoes, cut the, the potatoes, cut the potatoes
  - Gold spans: the potatoes, cut the potatoes

Method

Evaluation (Hou et al., 2018)

Recall, precision and F-score to measure the performance of anaphora resolution

recall = predicted correct links / gold anaphors
precision = predicted total links / gold anaphors

Method

Evaluation (Hou et al., 2018)

Recall, precision and F-score to measure the performance of anaphora resolution

recall = predicted correct links / gold anaphors
precision = predicted total links / gold anaphors

Recall, precision and F-score to measure the performance of relation classification

recall = predicted correct relation / gold relation
precision = predicted total relation / gold relation

Results

	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**

Average evaluation results over 3 runs of the proposed anaphora resolution model on our annotated test data for 200 epochs.

Conclusion

Anaphora Resolution: mention detection - anaphora resolution - relation classification

- Temporal features help for gold spans
  - Temporal features are not predictive for mention detection
- Difficulty: *bacon* $\rightarrow$ *bacon* $\rightarrow$ *fried bacon*
- the candidate spans *the pizza*, *pizza dough*, and *the pizza dough*

Conclusion

Find-2-Find: Multitask Learning for Anaphora Resolution and Object Localization

Cennet Oguz[1], Pascal Denis[2], Emmanuel Vincent[3] Simon Ostermann[1], Josef van Genabith[1]

[1]German Research Center for Artificial Intelligence (DFKI), Saarland Informatics
[2]Univ. Lille, Inria, CNRS, Centrale Lille, UMR 9189 CRIStAL, F-59000 Lille, France, [3]Université de Lorraine, CNRS, Inria, LORIA, F-54000 Nancy

Introduction

In multimodal understanding tasks, visual and linguistic ambiguities can arise.

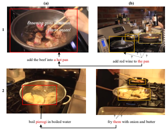


Figure 1. Examples of visual and linguistic ambiguities. Figure 1) represents the visual ambiguity related to which specific *pan* (in Figure 1a) is referenced with the phrase *the pan* because many pans occur on the stove. Figure 2) shows the linguistic ambiguity with the use of the pronoun *them* (in Figure 2b).

Figure 2. An example to display how visual-linguistic ambiguity occurs with a zero anaphor. The zero anaphor [ϕ] refers to two previous instructions as shown. The entities are aligned to the object with the arrows and the color codes.

- Visual ambiguity can occur when a model grounds a referring expression
- Linguistic ambiguity can occur from changes in entities in action flows

We define this chicken-and-egg problem as **visual-linguistic ambiguity**

Contributions

Our contributions are two-fold

- we present a new dataset *Find2Find*, for the joint evaluation of anaphora resolution and object localization.
- we present a new multitask learning system for modeling the two tasks of anaphora resolution and object localization jointly, using a fusion of visual and textual data.

Motivation

We propose Multitask Learning Object Localization and Anaphora Resolution

- Anaphora resolution addresses linguistic ambiguities

Modeling

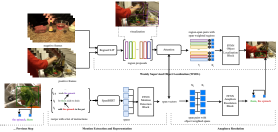


Figure 3. The architecture of the multitask learning framework of anaphora resolution and object localization.

Formula

Mention Detection

A span x_i consists of zero or more tokens of instruction I_i.

$$g_i = [x^*_{\text{start}(i)}, x^*_{\text{end}(i)}, \phi(i)]$$
$$\phi(i) = \text{width}(\text{end}(i) - \text{start}(i)).$$

start(i) and end(i) represent the starting and ending token indexes for g_i, respectively. $\phi(i)$ is the width feature

Object Localization

$$\text{FFNN}(g_i, r_i) = \begin{cases} 0 & x_i = \epsilon,\ \forall r_i \\ 0 & x_i \notin \{e_{i,1}, \ldots, e_{i,n}\},\ \forall r_i \\ 0 & x_i \in \{e_{i,1}, \ldots, e_{i,n}\},\ r_i \in \text{neg}_i \\ 1 & x_i \in \{e_{i,1}, \ldots, e_{i,n}\},\ r_i \in \text{pos}_i \end{cases}$$

Ten positive, i.e., pos_i, and ten negatives, i.e., neg_i, region representation vectors r_i to learn the best region from pos_i for the given span g_i.

