Predicting the response specificity for answering questions in Dialog

Wei-Jen Ko
Department of Computer Science
University of Texas at Austin

Sanat Sharma
Department of Computer Science
University of Texas at Austin

Abstract

Visual Question Answering and Dialog generation are key problems to solve for Artificial agents. Human beings elicit a less specific or more specific response based on context, but dialog systems often prefer generic responses. In this work, we investigate if predicting the specificity of dialog responses helps in better answering questions in dialog, both in a pure text and a visual grounded setting. We propose a model that adds a specificity predictor network/module to a baseline Question Answering model in order to account for specificity of responses. The specificity network generates a score for each candidate response and we perform a weighted average with the scores of the Question Answering base model to predict the most suitable response to a question. We experimented on answering questions in the DailyDialogue dataset and the Visual Dialog dataset, and showed that using the predicted specificity provides a small but significant improvement in both pure text and visual grounded settings.

1 Introduction

Specificity has been a well-researched topic in the field of linguistics and natural language processing in recent years. One line of work is about controlling the specificity of the generated responses in dialog generation.Zhang et al. (2018) proposed to learn a specificity-based probability term for each vocabulary using a Gaussian mixture model, and added semantic-based probability. Ko et al. (2019b) and See et al. (2019) learned embeddings to represent different levels of specificity, and conditioned the decoder on those embeddings, thus allowing the decoder to generate taking into account a particular specificity. (See et al., 2019) also experimented on using weighted decoding(Ghazvininejad et al., 2017) to control specificity. At test time, all those approaches require an input value to indicate the specificity level of the response we wish to generate. These works used fixed specificity values in their experiments. (See et al., 2019) showed that fixing the specificity at a value too generic or too specific will result in worse human perceived engagement scores.

However, in real conversations, the response specificity is not a fixed value but varies across sentences. The response specificity for human beings is not only dependent on the current sentence but also past conversation history. For example, human beings might be less specific in their responses if similar information has been shared in the past. In this work, our goal is to predict the specificity value of responses from the input, and then generate responses using the predicted specificity value.

First, we train a predictor to predict the specificity of the response given a input sentence. This is a challenging task, since the response specificity is non-deterministic. For example, the responder could choose to give a detailed response or provide less information for the same question. Also the responder may first respond to the input and then bring up a new topic. In this case, the specificity of the second part of the response has little relation to the input sentence. Directly training a model to predict the specificity on a dialogue dataset performs poorly. In this work, we limit the input sentence to questions, because since the responder is forced to talk about the question, the scope of the answer is more limited and the specificity is more predictable.

Then we attempt to use the predicted specificity to improve the dialogue response. We combine the scores of candidate responses in retrieval models and the score related to specificity to choose the appropriate response. We experimented on a pure text setting and an image grounded dataset, and

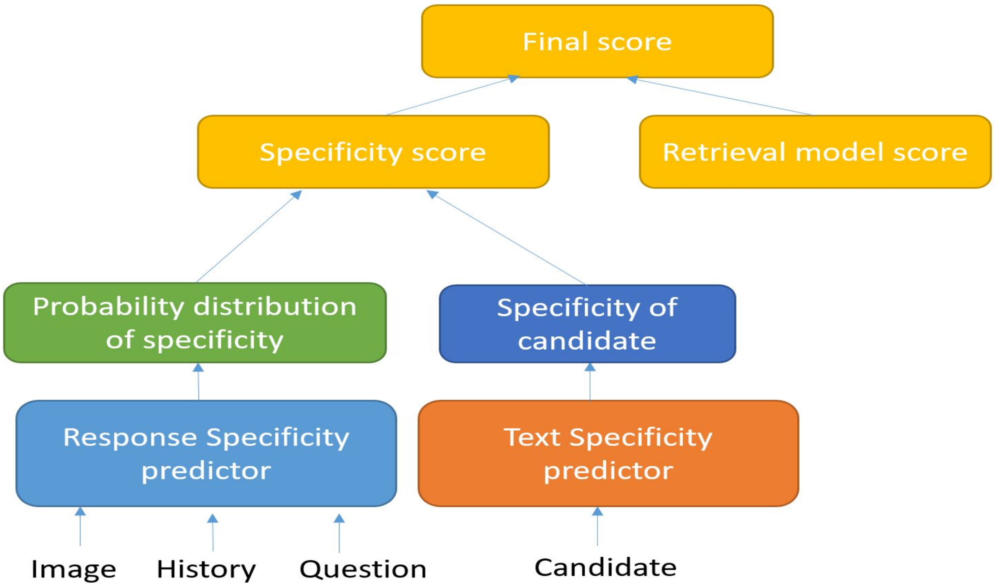


Figure 1: Structure of our method

showed that using the predicted specificity makes a small but significant improvement in both cases.

