A Weak Self-supervision with Transition-Based Modeling for Reference Resolution

Anonymous submission

Abstract

The reference resolution is a task to find the link between an entity and its source action in the same recipe. In this study, we introduce a weak self-supervision method with a transition-based model for reference resolution tasks for recipes, where the aim of the task is to make the syntax of the instructions used for reference resolution with self annotation. The results show that our approach to the problem outperforms the previous unsupervised methods with $\%8$ F1. Especially, our models show $> \%82$ accuracies of pronoun, and $> \%85$ accuracies for null entity resolution.

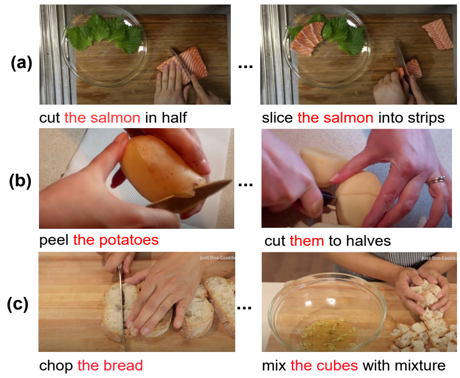

Figure 1: Examples of the references and entities in the recipe videos with the instructions.

1 Introduction

Recipe data has been rapidly growing in both visual and textual modalities and many studies have been using the subtitles of the instructional videos to obtain the joint embeddings of language and vision (Miech et al., 2019; Sun et al., 2019; Miech et al., 2020; Zhu and Yang, 2020), utilizing the descriptive sentences for video object grounding (Zhou et al., 2018a; Sadhu et al., 2020). On the other hand, videos are also used in many NLP tasks such as video question answering (Zeng et al., 2017; Le et al., 2020), machine translation (Sigurdsson et al., 2020; Gu et al., 2021), and so on. All these studies require one particular step to achieve good performance: resolving references to the objects. Since the given entities of a recipe are changed in the chain of actions, the inevitable linguistic ambiguities are presented in the recipe, see Figure 1. The lexical form for references might be with respect to the corresponding changes; the same nominal phrase might be used in the text even though the entity is changed in the visual domain Figure 1 a, a pronoun can be bound in place of the entity Figure 1 b, a new phrase might be replaced with the previous one Figure 1 c, etc. Hence, the reference resolution task in recipes (Kiddon et al., 2015) addresses learning of the source action that refers (also outputs) to the given entity in order to specify the changing status of entities. For example, *chop the bread* action refers to the *the cubes* in the action *mix the cubes with mixture*, Figure 1 c.

There has been a few attempts to address the reference resolution tasks that mainly formulate the reference resolution a graph optimization problem as determining the best edges between entities and the actions. Kiddon et al. (2015) use the self preferences between predicates and entitys of an action, and the conditional probability in between the entities and the previous actions (i.e., predicate, entity pair) to build the edges for obtaining an action graph. Furthermore, Huang et al. (2017) formulate the reference resolution problem as a graph optimization problem by adapting the likelihood measures from (Kiddon et al., 2015) to find the best edges between entities and the previous corresponding source action. The visual cues are used to constrain the entities to avoid the linguistic ambiguity. Huang et al. (2018) propose an entity-action pointer network to resolve the references by using visual object embeddings together with reference embeddings by using the given steps as the individual actions. However, we present the use of syntactic features of the instructions to obtain au-

tomatic annotation of the links (i.e., arcs) between actions and references for weak self-supervision learning, and a way of using the idea of transition-based dependency parsing method for the task.
Thus, two main contributions are presented here: (1) definition of referential tendency given by the choice of syntactic structure and type of referring expression in order to develop a weak self-supervision (2) an approach of using the method in transition-based dependency parsing for reference resolution in order to address the linguistic ambiguities of entities.

2 Problem Statement

2.1 Problem Statement

Each instruction text P consist of N number of ordered steps, where each step s, e.g. *pour olive oil on the Italian bread cubes and bake them in the oven*, includes T number of ordered actions e.g. 2 actions like *pour olive oil on the Italian bread cubes* and *bake them in the oven*. The given steps s are segmented into actions a and each action a_i in s_j defined as the pair of predicate p and the undergoing entity e.

$$\mathbf{P} = s_1, ..., s_N,\ 0 < N$$

$$\mathbf{s}_j = a_1, ..., a_T,\ 0 < T,\ \mathbf{a}_i = (p_i, e_i)$$

where p specifies the predicate of the action a_i, whereas e_i defines the corresponding entity. The entity resolution problem is a task to align the entity e_i of action a_i to its source action a_o which is one of the previous actions $a$in P and the latest action applied to undergoing entity e_i where $1 \leq n \leq i$, if any.

$$a_o = \alpha(e_i, a_1, ..., a_{i-1})$$

So, we formulate the reference link resolution to find the most likely relevant reference edge (i.e. $e_i \rightarrow a_o$) from source action to produced entity by the source action.
The new inputs (e.g., raw ingredients) are not considered as the produced entity. For example, the entity *an egg* of the first action in Figure 2 is the new input which is not produced in any previous actions in the recipe.

