Exophoric Pronoun Resolution in Dialogues with Topic Regularization

Xintong Yu[1], **Hongming Zhang**[2], **Yangqiu Song**[2], **Changshui Zhang**[1],
Kun Xu[3], **and Dong Yu**[3]
[1]Institute for Artificial Intelligence, Tsinghua University (THUAI);
[1]Department of Automation, Tsinghua University, Beijing, P.R.China
[2]Department of CSE, The Hong Kong University of Science and Technology
[3]Tecent AI lab
yuxt16@mails.tsinghua.edu.cn, zcs@mail.tsinghua.edu.cn,
{hzhangal, yqsong}@cse.ust.hk, {kxkunxu, dyu}@tencent.com

Abstract

Resolving pronouns to their referents has long been studied as a fundamental natural language understanding problem. Previous works on pronoun coreference resolution (PCR) mostly focus on resolving pronouns to mentions in text while ignoring the exophoric scenario. Exophoric pronouns are common in daily communications, where speakers may directly use pronouns to refer to some objects present in the environment without introducing the objects first. Although such objects are not mentioned in the dialogue text, they can often be disambiguated by the general topics of the dialogue. Motivated by this, we propose to jointly leverage the local context and global topics of dialogues to solve the out-of-text PCR problem. Extensive experiments demonstrate the effectiveness of adding topic regularization for resolving exophoric pronouns.

1 Introduction

Grounding pronouns to objects they refer to is a challenging yet crucial natural language understanding problem. The coreference relationship between a pronoun and its referents is categorized into *endophora* and *exophora* based on whether the referred objects appear in text or out of text, and the former case can be further divided into *anaphora* if the referents appear in the preceding text of the pronoun and *cataphora* if in the following text (Halliday and Hasan, 1976; Brown and Yule, 1983). Conventional studies on the pronoun coreference resolution (PCR) task in the NLP community mainly focus on anaphora (Hobbs, 1978; NIST, 2003; Pradhan et al., 2012) and some recent work analyzes cataphora in machine translation (Wong et al., 2020), while mostly ignoring the exophoric pronouns. However, in daily dialogues or conversations, speakers may often use exophoric pronouns to refer to objects in the situational context that all speakers and listeners are aware of without introducing them in the first place. This limits the use of current PCR models in many real-world dialogue/conversation scenarios, e.g., text interpretation (Hankamer and Sag, 1976; Yule, 1979) and downstream tasks such as dialogue generation (Kottur et al., 2018; Niu et al., 2019).

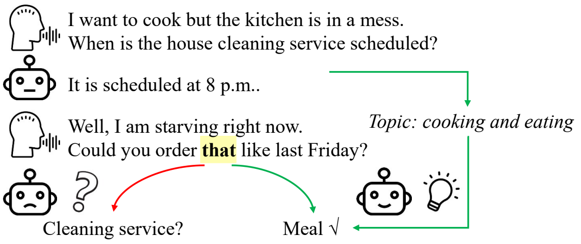

I want to cook but the kitchen is in a mess.
When is the house cleaning service scheduled?
It is scheduled at 8 p.m..
Well, I am starving right now.
Could you order **that** like last Friday?
Topic: cooking and eating
Cleaning service?
Meal √

Figure 1: An example of resolving **exophoric pronouns** in daily dialogues with and without the help of dialogue topics.

Figure 1 shows an example of exophora. A person talks with his AI assistant (Siri/Alexa), "Could you order that like last Friday?" In this scenario, "that" is an exophoric pronoun whose referent can not be found in the dialogue text. A smart enough AI system should be able to resolve the pronoun "that" to some food rather than cleaning service based on the context. Such resolution of exophora is a crucial step in natural language understanding for the AI dialogue system to generate meaningful and relevant responses.

Since traditional PCR tasks only focus on endophoric pronouns while ignoring exophoric ones, all existing models struggle when the correct referent is not in the textual context of the target pronoun. For example, most of the human-defined rules (Hobbs, 1978) (e.g., "them" can only refer to plural objects) and features (Ng, 2005) (e.g., the distance between the target pronoun and candidate noun phrase) become either less effective or inapplicable in the exophoric setting. Unlike human-designed patterns or feature-based methods, the end-to-end coreference models (Lee et al., 2018; Joshi et al., 2019) have the potential of resolving

Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing, pages 3832–3845
November 7–11, 2021. ©2021 Association for Computational Linguistics

pronouns to external objects as long as the names of objects are provided as candidates. Nonetheless, these models heavily rely on the representation of local context produced by deep models so they always tend to resolve pronouns to the mentions in near text. As Figure 1 shows, the models could easily be distracted by the noun phrase "cleaning service" in text and resolve "that" to the service.

To address the limitations of current models, we propose to take the overall dialogue topics into consideration. For the example in Figure 1, we can judge from the whole dialogue that the topic is about cooking and eating, so it is likely that the person needs some food. If the AI system correctly resolves "that" to the topic-related out-of-text object "meal," this may help the AI assistant to finally give a reasonable response, "I will order the take-away that you had last Friday."

To quantitatively define and evaluate exophora resolution, we leverage the VisPro dataset (Yu et al., 2019), which annotates PCR information on visual dialogues. It is the only PCR dataset with annotations of out-of-text referents to the best of our knowledge. While the original dataset provides images alongside dialogues, we observe that humans can resolve 96% of exophoric pronouns in VisPro with only dialogue texts, which perfectly matches our research goal. Therefore, we perform out-of-text PCR experiments on the texts of VisPro.

In this paper, we define the out-of-text PCR task and present a model, which jointly leverages the local context and global topics to better resolve pronouns to out-of-text objects. The model first identifies the overall dialogue topics and then assign larger scores to objects which are more relevant to the topics. By doing so, it less overfits the local context and learns to resolve pronouns based on global topics. Experimental results prove that the proposed model can significantly boost the performance of resolving exophoric pronouns without sacrificing the performance on in-text PCR. We also conduct an extensive analysis to show the contribution of different components. The data, code, and models are available at: `https://github.com/HKUST-KnowComp/Exo-PCR`.

2 Related Works

Coreference resolution is the task of identifying coreference relations among different mentions. As a vital natural language understanding component, a good coreference system could benefit many downstream tasks such as machine translation (Guillou, 2012; Wong et al., 2020), dialog systems (Strube and Müller, 2003), question answering (Dasigi et al., 2019), and summarization (Steinberger et al., 2007). Due to the weak semantic meaning of pronouns (Ehrlich, 1981), grounding pronouns to their referents (PCR) has been specially studied as a more challenging task than the general coreference resolution (Mitkov, 1998; Ng, 2005).

Previous PCR studies (Ng, 2005; Zhang et al., 2019) mostly focus on resolving pronouns to mentions in the near context. However, in informal text such as daily dialogues, it is common that pronouns may refer to out-of-text objects, which is crucial for dialogue understanding. Such pronouns have long been discussed as "pragmatically controlled anaphora" in linguistics (Hankamer and Sag, 1976; Yule, 1979; Brown and Yule, 1983), but there has been few discussion of exophoric pronouns in the NLP community. Hangyo et al. (2013) deal with exophora of zero pronouns, a special phenomenon in Japanese where an omitted argument of a predicate might refer to the "author" or the "reader" of the document. Aktacs et al. (2018) qualitatively analyze exophoric reference in twitter conversations, where the antecedent of a pronoun could appear in the attached media or the quoted tweet. Unlike previous works, we follow a more general linguistics definition of exohpora (Halliday and Hasan, 1976) and evaluate it quantitatively. One recent work (Yu et al., 2019) annotates a dataset VisPro containing in-text and out-of-text referents for pronouns in Visual Dialog (Das et al., 2017), and solve the PCR task by involving visual information. In this work, we propose to resolve exophora in VisPro with texts as the only input. Our model jointly uses local context and global topic information for exophora resolution, which does not require the support of visual signals and thus can be applied to all scenarios.

3 The Task

In this section, we introduce details about the dataset construction and the task definition.