We experiment on retrieval models instead of generation models because of the difficulty of evaluating the generated sentences. Responses have to be plausible for human to compare the appropriateness of specificity between sentences, and current state-of-the-art models for multi-turn dialogue are not good enough to generate sufficiently coherent and plausible responses.

Furthermore, most Visual Question Answering/Dialogue provide a list of possible answer choices for a question, which makes a retrieval task the go-to method of choice for current models.

2 Baseline Model

For the pure text model, we use we use Key-value memory networks(Miller et al., 2016). This is an variant of the memory network by performing attention over keys. Dialogue history is used as keys and the next dialog utterances is used as values.

For Visual dialog, We utilize a Neural Visual Question Answering model for our baseline, based on the work by (Das et al., 2016). Now we introduce the details of this model.

2.1 Recurrent Encoder

The model utilizes an Long Short Term Memory (LSTM) layer to encode both dialog history as well as the current question. Dialog history is an important component of the model since it allows the model to understand the previous questions and base the current response based on this context. Similarly, all potential answer choices are also encoded using an LSTM layer. For each question in the base model, the dialog history, the image and 100 potential responses are provided as part of the dataset.

Image features are extracted from the image using a VGG model(Simonyan and Zisserman, 2015) and fused together with the textual embeddings (dialog history and question embeddings) to construct the input representation for the model. The respective answer embedding forms the Answer representation.

The base model also uses a ’attention-over-history’ approach to keep track of the part of the dialog history that most closely impacts the current question.

2.2 Discriminative Decoder

The Discriminative Decoder computes a dot product similarity over the answer representation and the input representation to calculate similarity for each answer choice and the result over to a softmax layer. The Softmax layer utilizes the similarity values to predict a posterior distribution over the answer options.

Fig 2 shows the different components of the baseline model.

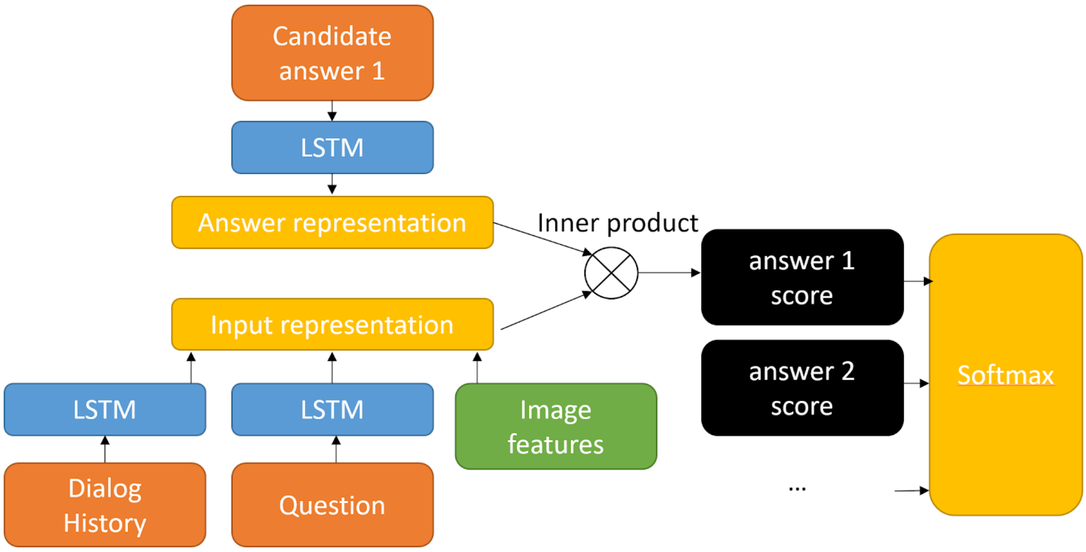


Figure 2: Neural Visual Question Answering Model

3 Method

3.1 Response specificity predictor

Our specificity predictor is a classification model based on a pretrained BERT(Devlin et al., 2019) model. BERT has shown to be extremely good at understanding semantic meaning of a sentence and construct context-aware embeddings. We utilize the BERT model and further fine-tune it on the response specificity prediction task. In the pure text case, the only new parameters we added during fine-tuning are a classification layer, which generates the classification probabilities from the final hidden vector corresponding to the first input token.