2.2 Evaluation

We compute the F-score for evaluation of reference resolution as it is denoted in the previous reference resolution studies (Kiddon et al., 2015; Huang et al., 2017, 2018) where precision P indicates how many of all the resolved references are correct with the formula $P = \frac{tp}{tp+fp}$ whereas recall R measure how many of the all references are correctly resolved with the formula $R = \frac{tp}{tp+fn}$ where tp designates the number of references that are correctly resolved, fp is the number of references that are not reference (e.g. raw ingredients) but recognized as reference, fn is the number of reference that are not detected as reference. We need to note here that only the *relevant edges* from both the predicted and the ground-truth references are considered. The relevant edges are ones between objects to action indices A_j where $j >= 0$.

3 Reference Resolution in Recipes

3.1 Reference Link Patterns in Instructions

Since the step combines more than one action together we define the syntax structures of steps in order to decompose the steps into sequential actions. To extract the entity references for weak self-supervision, we leverage the syntax structures of each action a_i where a_i consists of n number of entities, $n \geq 0$.

Single action. A predicate define the action with the including argument set. For an example, *pour the dry bread crumbs into a shallow dish* in Figure 2. Since one predicate is indicated to hold the action, it is named as the single action. Thus, is it not considered a case to decompose.

Consecutive action with explicit argument. The case with a step that includes more than one verbs, so more than one actions are grouped in one step. Two possible consecutive types are observed in the recipes: (1) sequential acts which continue with the same entity to complete the step in more than one action (2) parallel acts which shift the entity in the following actions in the same step. For example, the step *coat some onion rings in batter and transfer them* , Figure 2, includes sequential acts with the same entity. However, the step of parallel acts, such as *cut some slices of daikon and chop some green onions*, includes two predicates with two different entities.

Consecutive action with explicit argument. Two consecutive predicates occur in the step, the second predicate process the result of the first predicate with an implicit argument. For example, the step *move the onion rings to the bread crumbs and coat evenly*, Figure 2.

1. crack *an egg* into a bowl and break *it*
2. pour dry bread crumbs into a shallow dish
3. coat *onion rings* in batter and transfer *them*
4. move *the onion rings* and coat
...

Figure 2: An example of recipe

In order to apply self-supervision we use the given syntactic features of recipes defined above. When a bound pronoun is presented in the following actions of the consecutive actions, the first action of the corresponding consecutive action is defined as the source, as also defined in centering theory (Grosz et al., 1995; Brennan et al., 1987). If a null entity appears in a consecutive action we use this to link the null entity to the first action of the given consecutive action, inspired by (Kehler, 2000). From the Figure 2 in action 4 the null entity of *coat* refers to the *move the onion rings* in the same step. Additionally, in order to analyze the effect of the lexical similarities, the entities are linked to their closer action which contains the similar entity. The (cosine-) similarity threshold of the link is defined 0.9.

3.2 Transition-Based Reference Resolution

Since an entity might be used in different actions more than one time in the same recipe (e.g., boil the egg, peel the egg, cut the egg, put the egg in the bowl, etc.), the challenge in learning the references is finding the latest action applied to the current entity. Therefore, we apply a transition-based reference resolution (TBRR) method which is inspired by transition-based parsing (Nivre, 2004; Chen and Manning, 2014) because of keeping the order of actions. A configuration $c = (s, b, R)$ consists of a stack s, a buffer b, and a set of predefined relations R between entity-action pair a_i in an actions A_i. The initial configuration for a recipe $A_1, ..., A_n$ is $s = [root]$, $b = [a_1, ..., a_n]$, $R = \phi$. A configuration c is terminal if the buffer is empty. Denoting s_i $(i = 1, 2, ...)$ as the i-th top element on the stack, and b_i $(i = 1, 2, ...)$ as the i-th element on the buffer. We define three possible relations between arguments $\alpha = \{input, follower, output\}$ where;

- *input(*s_i*,* b_i*)* defines that b_i is a new entity, not an output of any previous actions and moves the b_i to s, precondition is $cos(s_i, b_i) < threshold$

- *follower(*s_i*,* b_i*)* defines that b_i is a ellipses or pronoun entity which is output of the s_i action and removes b_i from buffer

- *output(*s_i*,* b_i*)* defines that b_i is an entity which is output of the s_i action and removes s_i from stack, precondition is $cos(s_i, b_i) > threshold$

4 Experiments

4.1 Data

For unsupervised training, we use the YouCookII (Zhou et al., 2018b) dataset which consists of 2000 cooking videos with the annotation of instruction steps. Each video instruction includes 3 to 15 steps, where each step is an imperative sentence and temporally aligned to the corresponding video segment. The evaluation set (Huang et al., 2018) including 90 videos of YouCookII with their instruction steps that contains the reference annotation between entities and relevant actions.