3.1 Dataset Setting

We construct the exophoric PCR dataset on top of VisPro (Yu et al., 2019), which is the only dataset that provides rich exophoric pronoun annotations to the best of our knowledge. Although the original re-

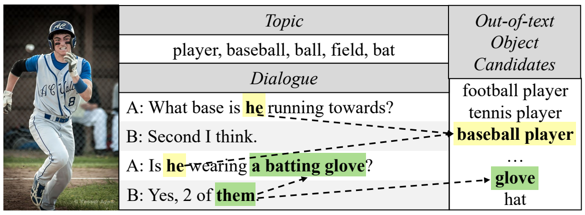

Topic
player, baseball, ball, field, bat
Dialogue
A: What base is he running towards?
B: Second I think.
A: Is he ~~wearing~~ a batting glove?
B: Yes, 2 of them.
Out-of-text Object Candidates
football player
tennis player
baseball player
...
glove
hat

Figure 2: An example of the task. Pronouns are linked with their in-text and out-of-text referents. Exophoric pronouns, endophoric pronouns, and their referents are marked with different colors. The topic words are predicted by an LDA model.

search focus of VisPro is to study the importance of visual information in resolving pronouns in visual-related dialogues, we observe that in many cases, the dialogue text is enough for humans to make the correct resolution. Take Figure 2 as an example. In the dialogue text, the pronoun “he” is exophoric because the referred person is not mentioned explicitly in the dialogue. Even without the image, we can still guess that the dialogue is about a baseball game from clues like “base” and “batting glove,” and thus the pronoun “he” is more likely to refer to “baseball player” rather than other candidates.

Quantitatively, we randomly select 100 exophoric pronouns in the development set of VisPro and find that 96% of them can be correctly resolved without the visual information. Therefore, VisPro can be used as a valid dataset for the exophoric pronoun resolution task. A more detailed analysis is provided in Appendix A.

3.2 Task definition

In this work, we focus on resolving pronouns to mentions inside dialogues and objects outside dialogues simultaneously.

Given a pronoun p in a dialogue $\mathcal{D}$, we first select the noun phrases previous to p in dialogue as candidates $\mathcal{M}$ for in-text referents. For example, the noun phrases “base” and “a batting glove” are candidates of antecedents for “them” in Figure 2.

To provide candidates for out-of-text referents for each dialogue, (Yu et al., 2019) randomly selects 30 noun phrases from image captions. However, such a setting is impractical when no caption is available (details are discussed in Appendix A). As exophoric pronouns may refer to any object in daily life, we collect all the objects that frequently appear in the situational context of dialogues in VisPro to form an object pool $\mathcal{O}$. The object pool contains 384 common object categories such as “hat” and “glove” shown in Figure 2. The details of the collection are described in Sec 5.4.

The goal of the task is to identify the correct antecedents in $\mathcal{M}$ and the correct out-of-text objects from $\mathcal{O}$ by minimizing the loss:

$$\mathcal{L}_{crf} = \mathcal{L}_i + \mathcal{L}_o, \tag{1}$$

where $\mathcal{L}_i$ is the loss function for the in-text coreference resolution and $\mathcal{L}_o$ for the out-of-text resolution. We then define them following the coreferenc loss in the end-to-end in-text coreference models (Lee et al., 2018):

$$\begin{aligned} \mathcal{L}_i &= -\log \frac{\sum_{c \in \mathcal{C}_m} e^{F(p,c;\mathcal{D})}}{\sum_{m \in \mathcal{M}} e^{F(p,m;\mathcal{D})}}, \\ \mathcal{L}_o &= -\log \frac{\sum_{c \in \mathcal{C}_o} e^{F(p,c;\mathcal{D})}}{\sum_{o \in \mathcal{O}} e^{F(p,o;\mathcal{D})}}, \end{aligned} \tag{2}$$

in which $F(\cdot)$ is the coreference score of pronouns p with mentions m or objects o, and $\mathcal{C}_m$ and $\mathcal{C}_o$ denote the correct referents in $\mathcal{M}$ and $\mathcal{O}$, respectively. For instance, for the pronoun “them” in Figure 2, the model is required to not only recognize its antecedent in text to be “a batting glove” but also link it to “glove” in the external object pool.

4 The Model

The goal of the coreference model is to provide the coreference score $F(p, d)$ between a pronoun p and a candidate d, which can either be a mention $m \in \mathcal{M}$ or an external object $o \in \mathcal{O}$. We divide the coreference score into three parts: the similarity score between p and d based on local context, the global topic relevance score of p, and that of d:

$$F(p, d) = F_l(p, d) + F_g(p) + F_g(d). \tag{3}$$

Specifically, F_l calculates the similarity between p and d via local context representations, while F_g acquires the relevance score between each text span and the global topics.

To capture the topic information of the dialogues, we employ topic prediction as an auxiliary task of the model. The overall model architecture is shown in Figure 3 and details are as follows.

4.1 Local Similarity Score

Following (Joshi et al., 2019; Lee et al., 2018; Bahdanau et al., 2015), for each span s, which could be either p, m, or o and contains T words $x_1, x_2, ..., x_T$, we first extract word embeddings from pre-trained language models as

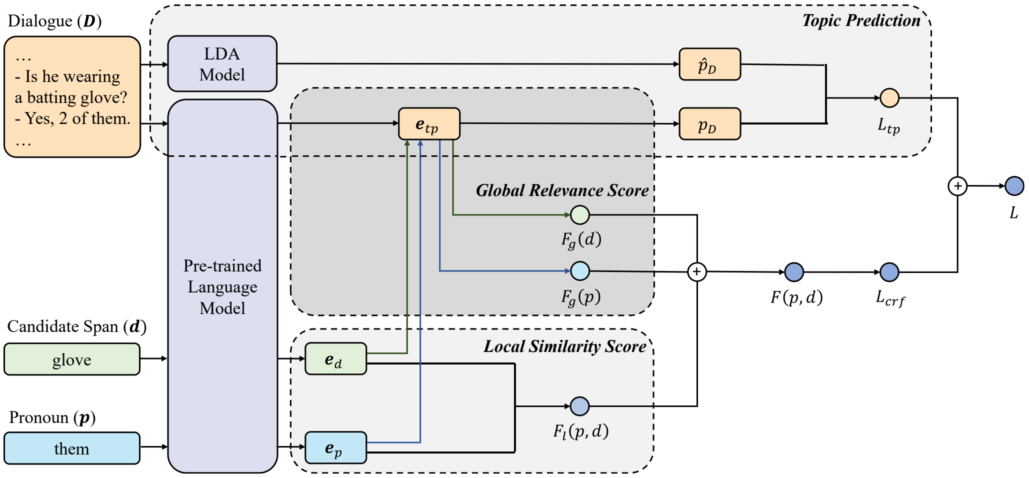


Figure 3: There are three main components in the proposed model: local similarity score calculation, global relevance score calculation, and topic prediction. The local score module calculates the similarity between a pronoun p and a candidate span d based on their textual representation. The global score module measures their relevance with the global dialogue topic. To help the topic embedding capture the topic information better, the topic prediction module uses the dialogue embedding to fit the topic vector predicted by LDA as an auxiliary task.

$\{\mathbf{x}_1, \mathbf{x}_2, ..., \mathbf{x}_T\}$. Then, we represent each span with the combination of the embeddings of the first token ($\mathbf{x}_1$), the last token ($\mathbf{x}_T$), the weighted sum of embeddings of all tokens in it ($\hat{\mathbf{x}}$), and the length feature of the span ($\phi(s)$):

$$\mathbf{e}_s = [\mathbf{x}_1, \mathbf{x}_T, \hat{\mathbf{x}}, \phi(s)] \,, \tag{4}$$

in which

$$\begin{aligned} \hat{\mathbf{x}} &= \sum_{t=1}^{T} \alpha_t \cdot \mathbf{x}_t, \\ \alpha_t &= \frac{\exp(\mathrm{NN}_\alpha(\mathbf{x}_t))}{\sum_{t=1}^{T} \exp(\mathrm{NN}_\alpha(\mathbf{x}_t))}. \end{aligned} \tag{5}$$

Here $[\cdot, \cdot]$ indicates the concatenation operation and NN the feed forward neural network.

