Controlling the Specificity of Clarification Question Generation

Yang Trista Cao
University of Maryland
ycao95@cs.umd.edu

Sudha Rao*
Microsoft Research
Sudha.Rao@microsoft.com

Hal Daumé III
University of Maryland
me@hal3.name

Abstract

Unlike comprehension-style questions, clarification questions look for some missing information in a given context. However, without guidance, neural models for question generation, similar to dialog generation models, lead to generic and bland questions that cannot elicit useful information. We argue that controlling the level of specificity of the generated questions can have useful applications and propose a neural clarification question generation model for the same. We first train a classifier that annotates a clarification question with its level of specificity (generic or specific) to the given context. Our results on the Amazon questions dataset demonstrate that training a clarification question generation model on specificity annotated data can generate questions with varied levels of specificity to the given context.

1 Introduction

In the field of natural language processing, the task of question generation has been predominantly defined as given a text, generate a question whose answer can be found in the given text (Heilman, 2011; Rus et al., 2010, 2011) to aid reading comprehension tasks. Recent advances in neural network modeling has triggered several sequence-to-sequence learning (Sutskever et al., 2014) based methods for question generation (Serban et al., 2016; Duan et al., 2017; Du et al., 2017).

In this work, however, we look at the task of clarification question generation i.e. generating questions that point at missing information in a given text. Recently, Rao and Daumé III (Rao and Daumé III, 2018) introduced a retrieval based model for this task, where given an unseen context, their model retrieves and ranks a set of candidate clarification questions from the training data by their relevance to the context. They followed this work by a generation model (Rao and Daumé III, 2019) which given a context, generates a useful clarification question from scratch. They find that training a vanilla sequence-to-sequence neural network model to generate a clarification question given a context results in over-generic questions, similar to recent findings in dialogue generation (Li et al., 2016). Therefore, they train their model to maximize over the usefulness of the generated question.

In this work, we hypothesize that if we label the clarification questions in the training data with their level of specificity to the context, then a vanilla sequence-to-sequence learning model can learn to control the level of specificity at test time. We define two levels of specificity: generic where the question is applicable to many contexts and specific where the question is applicable to relatively a few contexts. Figure 1 shows an example generic and specific question given a product description from Amazon.

The problem of measuring the level of specificity of text has received sparse attention. Louis and Nenkova (2011) first introduce a supervised binary classifier to identify whether the summary of a given text is specific or generic. Recently, Gao et al. (2019) propose a supervised regression model for identifying the specificity of sentences at a more finer grained level. While these works focus on identifying the specificity level of text, we go a step further and use the classifier as a guidance to control the level of specificity of the generated questions. To achieve this, we take a semi-supervised approach where we first train a model that automatically predicts a question’s specificity level (generic or specific) using a small amount of annotated data (Section 2). We use this classifier in turn to label all the questions

* This research was performed when the author was still at University of Maryland, College Park.
Figure 1: Product description from amazon.com paired with a generic and a specific clarification question.

in training data of our question generation model with its level of specificity to the context. Then motivated by Sennrich et al. (2016), we build a question generation model that incorporates the level of specificity as an additional input signal during training (Section 3). During test time, given a new context and a level of specificity (which is either generic or specific), our model generates a question at that level of specificity.

2 Model for Automatically Predicting Specificity Level

We annotate a set of 3000 questions from the Amazon dataset (Rao and Daumé III, 2019) with generic/specific labels using Amazon Mechanical Turk workers. Each question was annotated by three annotators and we take the majority as the label for that question.\footnote{In x\% of cases when there was no majority, we pick a label at random.} Given this annotated data, we want to train a machine learning model that can learn to predict the specificity level given a context and a question. We use some of the features described in Louis and Nenkova’s work (Louis and Nenkova, 2011) and introduce some new context-based features relevant to our setting. Based on these features, we train a logistic regression model to make a binary prediction (-1: generic, 1: specific) given a context and a question. We use the Support Vector Regression (SVR) model with Radial Basis Function (RBF) kernel. Gao et al. (2019), in their work of analyzing language in social media post, claim SVR with RBF has the best performance in predicting text specificity.

