Do Boat and Ocean Suggest Beach?
Dialogue Summarization with External Knowledge

Anonymous authors

Abstract

In human dialogues, utterances do not necessarily carry all the details. As pragmatics studies suggest [Grice, 1975], humans can infer meaning from the situational context even though the meaning is not literally expressed. It is crucial for natural language processing models to understand such inference process. In this paper, we address the problem of inferencing Concepts Out of the Dialogue Context (CODC) in the dialogue summarization task. We propose a novel framework comprised of a CODC inference module leveraging external knowledge from WordNet and a knowledge attention module aggregating the inferred knowledge into a neural summarization model. To evaluate the inference capability of different methods, we also propose a new evaluation metric based on CODC. Experiments suggest that current automatic evaluation metrics of natural language generation may not be enough to understand the quality of out-of-context inference in generation results, and our proposed summarization model can provide statistically significant improvements on both CODC inference and traditional automatic evaluation metrics, e.g., CIDEr. Human evaluation of the model's inference ability also demonstrates the superiority of the proposed model.

1. Introduction

Automatically summarizing conversations in our daily lives can benefit users for better organizing and retrieving their historical information. There have been several approaches to conversation summarization, including extractive approaches [Xie et al., 2008, Riedhammer et al., 2010] and abstractive approaches [Oya et al., 2014, Shang et al., 2018]. While extractive approaches focus on using the seen words in a conversation to summarize it, abstractive approaches usually use a generation based model to perform summarization.

Different from news articles, in daily dialogues, it is common that a speaker's utterance can suggest something that is not literally expressed but can be interpreted by a cooperative listener. For example, in Figure 1, if a dialogue mentions *boat* and *ocean*, the first impression of an ordinary person would be a boat sailing on the ocean. If, a modifier *abandoned* is added to *boat*, combined with *ocean*, the previous scene will be canceled and we will think
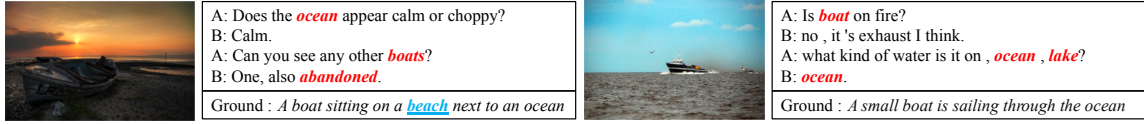
Figure 1: Two examples in DialSum. We highlight contextual concepts and out-of-context inference with **bold italics** and **underlined bold italics** respectively. "Ground" indicates ground truth summarizations.

of a beach or a shore because that is where the abandoned boat and ocean tend to cooccur. This kind of out-of-context inference is a language phenomenon in the field of pragmatics [Grice, 1975]. To comprehensively understand a dialogue where some context is omitted by the speakers, both thorough understanding of the context and necessary inference is required. However, to the best of our knowledge, neither evaluation metrics nor suitable methods are available for such kind of out-of-context inference currently, even if it is an easy task for human beings.

To study the phenomenon of out-of-context inference in dialogues, we narrow down the definition of out-of-context inference to a certain type so that we can automatically evaluate it. Here, we only focus on summarization's *new concepts* that are suggested by existing concepts in the dialogue context. This is also related to *lexical entrainment* [Brennan, 1996] which studies the lexical variability in language use and *lexical pragmatics* in relevance theory [Sperber and Wilson, 1986] which studies how to identify and infer concepts from words via broadening and narrowing contexts. To formally define the inference of new concepts in summarization, we distinguish a new concept from existing concepts following three rules according to WordNet [Miller, 1995]: (1) It should not be a general concept. (2) It should not be a synonym of an existing concept. (3) It should not be a hypernym (or super-concepts) of an existing concept. In this way, we can distinguish out-of-context inference from logical entailment or synonym as much as we can. Specifically, we name them as the Concept Out of the Dialogue Context (CODC.) Current evaluation metrics are usually based on lexical similarities or overlaps, e.g., BLEU [Papineni et al., 2002], ROUGE-L [Lin, 2004], METEOR [Lavie and Agarwal, 2007], CIDEr [Vedantam et al., 2015], etc. Such lexical metrics usually do not care about the recall of the novel words, and are not precise at evaluating the precision of the out-of-context inference. To tackle these limitations, we propose and study a new evaluation metric that incorporates CODC for the text generation.

Moreover, to improve the summarization results with the help of CODC inference, we propose an abstractive summarization framework that includes an out-of-context inference module to enhance the model's inference ability. The two-step framework we proposed basically contains an inference module and an extandable knowledge attention model. In the inference part, we use word relatedness features such as co-occurance, embedding similarities, and WordNet relatedness features to distinguish whether a candidate word is a plausible out-of-context inference or not. While in the knowledge attention module, we aggregate the retrieved knowledge from the inference module and apply attention mechanism to extract useful information in the decoding steps.