After acquiring the features of the spans, we then calculate the local similarity score between a pronoun p and a candidate span d as:

$$F_l(p, d) = \mathrm{NN}_r\left([\mathbf{e}_p, \mathbf{e}_d, \mathbf{e}_p \odot \mathbf{e}_d]\right). \tag{6}$$

where $\odot$ denotes the element-wise multiplication.

4.2 Global Relevance Score

Although the out-of-text referents of exophoric pronouns are not mentioned in the text, they can be inferred from the dialogue context. As the subject of dialogue context, the dialogue topics play a vital part in exophora resolution. For the daily dialogue example in Figure 1, we can infer from context words such as "cook," "kitchen," and "starving" that the dialogue topic is about cooking and eating, so the exophoric pronoun "that" is more likely to refer to "meal" rather than "cleaning service."

Similarly, in the VisPro example in Figure 2, if we only read the sentence containing "he," it is hard to infer the targeting object of "he" to be a baseball player, a tennis player, or a football player. On the contrary, if we consider the whole dialogue as context, we can recognize the topic to be a baseball game, in which a man "wearing a batting glove" is "running towards" a "base." Therefore, we can judge that this man must be a baseball player rather than a football or tennis player, so the exophoric pronoun "he" refers to the out-of-text object "baseball player."

Based on the above observations, we leverage the overall dialogue topic to help grounding pronouns to out-of-text objects. To effectively encode the topic information of the whole dialogue, we first obtain the overall embedding $\mathbf{e}_\mathcal{D}$ of a dialogue $\mathcal{D}$ with pre-trained language models. For LSTM-based models, we take the average embedding of all sentences as $\mathbf{e}_\mathcal{D}$. For BERT-based models, we take the embedding of the special token [CLS]. Then we pass it through a feed forward neural network to obtain the dialogue topic embedding:

$$\mathbf{e}_{tp} = \mathrm{NN}_{tp}(\mathbf{e}_\mathcal{D}). \tag{7}$$

After that, to indicate the relevance between a span s and the global topic of the dialogue, we calculate the topic relevance score as:

$$F_g(s) = \mathrm{NN}_g\left([\mathbf{e}_{tp}, \mathbf{e}_s, \mathbf{e}_{tp} \odot \mathbf{e}_s]\right). \tag{8}$$

No.	LDA Topic Words	Summarized Topic
15	car, street, sign, road, vehicle	cars in streets
16	tree, grass, fence, animal, leaf	animals on grass
23	player, baseball, ball, field, bat	baseball game
25	kitchen, food, cut, stove, pot	kitchen
28	orange, banana, fruit, store, apple	fruit

Table 1: Example topics with five topic words extracted by the LDA model on the VisPro training set with $n_{tp} = 40$. The last column presents topics summarized by human reading the extracted topic words.

In the end, we calculate the final coreference scores of pronouns p with in-text mentions m and out-of-text objects o as:

$$\begin{aligned} F(p,m) &= F_l(p,m) + F_g(p) + F_g(m), \\ F(p,o) &= F_l(p,o) + F_g(p) + F_g(o). \end{aligned} \tag{9}$$

With global relevance scores, models trained with VisPro are able to resolve exophoric pronouns based on dialogue topics. In real-life scenarios such as Figure 1, the key for understanding exophora is also the relevance between out-of-text objects and dialogue context. Thus the ability to resolve exophora with dialogue topics can also be transferred to such realistic cases.

4.3 Topic Prediction as Regularization

To help the topic embedding $\mathbf{e}_{tp}$ better represent the topic information of the dialogue, we propose to use topic prediction as an auxiliary task.

We first obtain the topic labels of dialogues by the most commonly used unsupervised topic model Latent Dirichlet Allocation (LDA) (Blei et al., 2001). The LDA model extracts n_{tp} topics from dialogues in the training set and represents each topic as a list of words with a high probability to appear under the topic. Table 1 presents some topics of VisPro dialogues extracted by the LDA model. From the topic words, we can summarize that the No.15 topic is about cars in streets and that the No.25 topic discusses a kitchen. The topic label $\hat{\mathbf{p}}_\mathcal{D}$ of a dialogue $\mathcal{D}$ can be defined as:

$$\hat{\mathbf{p}}_\mathcal{D} = \text{LDA}(\mathcal{D}) \in \mathbb{R}^{n_{tp}}, \tag{10}$$

where the j^{th} dimension of $\hat{\mathbf{p}}_\mathcal{D}$ represents the probability of the dialogue corresponding to the No.j topic. For instance, the LDA model predicts that the dialogue in Figure 2 belongs to the No.23 topic in Table 1 with 60% possibility and thus the 23^{th} dimension of $\hat{\mathbf{p}}_\mathcal{D}$ is 0.6.

As $\hat{\mathbf{p}}_\mathcal{D}$ sums up to 1 and each dialogue could associate with several topics, we fit $\hat{\mathbf{p}}_\mathcal{D}$ by $\mathbf{e}_{tp}$ with a L2 loss after a softmax function[1]:

$$\begin{aligned} \mathbf{p}_\mathcal{D} &= \text{softmax}\left(\text{NN}_p(\mathbf{e}_{tp})\right), \\ \mathcal{L}_{tp} &= \frac{1}{2}||\mathbf{p}_\mathcal{D} - \hat{\mathbf{p}}_\mathcal{D}||_2^2. \end{aligned} \tag{11}$$

We use the topic prediction loss as a regularization term to the total loss:

$$\mathcal{L} = \mathcal{L}_{crf} + \mathcal{L}_{tp} = \mathcal{L}_i + \mathcal{L}_o + \mathcal{L}_{tp}, \tag{12}$$

where $\mathcal{L}_i$ and $\mathcal{L}_o$ are defined in (2). As a result, the final loss function $\mathcal{L}$ can be optimized in an end-to-end manner.

5 The Experiment

In this section, we introduce the experiment details.

5.1 Dataset

We use VisPro (Yu et al., 2019) as the dataset, which contains 4,000 train, 500 development, and 500 test dialogues. The train, development, and test sets of VisPro contain 13,686, 1,726, and 1,781 pronouns with out-of-text referents and 13,986, 1,742, and 1,756 pronouns with in-text antecedents, respectively.

5.2 Evaluation Metrics

We use different metrics for in-text and out-of-text PCR due to the different numbers of candidates. For the in-text PCR, each pronoun has 10.3 candidates and 1.6 correct referents on average. Thus we follow the previous work (Yu et al., 2019) to employ Precision (P), Recall (R), and F1 score as the evaluation metrics. For the out-of-text PCR, as all 384 common object nouns are candidates and only one of them is correct, the F1 score is no longer suitable. For example, if the model predicts the correct answer to be the second place out of 384 candidates, it means that model can somehow understand the pronoun, while the F1 metric will count it as wrong. Therefore, we view out-of-text PCR as a ranking problem, where objects that a pronoun refers to should have a higher rank, and evaluate all models by the recall at 1, 5, and 10.

5.3 Baselines

We add our global relevance score module and topic prediction module on basis of the following

[1] We also tried other loss functions, such as KL-divergence between $\mathbf{p}_\mathcal{D}$ and $\hat{\mathbf{p}}_\mathcal{D}$, and cross entropy loss after a sigmoid function for each dimension of $\text{NN}_p(\mathbf{e}_{tp})$. Empirical studies show that the L2 loss achieves the best performance.

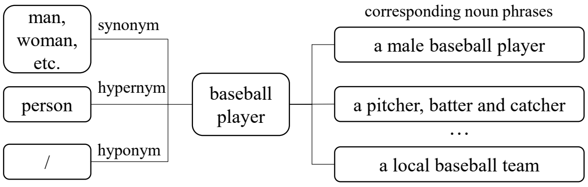


Figure 4: An example of the object category "baseball player." Each object category contains its synonyms, hypernyms, hyponyms, and corresponding noun phrases.

end-to-end coreference resolution models which only contains the local similarity score module[2]:

- End-to-end model with **LSTM** based on **ELMo** embedding (Lee et al., 2018) , which extracts features by a BiLSTM upon ELMo embeddings.
- End-to-end model with **BERT** embedding (Joshi et al., 2019).
- End-to-end model based on **SpanBERT** embedding (Joshi et al., 2020) , which can better represent text spans.