4.2 Method

To understand the importance of the lexical and contextualized representation we examine both since the cooking recipes belong to a domain where the usage of language is always very similar.

TBRR$_{lexical}$: The average embeddings FastText (Bojanowski et al., 2017) and GLoVe (Pennington et al., 2014) are concatenated to represent the inputs to classify the corresponding relation.

TBRR$_{context}$: The BERT (Devlin et al., 2018) is used to represent the local context of the entities with it whereas FastText used to encode the word features.

TBRR$_{swap}$: Since the actions might include more than one entity *mix egg yolk, yogurt, flour* if the buffer and stack contain the entities from the same action, we apply swap operation to take the previous action entities front.
To examine the effect of self-supervision, a simple feed-forward neural network is used to apply classification of the relations between the given stack and buffer entities. A linear layer is used to represent the stack entity and another linear layer used for buffer entity. Additionally, we also used the subtracted vector of the buffer and stack and a linear layer used in the model to encode it.

	1.0 Label	0.6 Label	0.2 Label	w/o Label	Transition-Based RR			
Previous Studies	F1	F1	F1	F1	Exp.	P	R	F1
VLRR	0.56	0.53	0.53	0.51	TBRR$_{lexical}$	0.65	0.52	0.58
PNRR(w/o Gnd)	0.59	0.59	0.53	0.49	TBRR$_{context}$	0.74	0.47	0.58
PNRR	0.62	0.61	0.51	0.49	TBRR$_{swap}$	0.79	0.47	0.60

Table 1: Results of the reference resolution of our model TBRR with the previous works VLRR and PNRR. The works are tested on the YouCookII dataset. The results of the previous works are delivered from their study, our results are produced by the average of three random train-test run.

5 Results and Analysis

5.1 Results

The aim of the study is to investigate using the effect of the centering theory (Grosz et al., 1995; Brennan et al., 1987) and ellipses (Kehler, 2000) in instructional language for weak self-supervision.
Table 1 shows the results of reference resolution with previous studies and our results. VLRR (Huang et al., 2017) proposes an unsupervised way for reference resolution by learning a joint visual-linguistic model. The PNRR (Huang et al., 2017) uses a pointer network (Vinyals et al., 2015) with hierarchical RNN encoder for the action flow. They both use GloVe (Pennington et al., 2014) for inputs. The fraction of labels on the table indicates the fraction of used labeled data. The full size 1.0 includes 60 recipes. Typically, we need to compare our results with the results which not use annotated data (the column w/o label). However, we also include the others to show the effectiveness of the study. Additionally, they also use the visual inputs of the videos for training the models.
As can be seen on the Table 1, our approaches outperform the others with $> \%8$ when we consider w/o label. VLRR model which uses visual input for learning the references with labeled data, our model constantly outperform $> \%2$. Additionally, our TBRR$_{swap}$ model shows better results than PNRR without visual inputs (w/o Gnd), but not PNRR with visual and labeled data when data fraction is > 0.2.
On the other hand, for the pronoun and null entities our approach shows good results. the lexical model (TBRR$_{lexical}$) model gives $\%82$ of all pronouns are resolved correctly, while the context model (TBRR$_{context}$) indicates $\%97.5$ of all pronouns are linked to correct source action. Moreover, $\%90.9$ of null entities resolved correctly with lexical model, and it is $\%85$ with context model. So, we can strongly claim that the application of centering theory improves the reference resolution.

5.2 Analysis of Transition-Based RR

When the lexical model is compared to the context model on the true positives, the context model gives better results with variances of the entities. For example the entity *the clam juice* of the action linked to the source action *Add the clam juice to the pan* correctly with the context model, whereas it is missed by the lexical model. However, as can be seen from the results this strength cannot create much difference since the context similarities are also high because of the strong domain bias. For example, the new ingredient *some green onions* is linked to the *some onions* as a false positive example with both. Furthermore, the lexical similarities between the different entities are creating a huge problem since the same entities are linked to each other thanks to the weak annotation. For example, *oil* of the action *put oil in the pan* and the *oil* of the action *mix oil, egg and yogurt* is different. However, the similarity is useful in the case of *knead the dough* and *Take a piece of dough*. Swap (TBRR$_{swap}$) model swaps the entities of the same actions. We see a significant effect of the swap since many actions include more than one entity such as *Add oil to the dough in the mixer* and the reference link can only be with previous actions. On the other hand, our model constantly fails with the relations like between *dough* and action *Add water to the flour in the mixer*.