Our contributions are summarized as follows:

2

1. We address the problem of out-of-context inference in dialogue summarization for the first time, and provide an elaborative definition into the problem.

2. We design a related metric based on Concepts Out-of Dialogue Context (CODC) to evaluate neural models' ability of inferencing plausible novel concepts.

3. We proposed Trans-KnowAttn, an abstractive dialogue summarization approach that incorporates an out-of-context inference (missing-link inference) module and a knowledge attention module to improve the inference capability in dialogue summarization.

2. Dialogue Summarization Task

We develop our dialogue summarization (DialSum) task based on the VisDial [Das et al., 2017] dataset. To construct their dataset, VisDial asks two workers on Amazon Mechanical Turk to chat with each other in real-time to discuss the shared image in the MSCOCO dataset [Lin et al., 2014]. The MSCOCO dataset contains human-annotated captions of about 120K images. Each image has five captions from five different annotators. Annotators usually describe the most prominent concepts in an image. Thus, for each image, we have both a dialogue from the VisDial dataset and five captions from the MSCOCO dataset. We align a dialogue with five captions as five alternative summarizations of the dialogue. By nature, there are many out-of-context inference phenomena behind those utterances, as the speakers already have the image in mind as their context so that they don't need to bring it up again in the dialogue. Assuming a gold summarization y is provided by an annotator, we evaluate models' inference ability by how many novel concepts (i.e., noun phrases that *do not* appear in the dialogue) in y can be mentioned by $\hat{y}$, the generated summary, without introducing extra noisy concepts. The number of examples in the training, developing, and testing set are 98,256, 12,282, and 12,083, respectively.

3. CODC-based Evaluation Metric

3.1 Definition of CODC

In this section, we introduce the new metric based on CODC for evaluating models' conceptual inference ability. Here, by concept, we mean a noun word. As most of examples in our dataset are general objects rather than specific named entities, we use WordNet [Miller, 1995] to check the *relations* among all the concepts.

For each example in DialSum, which contains a dialog x and a description y, we denote the concept sets extracted from x and y as $\mathcal{C}^x$ and $\mathcal{C}^y$ respectively. Then we define the concept out of dialogue context set as

$$\mathcal{C}^{y-x} = \{c_y \in \mathcal{C}^y | \forall c_x \in \mathcal{C}^x, I_1(c_x, c_y) = 1)\}, \tag{1}$$

where $I_1(c_x, c_y)$ is the function implementing the following three rules:

• The least depth of all synsets containing c_y is no less than δ_d, where we set $\delta_d = 4$ empirically (1,790 synsets in total are filtered out.) This rule is to filter out very general concepts that may be inferred by anything.

• The shortest path between any pair of synsets containing c_x and c_y is greater than 1. This rule is to filter out concepts that are synonyms of existing concepts in the dialogue.
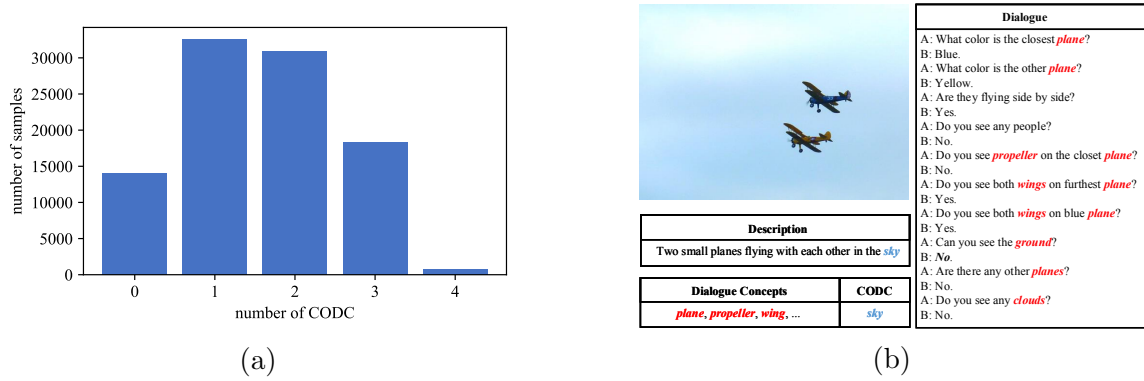
Figure 2: (a) Distribution of CODC. (b) An example of out-of-context inference in the dataset. We also show the corresponding figures from MSCOCO for better demonstration. The figures are not used in our experiments.

• Any synset of c_y is not a hypernym (super-concept) of any synset of c_x. This rule is to avoid entailment for the new concepts as essentially abstractive summarization is able to perform conceptual abstraction.

To better understand the dataset, we show the distribution of the counts of CODC per dialogue in Figure 2 (a), for one set of the descriptions in the training set. We can find that for more than 80% of the training dialogues, one or more CODC should be inferred. In Figure 2 (b), we show an real example from the dataset, where we can infer the word *sky* from the dialogue even if the word is not literally expressed. These observations show that the out-of-context concepts are common in this dialogue summarization task and further proves the importance of understanding how well models can generate summarization with out-of-context concepts.