In the image grounded case, we use VGG(Simonyan and Zisserman, 2015) network to extract image features. We concatenate these image features with the textual features from BERT, and then feed the combined feature vector into a multi-layer perceptron layer, followed by a Softmax layer for classification.

We experimented on both 1-of-n coding with cross entropy loss, and ordinal classification(Frank and Hall, 2001). Ordinal classification trains multiple binary classifiers. Each classifier represents a specificity threshold, classifying if the specificity is larger than or smaller than that threshold. The performance difference between the two classification methods are not significant.

Even though specificity is a continuous value, we choose to train a classification model instead of a regression model. This is because the probability mass of the response is spread through all levels of specificity, making the mean of all questions very close to the middle specificity level, so when we train a regression model, all the specificity predictions are very condensed at the middle.

All the true specificity values in this work, including the labels for training the response specificity predictor, and the specificity value of candidate responses that will be described later, are obtained using a domain-agnostic text specificity predictor.(Ko et al., 2019a) This predictor uses INSTANTIATION discourse relation pairs labeled on Penn Discourse treebank(Marcus et al., 1994) as supervision signal, and performs domain adaptation and distribution regularization, using unlabeled sentences in a new domain. We use all gold responses in the training set as the unlabeled sentences to train this system.

Fig 3 shows the response specificity predictor model.

3.2 Combining with retrieval model

To combine the two models, we use a weighted sum of the two scores.

$$S = \alpha S_{ret} + (1 - \alpha) S_{spec} \qquad (1)$$

The scores of the retrieval model S_{ret} is the original score the baseline model used to rerank responses. For the scores related to specificity S_{spec}, we first feed the question, dialog history

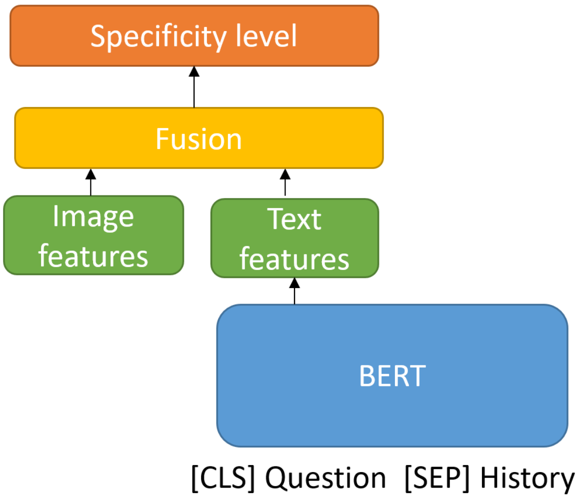


Figure 3: Response Specificity Predictor

and image into the response specificity predictor to get the probability of each specificity class. Then we calculate the specificity of each candidate response, and determine which specificity class the response is in. Then we use the probability of that specificity class predicted by the response specificity predictor as the score. To take into account the continuous nature of the specificity classes, we perform a convolution on the probabilities using a smoothing kernel. The formulation is

$$R_n = \frac{1}{1+2c}(P_n + c(P_{n+1} + P_{n-1})) \qquad (2)$$

where R_n indicates the final score of the n th specificity class, P_n is the probability of the n th specificity class by the predictor, and c is a tunable parameter.

$$S_{spec}[i] = R_{c[i]} \qquad (3)$$

$S_{spec}[i]$ is the score of the i th candidate, and $c[i]$ indicates the specificity level of the i th candidate.

4 Experiments

4.1 Dataset

For pure text, we use the DailyDialogue dataset(Li et al., 2017). We remove duplicated data in the dataset, and select only sentences including a question mark as input, and the following sentence as response.

Our training set has 27743 question-answer pairs, validation set has 2485 pairs, and test set has 2329 pairs.

For image grounded dialogue, we use the Visual Dialog dataset.(Das et al., 2016) The Visual Dialogue dataset is a subset of the MS COCO dataset (Lin et al., 2014). The dataset contains 123287 images in the training set and 2064 image in the validation set. Each image contains 10 rounds of dialogs related to the image. Because the test set is not released, we split the released validation set, using 1064 images for testing and 1000 images for validation. On average, each image in the dataset had 10 Question-Answer pairs associated with it, leading to roughly a million Question-Answer pairs that the model was trained on.