5.4 Implementation

Dataset Processing: To collect common object categories in VisPro, we first map 2,600 noun phrases annotated as out-of-text referents in VisPro to a compact list of 384 object categories by removing all modifiers and merging similar phrases. For instance, pronouns referring to "a male baseball player" or "a local baseball team" are both mapped to the object "baseball player." Moreover, some objects have similar or overlapping meanings with other objects (e.g., "pond" similar to "pool") but only one is labeled as the gold answer of a pronoun. It would be problematic if we directly label all others as wrong. To solve this problem, we use the synonyms, hypernyms, and hyponyms obtained from synset in Wordnet (Miller, 1995) in NLTK (Bird, 2006) as extra information attached to each object category. If a pronoun refers to a particular object in the external object pool, then the synonyms, hypernyms, and hyponyms of the targeting object are masked during the training and testing process. An example of an object category "baseball player" is shown in Figure 4. Note that other person categories which are not a synonym, hypernym or hyponym of "baseball player", such as "tennis player" and "football player", are not masked.

Last but not least, we split the pronouns with out-of-text referents by whether a pronoun simultaneously refers to some mentions in the dialogue. If a pronoun has both in-text and out-of-text referents, such as "them" in Figure 1, which refers to "a batting glove" in the dialogue as well as "glove" in the object pool, we denote it as "Discussed" in the dialogue. If a pronoun only has out-of-text referents, such as "he" in Figure 1, which only refers to the object "baseball player," we denote it as "Not Discussed" in the dialogue. While "Not Discussed" pronouns strictly match the definition of exophora, grounding the "Discussed" pronoun to out-of-text objects is also an important step towards linking dialogue text to the environment. In VisPro, 25.02% of all pronouns with out-of-text referents are "not discussed."

Training Details: We follow the hyperparameters set in (Joshi et al., 2019). The number of topics n_{tp} is set to 40 for LDA. The topic prediction module in the model contains one hidden layer of size 1,000. Gold mentions are provided for training and testing of the models. During testing, the in-text antecedents are chosen in the same way as (Lee et al., 2018). For the out-of-text part, objects o with scores $F(p, o) > 0$ are deemed as the prediction of out-of-text referents for the pronoun p and the selected objects are ranked according to the scores. Models are trained for ten epochs, and the best ones are selected based on their performance on the development set.

6 The Results

From the experimental results in Table 2, we can observe that BERT and SpanBERT based models outperform ELMo-LSTM based models, which is consistent with the observation in (Joshi et al., 2019) mainly because of their stronger context representation ability. On top of them, incorporating global topics improves recall for both exophoric and endophoric pronouns. Last but not least, for in-text PCR, adding topic information only slightly influences the precision while significantly improving the recall. As a result, it also achieves better overall F1 performance.

Further analyzing the performances of models on out-of-text PCR, we observe that the "Not Dis-

[2] We do not compare with CorefQA (Wu et al., 2020) because it selects in-text antecedents as a reading comprehension task, which cannot be applied to out-of-text objects. We do not compare with VisCoref (Yu et al., 2019) because it requires images as input, while our setting is text-only.

Model	Out-of-text PCR						In-text PCR		
	Not Discussed			Discussed					
	R@1	R@5	R@10	R@1	R@5	R@10	P	R	F1
ELMo-LSTM	61.54	66.19	66.80	70.86	71.25	71.25	88.15	66.05	75.51
+ topic (ours)	68.02	71.66	72.06	72.49	72.96	72.96	87.55	70.43	78.06
BERT-base	87.45	89.68	90.49	89.74	94.64	94.79	86.51	80.63	83.47
+ topic (ours)	**90.49**	**93.72**	**95.75**	92.46	96.43	96.89	85.79	83.66	84.72
SpanBERT-base	87.65	92.11	92.51	91.38	94.25	94.79	**89.08**	79.35	83.94
+ topic (ours)	90.28	93.32	93.93	**93.63**	**96.50**	**97.05**	83.97	**85.78**	**84.87**

Table 2: Results of experiments for out-of-text PCR evaluated by Recall (R) in the top 1, 5, and 10 predictions and in-text PCR measured by Precision (P), Recall (R), and F1 score. The best results are shown in **bold** font.

Model	Object Type	R@1	R@5	R@10
BERT-base	Infrequent	39.66	51.72	53.45
	Frequent	93.81	94.72	95.41
+ topic	Infrequent	46.55	65.52	74.14
	Frequent	96.33	97.48	98.62

Table 3: Recall of "Not Discussed" pronouns in the test set referring to "Infrequent" and "Frequent" objects.

cussed" pronouns are more challenging than the "Discussed" group for all models. This makes sense because if a pronoun refers to some noun phrases in text, the embedding of the pronoun will encode the information of those noun phrases via the language models. For instance, if the representation of "them" in Figure 2 encodes the context "a batting glove," it would be easier to identify the semantically related object "glove" as the out-of-text referent. In contrast, "Not Discussed" pronouns do not have any noun phrase antecedent in the dialogue and are thus more challenging. In such cases, the effect of incorporating global semantics becomes more significant than in "Discussed" cases. In the rest of this section, we present a detailed analysis with the BERT-base + topic model, which achieves the highest performance on "Not Discussed" pronouns and comparable performances on other settings, to show when our model performs well and when it fails.

6.1 Influence of Frequency

In the external object pool, the appearances of different objects varies. For instance, "man" appears 3,084 times in the training set, while "monkey" only appears once. To investigate the influence of such imbalance, we split the object list by their occurrence frequency, with occurrence less than 50 times as "Infrequent" objects, which make up 85.1% of list, and the rest as "Frequent" objects.

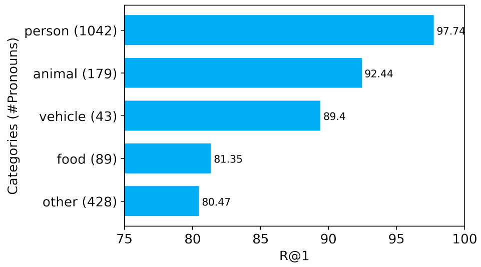


Figure 5: Performance and number of pronouns in the test set related to different out-of-text object categories.

As observed in Table 3, performances on infrequent objects are much lower than frequent ones, which indicates that although the models achieve high scores on frequent objects, they still fail to do well on the majority of relatively rare objects. This observation also shows that the exophoric PCR problem is still far from being solved. Compared to models focusing on local information, the proposed model, which incorporates the overall topics, boosts the performance by a large margin, especially on infrequent pronouns.

6.2 Influence of Object Categories

Besides the influence of frequency, we are also interested in how well our model can perform on different object categories. We record the performance of pronouns related to the four most common categories[3] (person, animal, vehicle, and food) in Figure 5, from which we can see that pronouns related to "person" and "animal" are most common and easiest to be resolved, which is consistent with our previous observation that our model performs better on frequent objects than on infrequent ones.

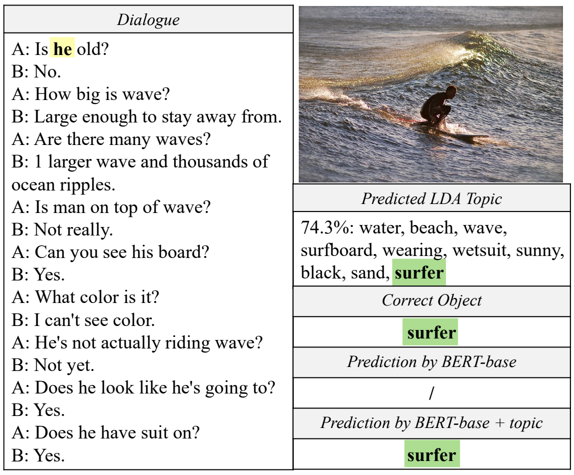
[3] Here a pronoun is deemed as related to a major category if the object it refers to is exactly that category or a hyponym of the category. For example, pronouns linked to "person" or "man" are both considered related to "person." We also report the number of related pronouns in the test set.