3.2 CODC Precision, Recall, and F1

As a model with strong inference ability should be able to infer correct new concepts without introducing wrong ones, inspired by the F1 evaluation metric, we design F1 over CODC to evaluate models' inference ability. Let $\mathcal{X} = \{x^{(1)}, x^{(2)}, ..., x^{(N)}\}$ be all dialogues in the evaluation set, where N is the size of the dataset, and $\mathcal{Y} = \{\mathcal{Y}^{(1)}, \mathcal{Y}^{(2)}, ..., \mathcal{Y}^{(N)}\}$ be the set of ground-truth descriptions. Note that, for each image, we may have K ground-truth descriptions such that $\mathcal{Y}^{(i)} = \{y_1^{(i)}, ..., y_K^{(i)}\}$. We then define the precision, recall, and F1 over CODC as follows:

$$P_{CODC} = \frac{\sum_{i=1}^{N} \max_k |h(x^{(i)}, y_k^{(i)}, \hat{y}^{(i)})|}{\sum_{i=1}^{N} |\mathcal{C}^{\hat{y}^{(i)} - x^{(i)}}|},$$

$$R_{CODC} = \frac{\sum_{i=1}^{N} |h(x^{(i)}, y_{k_i^*}^{(i)}, \hat{y}^{(i)})|}{\sum_{i=1}^{N} |\mathcal{C}^{y_{k_i^*}^{(i)} - x^{(i)}}|}, \tag{2}$$

$$F1_{CODC} = \frac{2P_{CODC}R_{CODC}}{P_{CODC} + R_{CODC}},$$

where $h(x, y, \hat{y})$ and k_i^* are defined as follows:

$$h(x, y, \hat{y}) = \{c_{\hat{y}} \in \mathcal{C}^{\hat{y}-x} | \exists c_y \in \mathcal{C}^{y-x}, I_2(c_y, c_{\hat{y}}) = 1\}, \tag{3}$$
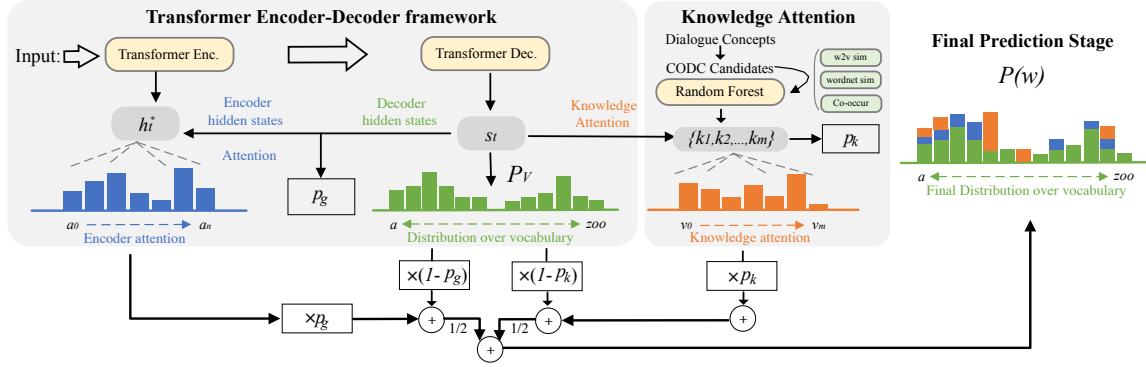
4

Figure 3: Overview of the Trans-KnowAttn framework.

$$k_i^* = \text{argmax}_{k \in \{1,\dots,K\}} \frac{|h(x^{(i)}, y_k^{(i)}, \hat{y}^{(i)})|}{|\mathcal{C}^{y_k^{(i)} - x^{(i)}}|}, \tag{4}$$

and $I_2(c_y, c_{\hat{y}})$ is the function to determine whether c_y and $c_{\hat{y}}$ are identical or entailed following either of two rules:

• There exists one pair of synsets of c_y and $c_{\hat{y}}$ being identical.

• There exists a synset of $c_{\hat{y}}$ being a hypernym of a synset of c_y.

Note that a summary can be considered as a plausible generation if it covers a reasonable amount of CODCs for **one of** the K descriptions. Based on this idea, for a certain dialogue, we only focus on the description where the generated sentence performs the best in precision or recall regarding the inference of CODC. In this case, *max* operation is selected in the calculation of precision and recall, which is different from the *average* operation that is typically used by previous automatic evaluation metrics in the setting of multi-reference. Also, by using the *max* operations, we argue that the theoretical upper bound for CODC precision, recall, and F1 are 1.0, as any ground-truth description will be scored 1.0 under the *max* operation.