We decided to utilize Visual Dialog over more popular Visual Questioning Answering datasets such as VQA 2.0(Antol et al., 2015) due to a couple factors. Firstly, VQA has been widely studied and researched, whereas Visdial being a new dataset, allows us to investigate new data. Secondly, Visdial has a bigger variety and distribution of longer correct answers, which we aimed to exploit in our work.

As presented by Antol et al. (2015), fig 4 displays the differences in the distribution of questions and answers in Visdial versus VQA.

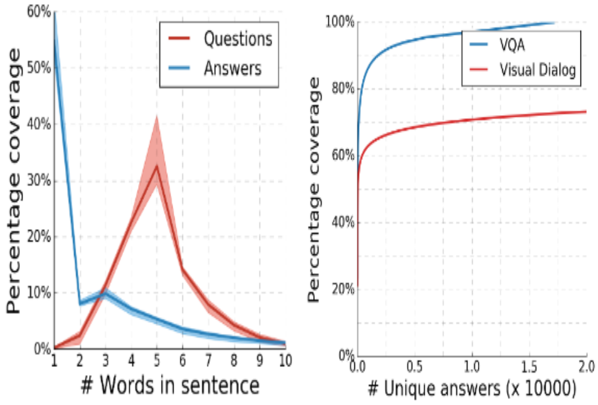


Figure 4: Visdial Dataset

4.2 Response specificity predictor

For specificity calculation, the specificity values generated by the network is a continuous value. For our classification task, we rank the specificity of the training set and split it into 5 classes, from least specific to most specific. We fine tune GoogleAI's pretrained BERT(Devlin et al., 2019) model with 12-layers, 768-hidden, 12-heads, and 110M parameters, using batch size 32 and Adam optimizer with learning rate 5e-5.

We determine the number of epochs using the validation set. In the image grounded setting, we add two fully connected layers after concatenating

Low specificity	Thank you . Are all these yours ? guess what ? I've got great news ! Could you sign each cheque here for me ? Did you hear about the robbery ? Is that everything that I have to do ?
High specificity	hello ! What are you reading about in the newspaper ? Ok . What do you think about our living room ? You are here on business , I think ? So , Paula, where are you from ? Did you get any honors or awards at your university ?

Table 1: Examples of questions predicted with the highest and lowest response specificity.

	accuracy
without images	70.66
with images	**70.69**

Table 2: Influence of images on specificity prediction

the image and text features. The hidden layer size for each of these fully connected layers is 128.

On the 5 way classification task, the accuracy for the daily dialog dataset we obtain is 37.7%, and for the visdial dataset, the accuracy is 70.6%. Table 1 shows some the questions that are predicted as most or least specific for subjective evaluation.

4.3 Influence of images on specificity prediction

We study the usefulness of images in predicting specificity, and compare the with and without images. Results are shown in table 2. Both models gets almost identical prediction accuracy, showing that image information does not really help the response specificity prediction.

This might be due to a few reasons. Firstly, we take a general image vector, containing the features of the entire image. On the other hand, while dealing with textual input, we utilize attention over the history to determine parts of relevant history. An attention based mechanism over the image might have yielded to better results. This might be accomplished using bounding boxes over various object regions on the image, and placing attention to the regions relevant for the current question. We initially tried this approach but soon found it to be unwieldy in terms of the training time required.

4.4 Retrieval results

Tables 3 and 4 show the retrieval results on the two datasets. We compare with an additional linear baseline, which means that we use a specificity score of value 1 to 5 proportional to the specificity level 1 to 5. This linear score is also combined with the score from the baseline model using a weighted sum.

For the baselines, we use Key-value memory networks(Miller et al., 2016) for Dailydialog. We use a Late Fusion Encoder and Discriminative decoder for Visdial. The α and c values are tuned on the validation set, and help to determine the weightage to give to the baseline model score versus our specificity score. We use $\alpha = 0.82$ and c=0.65 for Dailydialog and $\alpha = 0.75$ and c=0.6 for Visual dialog.

For the Daily dialog dataset our method outperforms the baselines on hits 1, 3, 5, and 10, indicating that our specificity component is useful. Since open domain dialog generation is known to prefer universal responses with low specificity, we compare with a linear baseline that tends to select more specific responses. Our method performs better than simply preferring more specific responses.