	Out-of-text R@1	Out-of-text ΔR@1	In-text F1	In-text ΔF1
Our full model	90.49	-	84.72	-
- topic prediction	88.46	-2.02	84.08	-0.63
- masking synonyms	56.88	-33.60	84.04	-0.67
- in-text training	87.25	-3.24	25.12	-59.60
- out-of-text training	48.99	-41.50	82.47	-2.24

Table 4: Ablation study results.

Model	Out-of-text Not Discussed R@1	Out-of-text Discussed R@1	In-text F1
BERT-base	87.45	89.74	83.47
+ topic	**90.49**	92.46	84.72
BERT-large	87.25	90.83	84.62
+ topic	88.46	92.00	85.08
SpanBERT-base	87.65	91.38	83.94
+ topic	90.28	**93.63**	84.87
SpanBERT-large	87.65	91.61	86.64
+ topic	89.68	93.40	**87.22**

Table 5: Performance comparison among BERT-base, BERT-large, SpanBERT-base, and SpanBERT-large embeddings. The best results are in **bold** font.



Dialogue

A: Is **he** old?
B: No.
A: How big is wave?
B: Large enough to stay away from.
A: Are there many waves?
B: 1 larger wave and thousands of ocean ripples.
A: Is man on top of wave?
B: Not really.
A: Can you see his board?
B: Yes.
A: What color is it?
B: I can't see color.
A: He's not actually riding wave?
B: Not yet.
A: Does he look like he's going to?
B: Yes.
A: Does he have suit on?
B: Yes.

Predicted LDA Topic

74.3%: water, beach, wave, surfboard, wearing, wetsuit, sunny, black, sand, **surfer**

Correct Object

surfer

Prediction by BERT-base

/

Prediction by BERT-base + topic

surfer

Figure 6: Case study for out-of-text PCR. Target pronouns and correct out-of-text objects with their hints are marked in different colors. Note that we only show the corresponding images here for clarity and that they are **not** provided to the models.

6.3 Ablation Study

We present the ablation study in Table 4, from which we can see that all components contribute to the ultimate success. For example, performance drops when removing the topic prediction loss as regularization, which indicates that the topic prediction module can help the embedding of the dialogue to capture the topic information better. Besides that, if we do not mask out the synonyms, hypernyms, and hyponyms of the object categories during training, the performance drops dramatically. It shows the importance of masking possible distractions to provide unique labels during training. Last but not least, one contribution of the proposed model is the joint training of both the in-text and out-of-text PCR and, the results show that removing either of them in the training process will result in a performance drop on both tasks. Similar improvement by joint training is also observed in (Bai et al., 2021), where the in-text PCR task is jointly trained with the character linking task that links the endophoric pronouns in TV show scripts to the characters.

6.4 BERT-base VS BERT-large

Table 5 compares the performance of models based on BERT-base, BERT-large, SpanBERT-base, and SpanBERT-large. Incorporating topic information consistently improves performance on out-of-text PCR for all models while achieving comparable scores on the in-text one. Besides, we surprisingly find out that compared to BERT-base and SpanBERT-base, even though BERT-large and SpanBERT-large achieve higher scores on in-text PCR, their performance on the out-of-text PCR slightly drops. An explanation is that they may easily overfit the local context and ignore the global topic information due to their deep model.

6.5 Case Study

Figure 6 shows a dialogue about a male surfer. The referents of the pronoun "he" is "Not Discussed" in the dialogue text. The model that can only access the local context cannot identify any object related to the pronoun. In contrast, the model with topic prediction assigns a high probability of 74.3% for the topic of the dialogue to be surfing judging from the related words such as "wave" and "board." Thus it identifies "surfer" as the out-of-text referent for the pronoun. More cases are shown in Appendix B.

6.6 Error Analysis

We first quantitatively study the error types of the BERT-base + topic model by randomly selecting 60 mistaken predictions in out-of-text PCR, including 30 cases for the "Not Discussed" pronouns and 30 for the "Discussed" ones. We observe that 1/3 of the cases are also difficult for humans to identify the correct objects without access to the correspond-

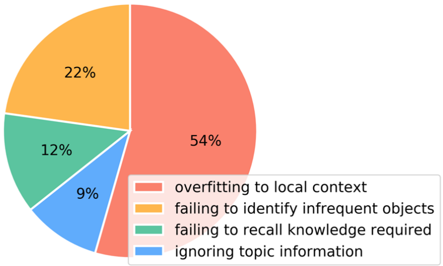


Figure 7: Error distribution in out-of-text PCR.

ing images. This is either because the dialogue text does not contain enough clues to infer the right answer, or multiple answers are reasonable but only one is annotated. For the other 2/3 cases, Figure 7 shows that more than half of errors are still from overfitting to local context and 10% from failure to use the topic information. Other 23% errors come from failure to associate pronouns with infrequent objects as discussed in Section 6.1 and the rest 13% are due to the lack of required knowledge. Error analysis demonstrates that the model can be further improved by avoiding overfitting to the local context and incorporating explicit knowledge. Some erroneous cases are provided in Appendix C.

7 Conclusion

In this paper, we focus on grounding pronouns in dialogues to out-of-text objects. We propose to incorporate the topics of the dialogues to help the PCR model identify the out-of-text referents better. Experiments show that the proposed model outperforms previous models on both in-text and out-of-text PCR tasks. Detailed analysis is presented to show the strength and limitations of the proposed model. While this work is a first step to explore exophora resolution on one dataset, future work may explore exophora resolution in different domains such as AI chat-bots for home assistants.

Acknowledgement

The authors of this paper were supported by the National Key Research and Development Program of China (No. 2018AAA0100701), a grant from the Guoqiang Institute, Tsinghua University, the NSFC Fund (U20B2053) from the NSFC of China, the RIF (R6020-19 and R6021-20) and the GRF (16211520) from RGC of Hong Kong, the MHKJFS (MHP/001/19) from ITC of Hong Kong, and the Tencent AI Lab Rhino-Bird Focused Research Program.

References

Berfin Aktacs, Tatjana Scheffler, and Manfred Stede. 2018. Anaphora resolution for Twitter conversations: An exploratory study. In *Proceedings of the First Workshop on Computational Models of Reference, Anaphora and Coreference*, pages 1–10, New Orleans, Louisiana.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *Proceedings of ICLR 2015*.

Jiaxin Bai, Hongming Zhang, Yangqiu Song, and Kun Xu. 2021. Joint coreference resolution and character linking for multiparty conversation. In *EACL*, pages 539–548. Association for Computational Linguistics.

Steven Bird. 2006. NLTK: the natural language toolkit. In *Proceedings of ACL 2006*.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2001. Latent dirichlet allocation. In *Proceedings of NIPS 2001*, pages 601–608.

Gillian Brown and George Yule. 1983. *Discourse analysis*. Cambridge university press.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2017. Visual dialog. In *Proceedings of CVPR 2017*, pages 1080–1089.

Pradeep Dasigi, Nelson F. Liu, Ana Marasovic, Noah A. Smith, and Matt Gardner. 2019. Quoref: A reading comprehension dataset with questions requiring coreferential reasoning. In *Proceedings of EMNLP-IJCNLP*, pages 5924–5931. Association for Computational Linguistics.

Kate Ehrlich. 1981. Search and inference strategies in pronoun resolution: an experimental study. In *Proceedings of ACL 1981*, pages 89–93.

Liane Guillou. 2012. Improving pronoun translation for statistical machine translation. In *Proceedings of EACL 2012*, pages 1–10.

M.A.K. Halliday and R. Hasan. 1976. Cohesion in english. *Longman*, pages 18–33.

Masatsugu Hangyo, Daisuke Kawahara, and Sadao Kurohashi. 2013. Japanese zero reference resolution considering exophora and author/reader mentions. In *EMNLP*, pages 924–934. ACL.