4. Knowledge-aware Summarization

In this section, we present Trans-KnowAttn, a knowledge-aware summarization framework which consists of an missing-link inference module and a flexible knowledge attention module that can be applied to general encoder-decoder models. An overview of the model is shown in Figure 3.

4.1 Missing-link Inference Module

The missing-link inference module infers plausible concepts closely related to the concepts mentioned in the dialogue, while are out of the dialogue context. For example, considering the dialogue and ground-truth description in Figure 1, given *beach* and *boat* that are mentioned in the conversation, we find a list that contains concepts that potentially includes *beach*. This process is different from simple Knowledge Base Completion from that we want to infer the missing links between a concept with a set of dialogue concepts.

5

Following the definitions in the CODC metric, we first build a bipartite knowledge graph $G = (V, E), V = (D, C)$, that records the co-occurrence information of dialogue concepts and CODCs in the training set. Here D represents the set of concepts in all training dialogues, and C represents the set of concepts in CODCs. The weight of each vertices $u(c_x), c_x \in D$ or $u(c_y), c_y \in C$ is assigned by their total number of occurrence in the training set. An edge $(c_x, c_y), c_x \in D, c_y \in C$ exists if there are at least one dialogue-description pairs such that $I_1(c_x, c_y) = 1$, as has been defined in Equation (1). The weight of the corresponding edge $u(c_x, c_y)$ is the number of co-occurrence of the concept pair in the training set. Based on the formulation of the co-occurrence graph, we formalize the inference process as follows:

1. Extract all concepts using rules defined in Section 3 from a dialogue as $c_x^{(1)}, c_x^{(2)}, \ldots, c_x^{(k)}$, where k is the number of extracted concepts for dialogue x.

2. Get the set $\cup_{i=1,\ldots,k} N(c_x^{(i)})$ as primitive knowledge candidates, where $N(c_x^{(i)})$ is the set of neighbors of vertex $c_x^{(i)}$ in G.

3. Calculate features of all candidates and train a classifier to determine whether a candidate is a plausible CODC or not. The groud-truth CODCs are calculated based on the definition in Section 3.

4. In the inference process, use the classifier from the above step and retrieve top m results ranked by the classifier, among the set $\cup_{i=1,\ldots,k} N(c_x^{(i)})$.

The features we use are described as follow:

Co-occurance. A simulated co-occurance feature of a candidate $n \in \cup_{i=1,\ldots,k} N(c_x^{(i)})$ given a set of dialogue concepts $\{c_x^{(1)}, \ldots, c_x^{(k)}\}$ is defined as : $P_{co}(n|\{c_x^{(1)}, \ldots, c_x^{(k)}\}) \propto \sum_{i=1}^{k} \log[\frac{u(n, c_x^{(i)})}{u(c_x^{(i)})}]$, which is an ad-hoc attribute that depicts the log-probability that a candidate concept n will co-occur given all dialogue concepts.

Pre-trained Word Embedding Similarities. We compute the average word embedding similarities of candidate concept n with all the dialogue concepts $\{c_x^{(1)}, \ldots, c_x^{(k)}\}$ as : $\frac{1}{k} \sum_{i=1}^{k} \frac{\mathbf{e}_{c_x^{(i)}} \cdot \mathbf{e}_n}{||\mathbf{e}_{c_x^{(i)}}|| \cdot ||\mathbf{e}_n||}$, where $\mathbf{e}_{c_x^{(i)}}$ and $\mathbf{e}_n$ are embedding vectors of $c_x^{(i)}$ and n obtained from existing pretrained vectors such as Word2Vec [Mikolov et al., 2013] and Glove [Pennington et al., 2014].

WordNet Synset Relatedness. We choose some typical similarity measurement based on WordNet as indicator of relatedness, Path Similarity, i.e., the inverse of the number of nodes visited in the path from one word to another via hypernym hierarchy, LCH Similarity [Leacock et al., 2002], and WuP Similarity [Wu and Palmer, 1994].

4.2 Knowledge Attention Network

Treating the inferred candidate list as knowledge, we provide a knowledge attention mechanism that can be added to general encoder-decoder frameworks. We choose Transformer as the base architecture of the seq2seq model. To generate summaries, the decoder computes the hidden states and attend to knowledge embedding list step by step, adjusting the current decoder state with the help of attended knowledge vector. Also, copy mechanism is used for copying useful candidate words directly from the knowledge candidate list. The

Method	BLEU-4	METEOR	ROUGE-l	CIDEr	P_{CODC}	R_{CODC}	$F1_{CODC}$
BertSum	23.49	22.89	49.38	79.94	36.48	38.90	37.65
S2S-Attn	29.90	24.51	52.45	96.55	44.32	42.46	43.37
PGN	30.12	24.58	52.66	97.97	45.36	42.49	43.88
Pair-encoder	**31.26**	25.34	53.26	101.04	45.10	44.39	44.74
Trans-Copy	31.09	25.54	53.38	102.81	46.20	44.55	45.36
Trans-KnowAttn	31.22	**25.93**	**53.70**	**104.00***	**46.31**	**45.66***	**45.98***

Table 1: Evaluations on conventional metrics and CODC metrics are presented, where bold scores are the best among all models. * after bold figures indicates the improvements are statistically significant with $p < 0.05$.

model differs from the standard attention encoder-decoder framework from that an extra layer of knowledge attention is added to the decoding part. More detailed formulations of the summarization model is provided in Appendix A.