6 Conclusion and Future Work

To conclude, we propose a transition based weakly supervised way of reference resolution in recipes and outperform the unsupervised methods even with a fraction of labeled data. So, our results indicate that the syntactic features of the instructions lead significant improvements on reference resolutions, and do not suggest blind segmentation of steps. And, transition-based approach might help to the studies like co-reference resolutions, anaphora resolution.

References

Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. 2017. Enriching word vectors with subword information. *Transactions of the Association for Computational Linguistics*, 5:135–146.

Susan E Brennan, Marilyn W Friedman, and Carl Pollard. 1987. A centering approach to pronouns. In *25th Annual Meeting of the Association for Computational Linguistics*, pages 155–162.

Danqi Chen and Christopher D Manning. 2014. A fast and accurate dependency parser using neural networks. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 740–750.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2018. Bert: Pre-training of deep bidirectional transformers for language understanding. *arXiv preprint arXiv:1810.04805*.

Barbara J Grosz, Aravind K Joshi, and Scott Weinstein. 1995. Centering: A framework for modelling the local coherence of discourse.

Weiqi Gu, Haiyue Song, Chenhui Chu, and Sadao Kurohashi. 2021. Video-guided machine translation with spatial hierarchical attention network. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing: Student Research Workshop*, pages 87–92.

De-An Huang, Shyamal Buch, Lucio Dery, Animesh Garg, Li Fei-Fei, and Juan Carlos Niebles. 2018. Finding" it": Weakly-supervised reference-aware visual grounding in instructional videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5948–5957.

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.

Andrew Kehler. 2000. Coherence and the resolution of ellipsis. *Linguistics and Philosophy*, 23(6):533–575.

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.

Thao Minh Le, Vuong Le, Svetha Venkatesh, and Truyen Tran. 2020. Hierarchical conditional relation networks for video question answering. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 9972–9981.

Antoine Miech, Jean-Baptiste Alayrac, Lucas Smaira, Ivan Laptev, Josef Sivic, and Andrew Zisserman. 2020. End-to-end learning of visual representations from uncurated instructional videos. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 9879–9889.

Antoine Miech, Dimitri Zhukov, Jean-Baptiste Alayrac, Makarand Tapaswi, Ivan Laptev, and Josef Sivic. 2019. Howto100m: Learning a text-video embedding by watching hundred million narrated video clips. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 2630–2640.

Joakim Nivre. 2004. Incrementality in deterministic dependency parsing. In *Proceedings of the workshop on incremental parsing: Bringing engineering and cognition together*, pages 50–57.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Arka Sadhu, Kan Chen, and Ram Nevatia. 2020. Video object grounding using semantic roles in language description. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10417–10427.

Gunnar A Sigurdsson, Jean-Baptiste Alayrac, Aida Nematzadeh, Lucas Smaira, Mateusz Malinowski, Joao Carreira, Phil Blunsom, and Andrew Zisserman. 2020. Visual grounding in video for unsupervised word translation. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition*, pages 10850–10859.

Chen Sun, Austin Myers, Carl Vondrick, Kevin Murphy, and Cordelia Schmid. 2019. Videobert: A joint model for video and language representation learning. In *Proceedings of the IEEE/CVF International Conference on Computer Vision*, pages 7464–7473.

Oriol Vinyals, Meire Fortunato, and Navdeep Jaitly. 2015. Pointer networks. *arXiv preprint arXiv:1506.03134*.

Kuo-Hao Zeng, Tseng-Hung Chen, Ching-Yao Chuang, Yuan-Hong Liao, Juan Carlos Niebles, and Min Sun. 2017. Leveraging video descriptions to learn video question answering. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Luowei Zhou, Nathan Louis, and Jason J Corso. 2018a. Weakly-supervised video object grounding from text by loss weighting and object interaction. *arXiv preprint arXiv:1805.02834*.

Luowei Zhou, Chenliang Xu, and Jason J Corso. 2018b. Towards automatic learning of procedures from web instructional videos. In *AAAI Conference on Artificial Intelligence*, pages 7590–7598.

Linchao Zhu and Yi Yang. 2020. Actbert: Learning global-local video-text representations. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 8746–8755.