Jorge Hankamer and Ivan Sag. 1976. Deep and surface anaphora. *Linguistic inquiry*, 7(3):391–428.

Jerry R Hobbs. 1978. Resolving pronoun references. *Lingua*, 44(4):311–338.

Mandar Joshi, Danqi Chen, Yinhan Liu, Daniel S. Weld, Luke Zettlemoyer, and Omer Levy. 2020. Spanbert: Improving pre-training by representing and predicting spans. *TACL*, 8:64–77.

Mandar Joshi, Omer Levy, Luke Zettlemoyer, and Daniel S. Weld. 2019. BERT for coreference resolution: Baselines and analysis. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5802–5807.

Satwik Kottur, José M. F. Moura, Devi Parikh, Dhruv Batra, and Marcus Rohrbach. 2018. Visual coreference resolution in visual dialog using neural module networks. In *Proceedings of ECCV 2018*, pages 160–178.

Kenton Lee, Luheng He, and Luke Zettlemoyer. 2018. Higher-order coreference resolution with coarse-to-fine inference. In *Proceedings of NAACL-HLT 2018*, pages 687–692.

George A. Miller. 1995. Wordnet: A lexical database for english. *Commun. ACM*, 38(11):39–41.

Ruslan Mitkov. 1998. Robust pronoun resolution with limited knowledge. In *Proceedings of ACL 1998*, pages 869–875.

Vincent Ng. 2005. Supervised ranking for pronoun resolution: Some recent improvements. In *Proceedings of AAAI 2005*, pages 1081–1086.

US NIST. 2003. The ace 2003 evaluation plan. *US National Institute for Standards and Technology (NIST)*, pages 2003–08.

Yulei Niu, Hanwang Zhang, Manli Zhang, Jianhong Zhang, Zhiwu Lu, and Ji-Rong Wen. 2019. Recursive visual attention in visual dialog. In *Proceedings of CVPR 2019*, pages 6679–6688.

Sameer Pradhan, Alessandro Moschitti, Nianwen Xue, Olga Uryupina, and Yuchen Zhang. 2012. Conll-2012 shared task: Modeling multilingual unrestricted coreference in ontonotes. In *Proceedings of EMNLP-CoNLL 2012*, pages 1–40.

Josef Steinberger, Massimo Poesio, Mijail A. Kabadjov, and Karel Jezek. 2007. Two uses of anaphora resolution in summarization. *Inf. Process. Manag.*, 43(6):1663–1680.

Michael Strube and Christoph Müller. 2003. A machine learning approach to pronoun resolution in spoken dialogue. In *Proceedings of ACL 2003*, pages 168–175.

KayYen Wong, Sameen Maruf, and Gholamreza Haffari. 2020. Contextual neural machine translation improves translation of cataphoric pronouns. In *Proceedings of ACL 2020*, pages 5971–5978.

Wei Wu, Fei Wang, Arianna Yuan, Fei Wu, and Jiwei Li. 2020. Corefqa: Coreference resolution as query-based span prediction. In *Proceedings of ACL 2020*, pages 6953–6963.

Xintong Yu, Hongming Zhang, Yangqiu Song, Yan Song, and Changshui Zhang. 2019. What you see is what you get: Visual pronoun coreference resolution in dialogues. In *Proceedings of EMNLP-IJCNLP 2019*, pages 5122–5131.

George Yule. 1979. Pragmatically controlled anaphora. *Lingua*, 49(2-3):127–135.

Hongming Zhang, Yan Song, and Yangqiu Song. 2019. Incorporating context and external knowledge for pronoun coreference resolution. In *Proceedings of NAACL-HLT 2019*, pages 872–881.

A Task Definition Compared to Prior Works

Our experiments are based on the dataset VisPro (Yu et al., 2019), which provides annotation of referents for pronouns in dialogues of the Visual Dialog dataset (Das et al., 2017). Figure 8 illustrates the different settings of our work compared to prior works.

In the original setting of Visual Dialog dataset (Figure 8(a)), each dialogue happens between two people chatting about an image, and each image is accompanied by a descriptive caption. Speaker A only has access to the caption and attempts to imagine the image by asking questions, while speaker B can access both the image and the caption and answers the questions. Thus the pronouns in the dialogues refer to either mention in the dialogue text or noun phrase in captions.

In the setting of VisPro (Figure 8(b)), to simulate the scenario where people use pronouns to directly refer to objects in the environment, the captions are separated from the dialogues. As the captions are descriptions of images, the mentions in captions must correspond to some objects in the images. Thus, when captions are no longer available, the pronouns that refer to noun phrases in captions can be deemed as referring to objects in the images.

Although VisPro first proposed the scenario where pronouns refer to out-of-text objects, it focused on visual-related cases and did not associate such cases with the general definition of exophora. Furthermore, in the definition of the visual pronoun coreference resolution task that Yu et al. (2019) proposed, the candidates of the out-of-text objects are 30 noun phrases randomly selected from captions. This small set of objects contains noun phrases from the corresponding caption as well as captions of other images to provide negative samples. However, such a setting is not so practical. For one thing, the out-of-text candidates are not fixed among different dialogues and the choices for negative samples are random, which makes it hard to compare between multiple dialogues. For another, such noun phrases are hard to obtain in practical cases where no caption for the environment is available, so the model trained under this task cannot be applied to dialogues outside the dataset.

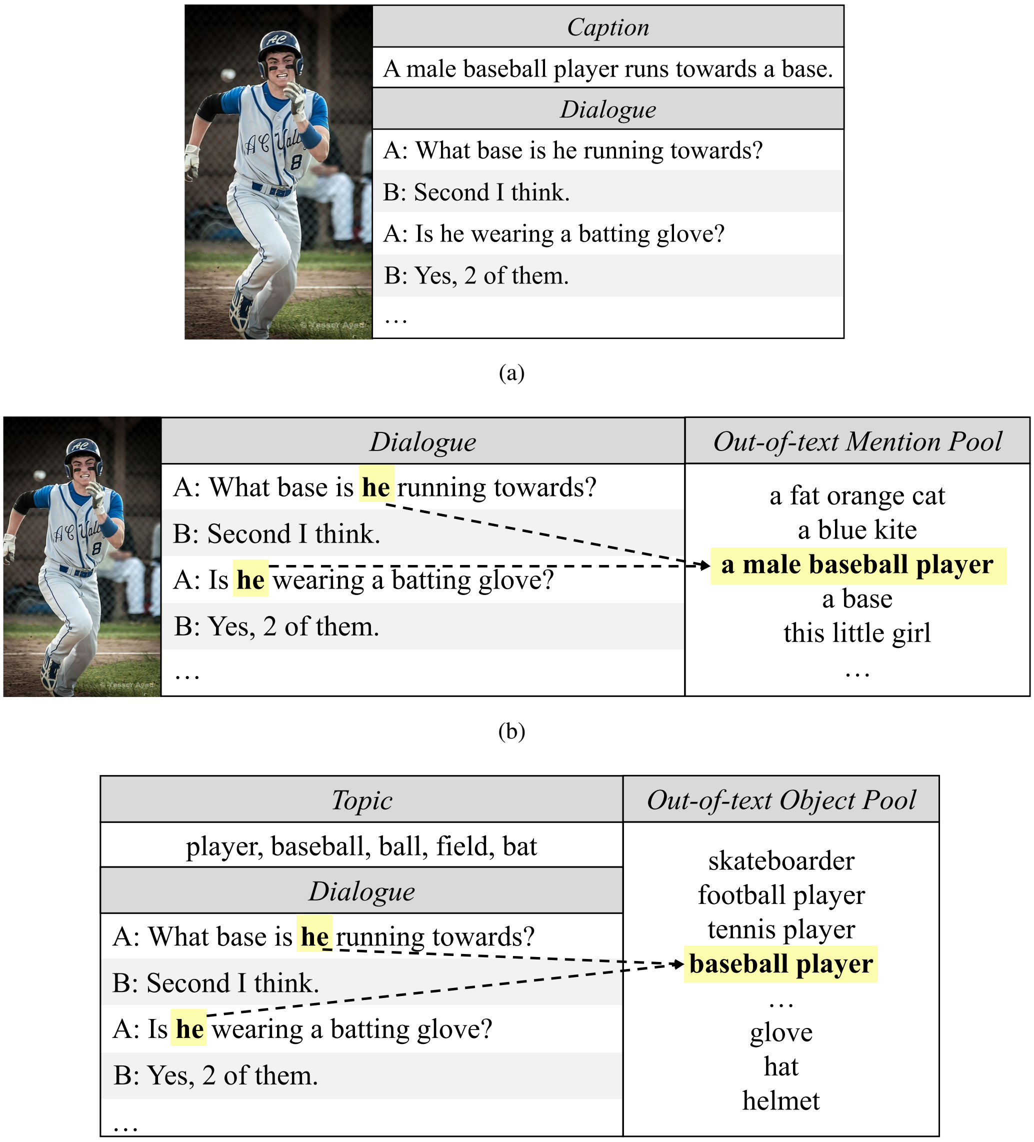
Based on the annotation of VisPro, we design a more practical experiment setting (Figure 8(c)). First, we assume that the visual background of dialogues is not always available, and thus aim to resolve exophoric pronouns based on only the dialogue text. Second, since exophoric pronouns might refer to any object in daily life, we collect all the common objects in VisPro to form a candidate pool of 384 object categories. Since the candidate pool is fixed for all dialogues, we can reasonably compare the performance between different dialogues and models. The model trained under our setting can thus be applied to real-life dialogues.