5. Experiments

5.1 Baselines

We select S2S-Attn [See et al., 2017], PGN [See et al., 2017], Trans-Copy [Vaswani et al., 2017], PairEncoder [Pan et al., 2018], and BertSum [Liu and Lapata, 2019]. More details of the baseline models are presented in the Appendix B.

5.2 Experimental Settings

Inference Module: In the inference module, a Random Forest classifier from sklearn [1] is used to distinguish whether a candidate word is a plausible CODC or not, and we select top $m = 13$ candidates ranked by the classifier as knowledge that are fed into the knowledge attention module. Details of the ablation study of this classification task are shown in Section 5.5.

Knowledge Attention Summarization Model: For all models except for PairEncoder, the input sequences are concatenated dialogue text with marks of `<q>` and `<a>` to iden-tity different turns with two speakers. For PairEncoder, the input is a list of utterance pairs. All five references are used in the training process. For models except for BertSum, we use a vocabulary of 20K words out of total 28K words. For all RNN-based models, 256-dimensional RNN hidden states and 256-dimensional word embeddings are applied for both knowledge word embeddings and encoder-decoder embeddings. For Trans-Copy, the dimension of word embeddings is set to be 256. For the BertSum model, we follow the same experimental setting as its original paper [Liu and Lapata, 2019].

5.3 Evaluation

Automatic Evaluation: Besides CODC metrics, conventional lexical metrics BLEU [Pap-ineni et al., 2002], ROUGE-L [Lin, 2004], METEOR [Lavie and Agarwal, 2007], and CIDEr [Vedantam et al., 2015] are used for evaluation.

---

1. https://scikit-learn.org

	General Quality	OOC-Inference
Trans-Copy	0.484	0.473
Trans-KnowAttn	**0.516**	**0.527**

	P_{CODC}	R_{CODC}	$F1_{CODC}$
all	8.83	46.71	14.85
-glove	9.21	48.48	15.48
-w2v	**9.30**	**48.91**	**15.63**
-cooccurance	5.84	30.13	9.78
-wordnet	8.62	45.66	14.51

Table 2: (a) Results of human evaluation. OOC-inference indicates out-of-context inference. (b) Effects of different features. - before the name of the features indicates removing this feature in the classification.

Human Evaluation: To better understand the effects of the knowledge attention module, we conducted human annotation on the overall quality and the out-of-context inference (missing-link inference) ability for 100 randomly selected summaries, generated by Trans-Copy and Trans-KnowAttn, respectively. Each pair of summaries were evaluated by 5 annotators. We run the evaluation using workers from Amazon Mechanical Turk (AMT.) Workers were given full illustration about the definition of out-of-context inference, and were asked to judge which summarization is better in terms of overall abstraction quality and out-of-context inference ability. The summaries were randomly shuffled to prevent unexpected bias. There was also a neutral option for the annotators if they find no difference between the two given summaries.

5.4 Results and Evaluation

Results of conventional metrics and CODC metrics are shown in Table 1. For the main metric, CIDEr, there's a 1.2 points improvement from Trans-Copy to Trans-KnowAttn. In addition, the improvements on the CODC Recall and F1 are also significant, indicating that the inferential ability of the model is improved by the inference module. We carry out statistical test using Randomization Test [Cohen, 1995]. All the improvements except for BLEU-4 and P_{CODC} are statistically significant. The human evaluation results are presented in Table 2 (a), where the scores for each model are proportional to the number of time a model is considered superior to the other. The scores in each columns are normalized so that the sum equals 1. We show that Trans-KnowAttn brings consistent improvements in terms of both inference ability and overall summarization quality.

5.5 Ablations

Inference Module: We check the importance of different features by removing each of them in the random forest classifier. As shown in Table 2 (b), we report the retrieval results on the validation set under top 10 retrieval, and the CODC scores are calculated on the retrieved knowledge. It shows that keeping one of the w2v or glove similarities would boost the retrieval performance. Also, co-occurance feature is the most prominent one among all. In the end, we select the -w2v setting for the summarization.

