Grounded Multimodal Procedural Entity Recognition for Procedural Documents: A New Dataset and Baseline

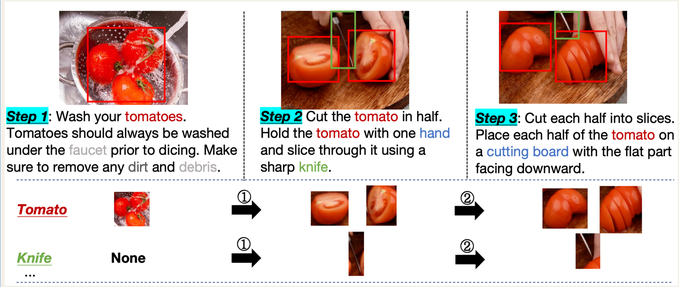
Haopeng Ren[1,2], Yushi Zeng[1,2],Yi Cai[1,2], Zhenqi Ye[1,2], Li Yuan[1,2], Pinli Zhu[1,2]
[1]School of Software Engineering, South China University of Technology, Guangzhou, China
[2]Key Laboratory of Big Data and Intelligent Robot (SCUT), Ministry of Education
hpren_scut@foxmail.com, ycai@scut.edu.cn

Introduction

Background: In our daily life, much of commonsense knowledge is in the form of sequences of actions to achieve particular goals (e.g., cooking recipes, crafting and maintenance manuals), which is called Procedural Knowledge and benefits multiple downstream applications: sequence ording, question answering and opertion diagnosis.

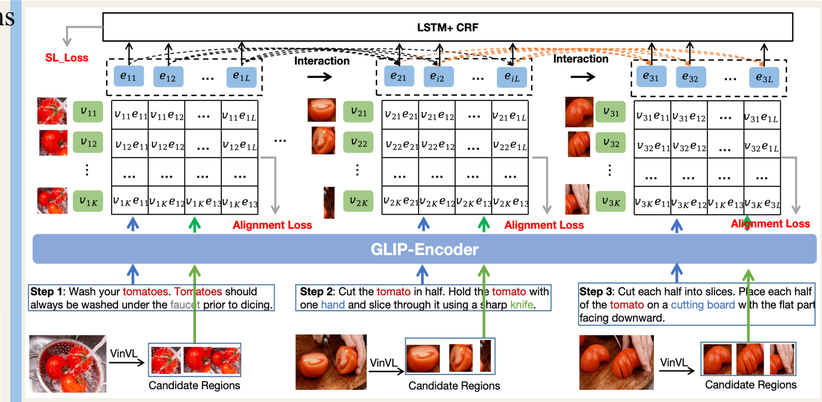
Challenges:

1. Grounded Multimodal Procedural Entity Recognition (GMPER) is based on long multimodal procedural documents with **multiple steps and complex interactions** between procedural entities.
2. **The state of the same visual procedural entities**, such as shape, color and forms (e.g., solid, liguid and gaseous) will **dynamically changes** as the procedural progresses.



Two Cases of Object Detection by Grounded Language-Image Pretrained Model (GLIP) (Li et al., 2022) with the Prompt Text "Tomato or Persimmon"; The highest confidence score for detecting object "tomato" in left image is 0.76, while that in the right image is 0.57.

Model Details



1. Problem Definition and Notations

Given a multimodal procedural document with a sequence of steps D = $\{s_1, s_2, \ldots, s_{Ld}\}$ and a corresponding sequence of images V = $\{v_1, v_2, \ldots, v_{Ld}\}$, the goal of the Grounded Multimodal Procedural Entity Recognition (GMPER) task is to extract a set of entity tuples: Y = $\{(e_1, r_1),\ldots,(e_t, r_t)\}$.

2. Multimodal Feature Representation

2.1 Text&Image Representation

Pretrained multimodal encoder in GLIP (Li et al., 2022) is used to extract features for both the text and image in each step.

2.2 Candidate Region Representation

A widely-adopted object detection model VinVL (Zhang et al., 2021b) is utilized to extract the candidate semantic visual region. Then, we rank candidate visual regions based on their detection probabilities and obtain the Top-K candidate visual regions.