B Case Study for Out-of-Text PCR

We randomly select some cases from the test split of VisPro and present them in Figure 9. Cases (a)-(d) are "Not Discussed" pronouns which only have out-of-text referents, and cases (e)(f) are "Discussed" pronouns which have both out-of-text and in-text referents. For the "Discussed" pronouns in (e)(f), even though the referred objects are mentioned in the text, the BERT-base model still overfits to distracting words and gives the false prediction "person." On the contrary, our model leverages the topic information and predicts the correct objects.

C Erroneous Case Study for Out-of-Text PCR

Figure 10 presents some typical erroneous cases. In Figure 10(a), the model predicts that "they" refers to "person" instead of "sheep," which hits three error types. First, the topic model correctly infers that the dialogue is about some animals on grass but the coreference model ignores this information. Second, based on the word "sheared" and knowledge that sheep need to be sheared, humans can infer that the pronoun refers to "sheep." However, the model fails to learn such knowledge from the pre-training of the language model. Last, the prediction of "person" indicates that it overfits to the word "people" in dialogue text even though it says that there are 0 people. Figure 10(b) shows another case where the model fails to recall the knowledge that only a person could wear a ring or a watch and thus fail to infer that "he" refers to a person.



Caption
A male baseball player runs towards a base.
Dialogue
A: What base is he running towards?
B: Second I think.
A: Is he wearing a batting glove?
B: Yes, 2 of them.
…

(a)

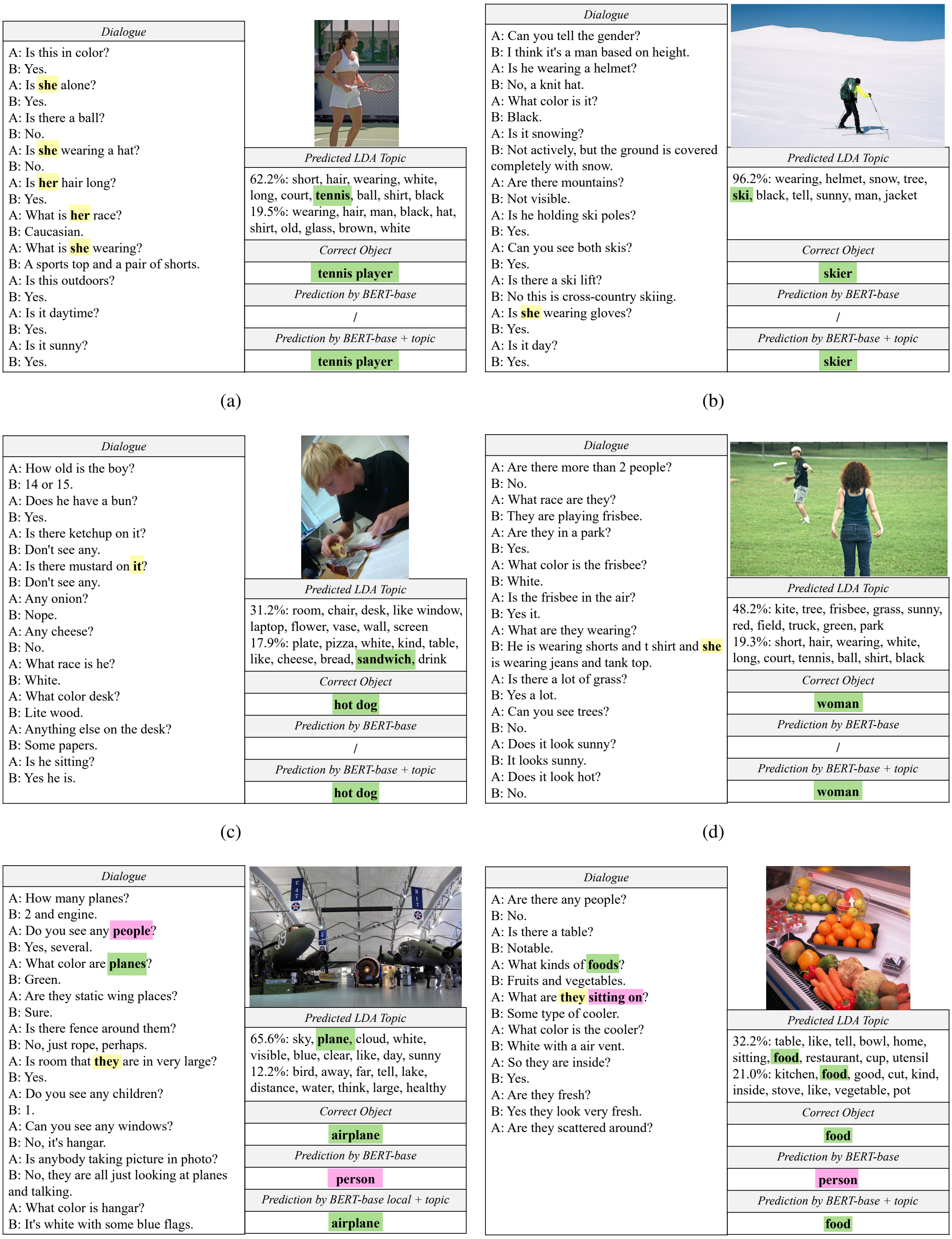
Dialogue	*Out-of-text Mention Pool*
A: What base is **he** running towards?	a fat orange cat
B: Second I think.	a blue kite
A: Is **he** wearing a batting glove?	**a male baseball player**
B: Yes, 2 of them.	a base
…	this little girl …

(b)

Topic	*Out-of-text Object Pool*
player, baseball, ball, field, bat	skateboarder
Dialogue	football player
A: What base is **he** running towards?	tennis player
B: Second I think.	**baseball player**
A: Is **he** wearing a batting glove?	…
B: Yes, 2 of them.	glove hat
…	helmet

(c)

Figure 8: Examples of different settings in (a) Visual Dialog, (b) VisPro, and (c) ours.



(a)

Dialogue

A: Is this in color?
B: Yes.
A: Is **she** alone?
B: Yes.
A: Is there a ball?
B: No.
A: Is **she** wearing a hat?
B: No.
A: Is **her** hair long?
B: Yes.
A: What is **her** race?
B: Caucasian.
A: What is **she** wearing?
B: A sports top and a pair of shorts.
A: Is this outdoors?
B: Yes.
A: Is it daytime?
B: Yes.
A: Is it sunny?
B: Yes.