Summarization Results: We study the overall $F1_{CODC}$ and CIDEr scores of the generated summaries, given the number of retrieved knowledge words that are fed into the knowledge attention model, as shown in Figure 4 (a). When the number of retrieved words
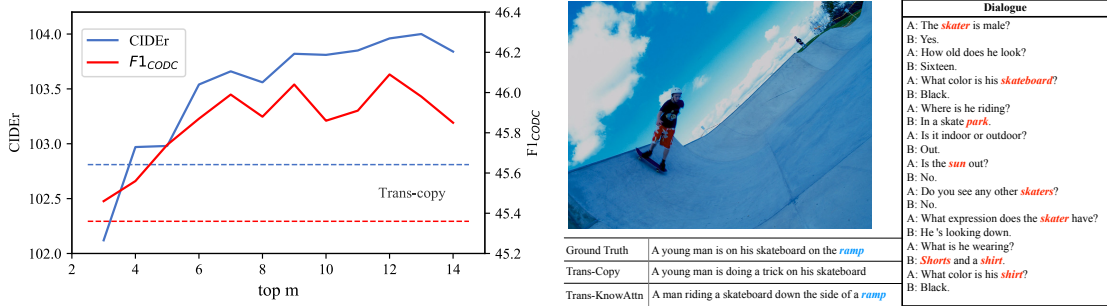
Figure 4: (a) Effects on the performance of the Trans-KnowAttn model given different number of retrieved knowledge words k. Dashed lines indicates the corresponding scores of Trans-copy baseline. (b) A case of CODC inference. We highlight contextual concepts and CODC inference with ***bold italics*** and ***bold italics*** respectively.

m exceeds 4, the overall CIDEr score can outperform the Trans-Copy baseline. For $F1_{CODC}$, it reaches the maximum when $m = 12$.

5.6 Case Study

We show an example of the generated results in Figure 4 (b). We observed concepts such as "*skater*," "*skateboard*," and "*skate park*" from the dialogue, from which human will easily inference a "*ramp*" from the context which is a common environment where people play skateboards. The Trans-Copy model fails to inference such new but highly relevant concept while Trans-KnowAttn will successfully yield such inference. More case studies will be presented in the appendix.

6. Related works

We introduce the related work in two-fold: neural text summarization and pragmatics tasks in NLP.

6.1 Neural Text Summarization

Conversation summarization has been studied extensively. It can be extractive approaches [Zechner, 2001, Nenkova and Bagga, 2003, Maskey and Hirschberg, 2005, Xie et al., 2008, Riedhammer et al., 2010] and abstractive approaches [Oya et al., 2014, Banerjee et al., 2015, Shang et al., 2018]. Nowadays, more attention has been paid to neural network based approaches due to the successful application of deep learning in natural language generation. Most of the neural network based approaches [Rush et al., 2015, Chopra et al., 2016, Nallapati et al., 2016, See et al., 2017, Paulus et al., 2018, Wang et al., 2019] follow the sequence-to-sequence framework [Sutskever et al., 2014]. Especially, attention mechanisms [Bahdanau et al., 2015] are widely used in the sequence-to-sequence framework to improve the generation quality [Luong et al., 2015, Rush et al., 2015, Xu et al., 2015]. Self-supervised pretrained language models, such as BERT [Devlin et al., 2019], are also applied to natural language generation tasks including summarization. Liu and Lapata [2019] intro-

duced a document-level encoder based on BERT which would better capture the semantics of a document. More recently, external knowledge in knowledge graphs and local sementic knowledge graph are also used to improve the correctness of factual statements and quality in abstractive summarization [Zhu et al., 2020, Huang et al., 2020]. Also, multi-view seq2seq models are designed [Chen and Yang, 2020] by combining conversational structures from different views for better representation of human conversations.

6.2 Pragmatics

Pragmatics [Grice, 1975] has been studied in both linguistics and natural language processing for a long time. It generally studies the ways in which the context contributes to the meaning. Early approaches only consider cases or rule based methods to evaluate pragmatics in language understanding and generation problems such as machine translation or dialogue systems [Rothkegel, 1986, Carberry, 1989, Iida et al., 1990]. Recent research focuses on using computational methods and automatic evaluation metrics in language games to evaluate the ability to inference through context [Frank and Goodman, 2012], which is usually called Rational Speech Acts (RSA) model. Wang et al. [2016] developed another language game and found that pragmatics may be hurt when people are using less precise and consistent languages. Fried et al. [2018] showed that explicit inference can also help learning based RSA models.

Besides typical RSA models that committed to Grice's original target, there are also several other interesting directions to explore in computational ways. Kazemzadeh et al. [2014] introduced a way to evaluate pragmatics, where an image is to used to identify an object inside. Then one player is asked to provide referring expression for the object while the other is asked to localize the object based on the expression. Many models have been developed for this task [Mao et al., 2016, Andreas and Klein, 2016, Monroe et al., 2017, Vedantam et al., 2017]. In addition, Lewis et al. [2017] introduced a negotiation generation task which implicitly needs pragmatics inference in the learning system. Another data (or task) is called CommitmentBank, which evaluates the inference of a speaker's commitment towards the content of the complement under different entailment-canceling environments [Jiang and de Marneffe, 2019a,b]. Recently, Shen et al. [2019] improves text generation with techniques of computational pragmatics, which are comprised of information preservation and explicit modeling of distractors.