3. Multimodal Sequenctial Feature Fusion

Considering that the procedural entities or regions discovered in the previous step can provide the important clues for the identification of entities and regions in the following steps, a sequential element attention machanism for each step:

$$\boldsymbol{t}_{i,j}^{fuse} = [\sum_{j=0}^{m} \alpha_{i-1,j}\boldsymbol{t}_{i-1,j}; \boldsymbol{t}_{i,j}]$$

$$\alpha_{i-1,j} = \frac{e^{\boldsymbol{t}_{i-1,j}\boldsymbol{t}_{i,j}}}{\sum_{k=0}^{m} e^{\boldsymbol{t}_{i-1,k}\boldsymbol{t}_{i,j}}}$$

4. GMPER

Three subtasks are conducted, including Procedural Entity Recognition (PER), Binary Groundable Classification (BGC) and Grounded Procedural Entity (GPE)

Experiment Results

We conduct extensive experiments on our annotated dataset Wiki-GMPER. Comparing with existing text-only NER methods and existing MNER and GMNER methods, our proposed methods achieve better performance in all experimental settings.

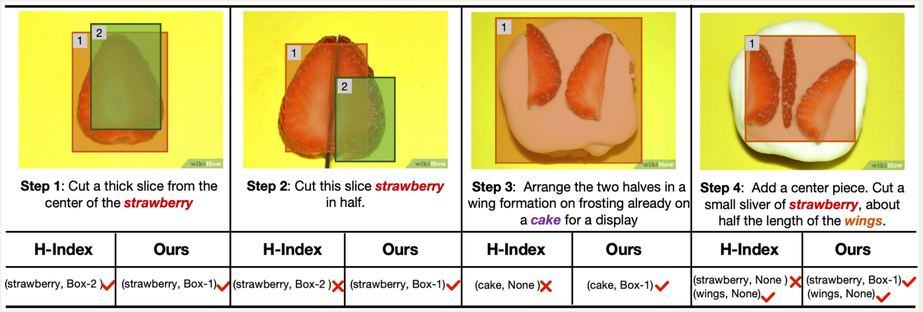
Dataset Statistics	Train	Test	Validation
# Doc.	809	299	122
# Step	4794	1341	701
Avg Step of Doc.	5.90	5.00	5.75
# Entity	19869	5836	2738
# Groundable	11029	2854	1252
# Ungroundable	8840	2982	1486

Table 1: The statistics of our annotated dataset Wiki-GMPER

Model		Pre.	Rec.	F1
Text Only	BiLSTM-CRF-None	16.45	14.08	15.17
	BERT-None (Kenton and Toutanova, 2019)	19.96	20.53	20.24
	BERT-CRF-None	19.86	21.82	20.79
	BARTNER-None (Yan et al., 2021)	20.30	22.92	21.53
Text+Image	UMT-RCNN-EVG (Yu et al., 2020)	32.47	33.91	33.18
	UMT-VinVL-EVG (Yu et al., 2020)	38.14	39.82	38.96
	UMGF-VinVL-EVG (Zhang et al., 2021a)	37.70	39.89	38.76
	ITA-VinVL-EVG (Wang et al., 2022a)	38.85	40.76	39.78
	BARTNER-VinVL-EVG (Yu et al., 2023)	34.08	39.76	36.70
	H-Index (Yu et al., 2023)	41.45	43.37	42.38
	SeqGMPER-None (Ours)	40.20	40.86	40.53
	SeqGMPER (Ours)	**44.86**	**43.74**	**44.28**

Table 2: Ablation experiments (%) on dataset FewRel

Visualization Study



To intuitively explain the effectiveness of our proposed model, we conduct the case studies on GMPER task for H-Index (Yu et al., 2023) and our proposed model SeqGMPER. As shown in the above Figure, we can observe that both SeqGMPER and HIndex can correctly recognize the procedural entity "strawberry" in step 1. However, as the shape of "strawberry" changes in the following steps (i.e., step 2, 3 and 4), H-Index gradually fails to localize its bounding boxes.

Contributions

1. We explore a new problem named Grounded Multimodal Procedural Entity Recognition (GMPER), aiming to automatically recognize textual procedural entities and link the corresponding visual regions in images from multimodal procedural documents.
2. We design a textual and visual sequential feature fusion method to capture the state changes of entities as the sequence or procedure progresses, which effectively assist the detection of both textual and visual entities from multimodal procedural documents.
3. Extensive experiments and visualization analysis are conducted.

Reference

[1] Jianfei Yu, Jing Jiang, Li Yang, and Rui Xia. 2020. Improving multimodal named entity recognition via entity span detection with unified multimodal transformer. In Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics, pages 3342–3352.