Anaphora Resolution

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

where the feature vector $\phi_{dist}(i, j)$ is the distance distance(start(j) − start(i)) between the index of the instruction span i and span j.

Object Localization Data

- Use the YouCook2 [3] for object localization
- Use temporal boundaries for extracting video clips of each instruction
- Divide each video clip into 3 equal parts
- Pick only one frame of each of the 3 equal parts

Anaphora Resolution Data

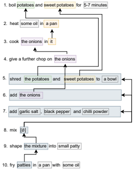


Figure 4. An example of the annotation

	Train	Test
Null Pronoun	1,002	282
Noun Phrases	8,314	2,560
Instruction	4,633	1,422
Recipe	400	100

Table 1. Annotated Data Statistics

Figure 5. Examples for showing the temporal changes and the referring expressions.

Results

	Nominal Anaphora Res.			Zero Anaphora Res.			Anaphora Res.		
Methods	Precision	Recall	F1-score	Precision	Recall	F1-score	Precision	Recall	F1-score
w/o Object L.									
Cand. Mention	54.76	46.65	50.38	73.38	68.76	71.00	63.03	54.06	58.20
Gold Mention	58.76	52.25	55.31	75.38	71.18	73.22	64.16	58.15	61.01
w Object L.									
Cand. Mention	52.03	50.49	51.25	77.68	69.97	73.63	62.46	56.19	59.16
Gold Mention	58.24	55.43	**56.80**	80.10	76.02	**78.01**	64.92	61.93	**63.39**

Table 2. Results of the anaphora resolution with and without object localization for gold and candidate mentions.

Methods	Nominal	Null	All
Random	13.98	16.66	14.07
DVSA w Gold Mentions	19.90	-	19.90
AR w Cand. Mentions	**21.02**	**24.46**	**20.79**
AR w Gold Mentions	**21.17**	**25.98**	**22.36**

Table 3. The Top-1 results of object localization with gold and candidate mentions.

References

[1] Cennet Oguz, Ivana Kruijff-Korbayova, Emmanuel Vincent, Pascal Denis, and Josef van Genabith. Chop and change: Anaphora resolution in instructional cooking videos. In *Findings of the Association for Computational Linguistics: AACL-IJCNLP 2022*, pages 364–374, Online only, November 2022. Association for Computational Linguistics.

[2] Yiwu Zhong, Jianwei Yang, Pengchuan Zhang, Chunyuan Li, Noel Codella, Liunian Harold Li, Luowei Zhou, Xiyang Dai, Lu Yuan, Yin Li, et al. Regionclip: Region-based language-image pretraining.

References

Dosovitskiy, A., Beyer, L., Kolesnikov, A., Weissenborn, D., Zhai, X., Unterthiner, T., Dehghani, M., Minderer, M., Heigold, G., Gelly, S., Uszkoreit, J., and Houlsby, N. (2021). An image is worth 16x16 words: Transformers for image recognition at scale. *ArXiv*, abs/2010.11929.

Hou, Y., Markert, K., and Strube, M. (2018). Unrestricted bridging resolution. *Computational Linguistics*, 44(2):237–284.

Lee, K., He, L., Lewis, M., and Zettlemoyer, L. (2017). End-to-end neural coreference resolution. *arXiv preprint arXiv:1707.07045*.

Recasens, M., Hovy, E., and Martí, M. A. (2011). Identity, non-identity, and near-identity: Addressing the complexity of coreference. *Lingua*, 121(6):1138–1152.

Yu, J. and Poesio, M. (2020). Multi-task learning based neural bridging reference resolution. *arXiv preprint arXiv:2003.03666*.

Zhou, L., Xu, C., and Corso, J. J. (2018). Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598.