7. Conclusion

In this paper, we address the problem of out-of-context inference in dialogue summarization for the first time, and proposed Trans-KnowAttn to improve the inference ability of dialogue summarization. Based on our observation, such kind of inference is required for most of the dialogues in the DialSum dataset. Experiments demonstrate that the proposed model can outperform previous state-of-the-art in terms of both traditional lexical metrics, and our newly proposed $F1_{CODC}$ significantly. We argue that with the help of out-of-context inference, neural models can better understand the pragmatics in human dialogues and thus improve the overall quality of summarization.

References

Jacob Andreas and Dan Klein. Reasoning about pragmatics with neural listeners and speakers. In *EMNLP*, pages 1173–1182, 2016.

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

Siddhartha Banerjee, Prasenjit Mitra, and Kazunari Sugiyama. Multi-document abstractive summarization using ilp based multi-sentence compression. In *IJCAI*, pages 1208–1214, 2015.

Susan E. Brennan. Lexical entrainment in spontaneous dialog. In *International Symposium on Spoken Dialogue, ISSD*, pages 41–44, 1996.

Sandra Carberry. A pragmatics-based approach to ellipsis resolution. *Computational Linguistics*, 15(2):75–96, 1989.

Jiaao Chen and Diyi Yang. Multi-view sequence-to-sequence models with conversational structure for abstractive dialogue summarization. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 4106–4118, Online, November 2020. Association for Computational Linguistics. doi: 10.18653/v1/2020.emnlp-main.336.

Sumit Chopra, Michael Auli, and Alexander M. Rush. Abstractive sentence summarization with attentive recurrent neural networks. In *NAACL HLT*, pages 93–98, 2016.

Paul R Cohen. *Empirical methods for artificial intelligence*, volume 139. MIT press Cambridge, MA, 1995.

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

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. BERT: pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*, pages 4171–4186, 2019.

Michael C. Frank and Noah D. Goodman. Predicting pragmatic reasoning in language games. *Science*, page 998, 2012.

Daniel Fried, Jacob Andreas, and Dan Klein. Unified pragmatic models for generating and following instructions. In *NAACL-HLT*, pages 1951–1963, 2018.

H. P. Grice. Logic and conversation. In Peter Cole and Jerry L. Morgan, editors, *Syntax and Semantics: Vol. 3: Speech Acts*, pages 41–58. Academic Press, 1975.

Luyang Huang, Lingfei Wu, and Lu Wang. Knowledge graph-augmented abstractive summarization with semantic-driven cloze reward. In *ACL*, pages 5094–5107, Online, July 2020. Association for Computational Linguistics.

Hitoshi Iida, Takayuki Yamaoka, and Hidekazu Arita. Three typed pragmatics for dialogue structure analysis. In *COLING*, pages 370–372, 1990.

Nanjiang Jiang and Marie-Catherine de Marneffe. Do you know that florence is packed with visitors? evaluating state-of-the-art models of speaker commitment. In *ACL (1)*, pages 4208–4213, 2019a.

Nanjiang Jiang and Marie-Catherine de Marneffe. Evaluating BERT for natural language inference: A case study on the commitmentbank. In *EMNLP/IJCNLP (1)*, pages 6085–6090, 2019b.

Sahar Kazemzadeh, Vicente Ordonez, Mark Matten, and Tamara L. Berg. Referitgame: Referring to objects in photographs of natural scenes. In *EMNLP*, pages 787–798. ACL, 2014.

Alon Lavie and Abhaya Agarwal. METEOR: an automatic metric for MT evaluation with high levels of correlation with human judgments. In *WMT@ACL*, pages 228–231, 2007.

Claudia Leacock, George A. Miller, and Martin Chodorow. Using corpus statistics and wordnet relations for sense identification. *Journal of Computational Linguistics*, 24(1): 147–165, 2002.

Mike Lewis, Denis Yarats, Yann Dauphin, Devi Parikh, and Dhruv Batra. Deal or no deal? end-to-end learning of negotiation dialogues. In *EMNLP*, pages 2443–2453, 2017.

Chin-Yew Lin. ROUGE: A package for automatic evaluation of summaries. In *Text Summarization Branches Out*, Barcelona, Spain, July 2004.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. Microsoft COCO: common objects in context. In *ECCV*, pages 740–755, 2014.

Yang Liu and Mirella Lapata. Text summarization with pretrained encoders. In *EMNLP-IJCNLP*, pages 3721–3731, 2019.

Minh-Thang Luong, Hieu Pham, and Christopher D Manning. Effective approaches to attention-based neural machine translation. In *EMNLP*, pages 1412–1421, 2015.

Junhua Mao, Jonathan Huang, Alexander Toshev, Oana Camburu, Alan L. Yuille, and Kevin Murphy. Generation and comprehension of unambiguous object descriptions. In *CVPR*, pages 11–20, 2016.