For the Visdial dataset, our method performs better on the hits@1, hits@5 and MRR metrics, showing that the top ranked answers are more correct. However, the Mean rank increased. This shows that ours method sometimes ranks the correct candidate to a much higher rank. During its evaluation phase, our evaluation technique uses Normalized discounted cumulative gain(NDCG) over top k responses to count similar responses (eg yeah or yes should both be given credit). The dataset provides this information by using human annotators to map similar option choices for each question.

	hits@1	3	5	10
no specificity	41.0	56.8	62.0	69.3
with specificity	**42.5**	**57.3**	**62.8**	**70.0**
linear baseline	41.1	57.0	62.2	69.4

Table 3: Daily Dialogue Results

	hits@1	hits@5	MRR	Mean
no specificity	44.6	76.2	0.590	**5.19**
with specificity	**44.9**	**76.3**	**0.592**	5.21
linear baseline	44.7	76.2	0.590	5.21

Table 4: Visual Dialogue Results

Qualitatively, we observe that when the two models, with and without specificity, select different answers, the model with specificity usually selects a less specific answer, such as yes/no. While not ideal, this is because a large proportion of answers in the Visdial dataset tend to be less specific or short. Given the skew of data towards less specific answers, the response specificity predictor often gives higher probability for low specificity levels. However, unlike our model, the linear model which always makes the model preferring more general responses for any input, does not improve over the original baseline. This shows that our response specificity predictor successfully learns when to generate a less specific response.

Even though the improvements of adding specificity are small, we performed significant tests on the results of both datasets, and the p value are both < 0.05, indicating the improvements are statistically significant.

4.5 Human Evaluation

Evaluation techniques such as MRR (Mean Reciprocal Rank) while useful do not fully correspond to how human beings rank/evaluate options, especially with artificial data. Hence, we utilize a small corpus taken from the validation set to perform human evaluation, in order to fully compare our model's results with respect to the base model.

We utilize 10 images chosen at random from the validation set, where the base model and our model differ in their selected answers. We gave all of our human evaluators a survey where they could either choose the first answer or the second answer. In case the evaluators felt that neither of the answers was correct or answered the question, they could click on a third option which stated ”Neither answer satisfactory”.

The evaluators were not told which answer choices corresponded to which model in order to remove bias from their judgment. We had 25 human evaluators take the survey and found that certain questions were answered correctly by one of the answer choices while some questions had a varied distribution in what answer choice the humans preferred.

Fig 5 shows some images where our model performs better than the base. However, in some cases, the base model predicts better options, as seen in Fig 6. Users preferred our model's response **66.6%** of the times compared to the non-specificity utilizing model.

Overall, we saw that both models struggle with counting tasks, as shown in Fig 7. This might be due to not efficiently using image features required for counting. Currently both models utilize VGG features for the entire image, which might lose semantic information such as the number of occurrences of an object. Constructing object boundaries and using attention to delve into features of specific regions of the image might lead to better performance for counting tasks.

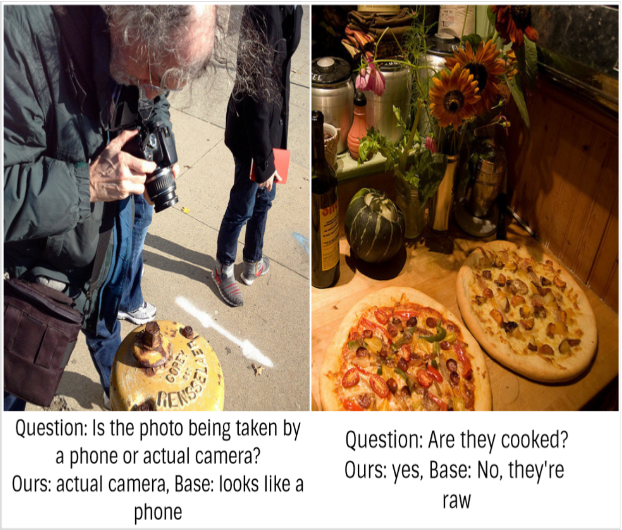
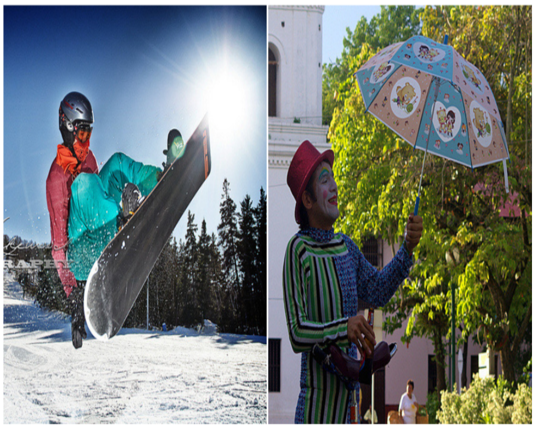

Question: Is the photo being taken by a phone or actual camera?
Ours: actual camera, Base: looks like a phone
Question: Are they cooked?
Ours: yes, Base: No, they're raw

Figure 5: Our Model performs favorably

In future iterations of our work, we aim to utilize a larger set of images for evaluation (in the magnitude of hundreds) in order to account for any human error or variance while choosing an option.