Predicted LDA Topic

62.2%: short, hair, wearing, white, long, court, **tennis**, ball, shirt, black
19.5%: wearing, hair, man, black, hat, shirt, old, glass, brown, white

Correct Object

tennis player

Prediction by BERT-base

/

Prediction by BERT-base + topic

tennis player

(b)

Dialogue

A: Can you tell the gender?
B: I think it's a man based on height.
A: Is he wearing a helmet?
B: No, a knit hat.
A: What color is it?
B: Black.
A: Is it snowing?
B: Not actively, but the ground is covered completely with snow.
A: Are there mountains?
B: Not visible.
A: Is he holding ski poles?
B: Yes.
A: Can you see both skis?
B: Yes.
A: Is there a ski lift?
B: No this is cross-country skiing.
A: Is **she** wearing gloves?
B: Yes.
A: Is it day?
B: Yes.

Predicted LDA Topic

96.2%: wearing, helmet, snow, tree, **ski**, black, tell, sunny, man, jacket

Correct Object

skier

Prediction by BERT-base

/

Prediction by BERT-base + topic

skier

(c)

Dialogue

A: How old is the boy?
B: 14 or 15.
A: Does he have a bun?
B: Yes.
A: Is there ketchup on it?
B: Don't see any.
A: Is there mustard on **it**?
B: Don't see any.
A: Any onion?
B: Nope.
A: Any cheese?
B: No.
A: What race is he?
B: White.
A: What color desk?
B: Lite wood.
A: Anything else on the desk?
B: Some papers.
A: Is he sitting?
B: Yes he is.

Predicted LDA Topic

31.2%: room, chair, desk, like window, laptop, flower, vase, wall, screen
17.9%: plate, pizza, white, kind, table, like, cheese, bread, **sandwich**, drink

Correct Object

hot dog

Prediction by BERT-base

/

Prediction by BERT-base + topic

hot dog

(d)

Dialogue

A: Are there more than 2 people?
B: No.
A: What race are they?
B: They are playing frisbee.
A: Are they in a park?
B: Yes.
A: What color is the frisbee?
B: White.
A: Is the frisbee in the air?
B: Yes it.
A: What are they wearing?
B: He is wearing shorts and t shirt and **she** is wearing jeans and tank top.
A: Is there a lot of grass?
B: Yes a lot.
A: Can you see trees?
B: No.
A: Does it look sunny?
B: It looks sunny.
A: Does it look hot?
B: No.

Predicted LDA Topic

48.2%: kite, tree, frisbee, grass, sunny, red, field, truck, green, park
19.3%: short, hair, wearing, white, long, court, tennis, ball, shirt, black

Correct Object

woman

Prediction by BERT-base

/

Prediction by BERT-base + topic

woman

(e)

Dialogue

A: How many planes?
B: 2 and engine.
A: Do you see any **people**?
B: Yes, several.
A: What color are **planes**?
B: Green.
A: Are they static wing places?
B: Sure.
A: Is there fence around them?
B: No, just rope, perhaps.
A: Is room that **they** are in very large?
B: Yes.
A: Do you see any children?
B: 1.
A: Can you see any windows?
B: No, it's hangar.
A: Is anybody taking picture in photo?
B: No, they are all just looking at planes and talking.
A: What color is hangar?
B: It's white with some blue flags.

Predicted LDA Topic

65.6%: sky, **plane**, cloud, white, visible, blue, clear, like, day, sunny
12.2%: bird, away, far, tell, lake, distance, water, think, large, healthy

Correct Object

airplane

Prediction by BERT-base

person

Prediction by BERT-base local + topic

airplane

(f)

Dialogue

A: Are there any people?
B: No.
A: Is there a table?
B: Notable.
A: What kinds of **foods**?
B: Fruits and vegetables.
A: What are **they** **sitting on**?
B: Some type of cooler.
A: What color is the cooler?
B: White with a air vent.
A: So they are inside?
B: Yes.
A: Are they fresh?
B: Yes they look very fresh.
A: Are they scattered around?

Predicted LDA Topic

32.2%: table, like, tell, bowl, home, sitting, **food**, restaurant, cup, utensil
21.0%: kitchen, **food**, good, cut, kind, inside, stove, like, vegetable, pot

Correct Object

food

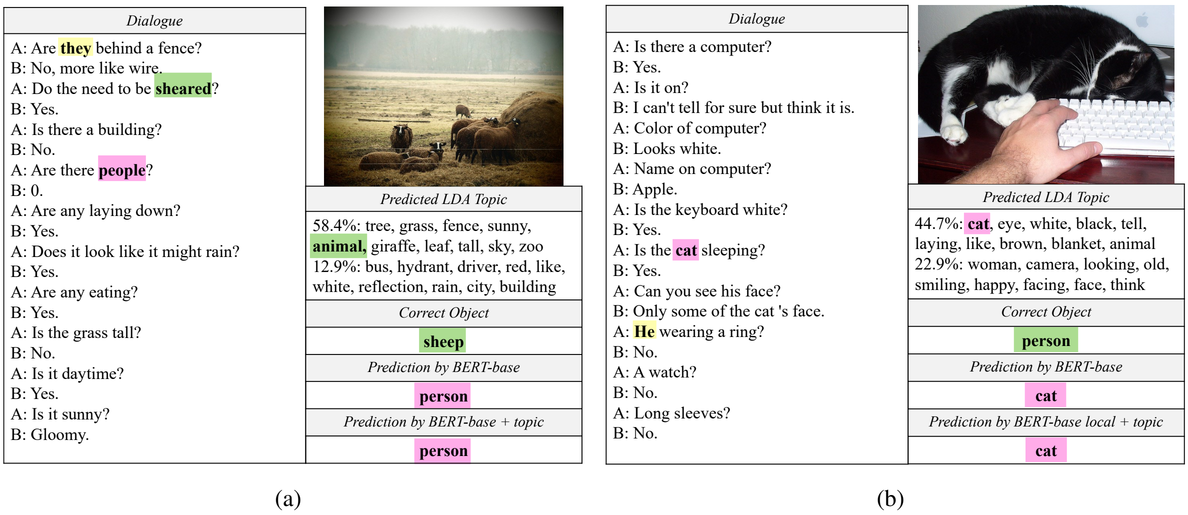
Prediction by BERT-base

person

Prediction by BERT-base + topic

food

Figure 9: Case study for (a)-(d) “Not Discussed” and (e)(f) “Discussed” out-of-text PCR. Target pronouns, correct out-of-text objects with their hints, and false prediction with distracting words are marked in different colors. Note that the images are not provided to the models.



Dialogue
A: Are **they** behind a fence? B: No, more like wire. A: Do the need to be **sheared**? B: Yes. A: Is there a building? B: No. A: Are there **people**? B: 0. A: Are any laying down? B: Yes. A: Does it look like it might rain? B: Yes. A: Are any eating? B: Yes. A: Is the grass tall? B: No. A: Is it daytime? B: Yes. A: Is it sunny? B: Gloomy.

Predicted LDA Topic
58.4%: tree, grass, fence, sunny, **animal,** giraffe, leaf, tall, sky, zoo 12.9%: bus, hydrant, driver, red, like, white, reflection, rain, city, building
Correct Object
sheep
Prediction by BERT-base
person
Prediction by BERT-base + topic
person

(a)

Dialogue
A: Is there a computer? B: Yes. A: Is it on? B: I can't tell for sure but think it is. A: Color of computer? B: Looks white. A: Name on computer? B: Apple. A: Is the keyboard white? B: Yes. A: Is the **cat** sleeping? B: Yes. A: Can you see his face? B: Only some of the cat 's face. A: **He** wearing a ring? B: No. A: A watch? B: No. A: Long sleeves? B: No.

Predicted LDA Topic
44.7%: **cat**, eye, white, black, tell, laying, like, brown, blanket, animal 22.9%: woman, camera, looking, old, smiling, happy, facing, face, think
Correct Object
person
Prediction by BERT-base
cat
Prediction by BERT-base local + topic
cat

(b)

Figure 10: Erroneous case study for "Not Discussed" out-of-text PCR. Target pronouns, correct out-of-text objects with their hints, and false prediction with distracting words are marked in different colors. Note that the images are not provided to the models.