Sameer Maskey and Julia Hirschberg. Comparing lexical, acoustic/prosodic, structural and discourse features for speech summarization. In *INTERSPEECH*, pages 621–624, 2005.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. Distributed representations of words and phrases and their compositionality. In *NIPS*, pages 3111–3119, 2013.

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

Will Monroe, Robert X. D. Hawkins, Noah D. Goodman, and Christopher Potts. Colors in context: A pragmatic neural model for grounded language understanding. *TACL*, 5: 325–338, 2017.

Ramesh Nallapati, Bowen Zhou, Cícero Nogueira dos Santos, **c**Caglar Gül**c**cehre, and Bing Xiang. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *SIGNLL*, pages 280–290, 2016.

Ani Nenkova and Amit Bagga. Facilitating email thread access by extractive summary generation. In *RANLP*, pages 287–296, 2003.

Tatsuro Oya, Yashar Mehdad, Giuseppe Carenini, and Raymond Ng. A template-based abstractive meeting summarization: Leveraging summary and source text relationships. In *INLG*, pages 45–53, 2014.

Haojie Pan, Junpei Zhou, Zhou Zhao, Yan Liu, Deng Cai, and Min Yang. Dial2desc: End-to-end dialogue description generation. *CoRR*, abs/1811.00185, 2018.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. Bleu: a method for automatic evaluation of machine translation. In *ACL*, pages 311–318, 2002.

Romain Paulus, Caiming Xiong, and Richard Socher. A deep reinforced model for abstractive summarization. In *ICLR*, 2018.

Jeffrey Pennington, Richard Socher, and Christopher D Manning. Glove: Global vectors for word representation. In *EMNLP*, pages 1532–1543, 2014.

Korbinian Riedhammer, Benoit Favre, and Dilek Hakkani-Tür. Long story short–global unsupervised models for keyphrase based meeting summarization. *Speech Communication*, 52(10), 2010.

Annely Rothkegel. Pragmatics in machine translation. In *COLING*, 1986.

Alexander M. Rush, Sumit Chopra, and Jason Weston. A neural attention model for abstractive sentence summarization. In *EMNLP*, pages 379–389, 2015.

Abigail See, Peter J. Liu, and Christopher D. Manning. Get to the point: Summarization with pointer-generator networks. In *ACL*, pages 1073–1083, 2017.

Guokan Shang, Wensi Ding, Zekun Zhang, Antoine Tixier, Polykarpos Meladianos, Michalis Vazirgiannis, and Jean-Pierre Lorré. Unsupervised abstractive meeting summarization with multi-sentence compression and budgeted submodular maximization. In *ACL*, pages 664–674, 2018.

Sheng Shen, Daniel Fried, Jacob Andreas, and Dan Klein. Pragmatically informative text generation. In *NAACL-HLT*, pages 4060–4067, 2019.

Dan Sperber and Deirdre Wilson. *Relevance: Communication and Cognition.* Harvard University Press, Cambridge, MA, USA, 1986.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. Sequence to sequence learning with neural networks. In Z. Ghahramani, M. Welling, C. Cortes, N. D. Lawrence, and K. Q. Weinberger, editors, *NIPS*, pages 3104–3112. 2014.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. Attention is all you need. In *NIPS*, pages 6000–6010, 2017.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. Cider: Consensus-based image description evaluation. In *CVPR*, pages 4566–4575, 2015.

Ramakrishna Vedantam, Samy Bengio, Kevin Murphy, Devi Parikh, and Gal Chechik. Context-aware captions from context-agnostic supervision. In *CVPR*, pages 1070–1079, 2017.

Sida I. Wang, Percy Liang, and Christopher D. Manning. Learning language games through interaction. In *ACL*, 2016.

Wenbo Wang, Yang Gao, He-Yan Huang, and Yuxiang Zhou. Concept pointer network for abstractive summarization. In *EMNLP-IJCNLP*, pages 3067–3076, 2019.

Zhibiao Wu and Martha Palmer. Verbs semantics and lexical selection. In *ACL*, pages 133–138, 1994.

Shasha Xie, Yang Liu, and Hui Lin. Evaluating the effectiveness of features and sampling in extractive meeting summarization. In *SLT*, pages 157–160, 2008.

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron C. Courville, Ruslan Salakhutdinov, Richard S. Zemel, and Yoshua Bengio. Show, attend and tell: Neural image caption generation with visual attention. In *ICML*, pages 2048–2057, 2015.

Klaus Zechner. Automatic generation of concise summaries of spoken dialogues in unrestricted domains. In *SIGIR*, pages 199–207, 2001.

Chenguang Zhu, William Hinthorn, Ruochen Xu, Qingkai Zeng, Michael Zeng, Xuedong Huang, and Meng Jiang. Boosting factual correctness of abstractive summarization with knowledge graph. *CoRR*, abs/2003.08612, 2020.