5 Conclusion

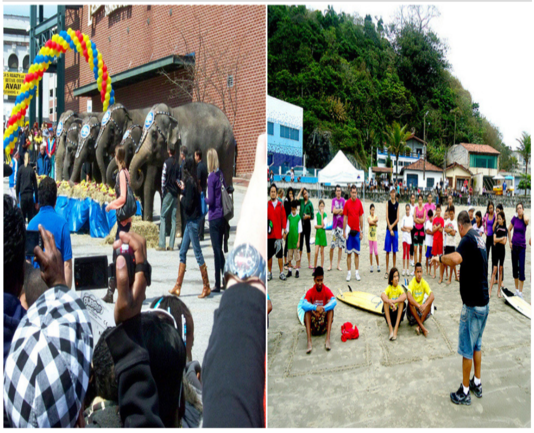
In this work we tied the concept of specificity with Visual Question Answering (and dialog generation), thus trying to mimic how humans respond



Question: Is it daytime or nighttime?
Ours: night time, Base: daytime

Question: Is he under the umrella?
Ours: yes, he is; Base: No, he's holding it up in front of him.

Figure 6: Failure cases



Question: How many elephants are there?
Ours: hundreds, Base: a lot, about 30

Question: How many students?
Ours: about 6, Base: 6 adult men

Figure 7: Failure on counting tasks

to questions. We created a specificity predictor model that helps predict an answer choice for a question based on the predicted specificity of the correct answer. We showed how the specificity model can be used as a module with any state of the art Visual Question Answering models and improve overall accuracy.

We evaluated and tested our model on the Dialy Dialog dataset as well as the Visual Dialog dataset, and showed statistically significant increases in accuracy in both, when using our model.

References

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *The IEEE International Conference on Computer Vision (ICCV)*.

Abhishek Das, Satwik Kottur, Khushi Gupta, Avi Singh, Deshraj Yadav, José M. F. Moura, Devi Parikh, and Dhruv Batra. 2016. Visual dialog. *CoRR*, abs/1611.08669.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL*.

Eibe Frank and Mark Hall. 2001. A simple approach to ordinal classification. In *ECML*.

Marjan Ghazvininejad, Xing Shi, Jay Priyadarshi, and Kevin Knight. 2017. Hafez: an interactive poetry generation system. In *ACL*.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019a. Domain agnostic real-valued specificity prediction. In *AAAI*.

Wei-Jen Ko, Greg Durrett, and Junyi Jessy Li. 2019b. Linguistically-informed specificity and semantic plausibility for dialog generation. In *NAACL*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, and Shuzi Niu Ziqiang Cao. 2017. Dailydialog: A manually labelled multi-turn dialogue dataset. In *IJCNLP*.

Tsung-Yi Lin, Michael Maire, Serge J. Belongie, Lubomir D. Bourdev, Ross B. Girshick, James Hays, Pietro Perona, Deva Ramanan, Piotr Dollár, and C. Lawrence Zitnick. 2014. Microsoft COCO: common objects in context. *CoRR*, abs/1405.0312.

Mitchell Marcus, Grace Kim, Mary Ann Marcinkiewicz, Robert MacIntyre, Ann Bies, Mark Ferguson, Karen Katz, and Britta Schasberger. 1994. The penn treebank: Annotating predicate argument structure. In *Proceedings of the Workshop on Human Language Technology*, HLT '94, pages 114–119, Stroudsburg, PA, USA. Association for Computational Linguistics.

Alexander Miller, Adam Fisch, Jesse Dodge, Amir-Hossein Karimi, Antoine Bordes, and Jason Weston. 2016. Key-value memory networks for directly reading documents. In *EMNLP*.

Abigail See, Stephen Roller, Douwe Kiela, and Jason Weston. 2019. What makes a good conversation? how controllable attributes affect human judgments. In *NAACL*.

K. Simonyan and A. Zisserman. 2015. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*.

Ruqing Zhang, Jiafeng Guo, Yixing Fan, YanyanLan, Jun Xu, and Xueqi Cheng. 2018. Learning to control the specificity in neural response generation. In *ACL*.