3 Specificity-Controlled Question Generation Model

The key idea behind sequence-to-sequence approaches is that given large amounts of input, output sequence pairs, the model learns internal representations such that at test time, given an input sequence, it generates the appropriate output sequence. We use the specificity classifier described in the previous section to label all the questions in the training (and tune) data with generic/specific labels. We use these labels to append each context with the <specific> tag when the question paired with the context is labeled as specific and with the <generic> tag when the question paired with the context is labeled as generic. We train an attention-based sequence-to-sequence learning model (Luong et al., 2015) on (context+specificity, question) pairs using maximum likelihood objective. At test time, given a new context appended with the desired level of specificity, we generate a question at that level of specificity.

4 Results and Conclusion

We evaluate our specificity classifier using 10-fold cross-validation on our labeled set of 3000 questions. We find that our specificity classifier is able to predict the level of specificity of the question to the context with 76% accuracy. In comparison, a majority baseline achieves an accuracy of 65%. We also find question length and question word embeddings to be strong indicators of specificity.

For our specificity-controlled question generation model, we answer the following research questions using our experimentation:

1. Does our proposed model generate specific and generic clarification questions when we append the context with <specific> and <generic> tag respectively?
2. What is the effect of applying our idea to a vanilla MLE trained model versus applying it to the state-of-the-art GAN-based model (Rao and Daumé III, 2019)?
3. Does generating more specific questions adversely affect grammaticality or relevancy?
Table 1: Human judgments are obtained on 300 questions from the Home & Kitchen category of Amazon. (s) denotes specific reference/output and (g) denotes generic reference/output. Relevancy scores are in the range [0-1], grammaticality in [0-1] and specificity in [1-4]. The difference between the bold and the non-bold numbers is statistically significant with p < 0.05 (reference excluded). BLEU and METEOR scores are calculated on the entire test set by comparing output with an average of 3 references under (s) setting and 6 references under (g) setting.

**Dataset:** We evaluate our proposed model on the Home & Kitchen category of the Amazon dataset (McAuley and Yang, 2016) consisting of 91,874 training, 11,646 tune and 11,264 test questions.

**Baselines:** We compare our model to two baselines: MLE, a sequence-to-sequence model trained using maximum-likelihood estimation and GAN-UTILITY (Rao and Daumé III, 2019), the previous state-of-the-art model on Amazon dataset. SPEC-MLE is our model applied to the MLE-trained model and SPEC-GAN-UTILITY is our model applied to the GAN-UTILITY-trained model.

**Metrics:** Inspired by Rao and Daumé III’s (2019) human-based evaluation methodology, we ask humans to judge outputs for relevancy, fluency, specificity and seeking new information. We use BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005) for our automatic metric-based evaluation. When we append the context with the ⟨specific⟩ tag (setting (s)), we compare outputs to specific references and when we append context with ⟨generic⟩ tag (setting (g)), we compare outputs to generic references (using our classifier to identify specific vs generic reference questions).

**Analysis:** According to human judgments (left half of Table 1), SPEC-MLE and SPEC-GAN-UTILITY generate questions that are significantly more specific under setting (s) and significantly more generic under setting (g) compared to other models. All models are statistically indistinguishable under seeking new information criteria and get a score of around 0.80 (range [0-1]). However, SPEC-MLE and SPEC-GAN-UTILITY get reasonable but statistically significantly lower relevance and grammatical scores under setting (s) suggesting that increased specificity comes at a cost of slightly lower relevancy and fluency. Sample model outputs are included in the supplementary material.

Under automatic metrics (right half of Table 1), SPEC-MLE gets significantly higher BLEU and METEOR scores compared to MLE and GAN-UTILITY suggesting that it generates generic and specific questions that are more similar to the references. Interestingly, SPEC-MLE beats SPEC-GAN-UTILITY suggesting that our approach is more effective when applied on the simpler MLE trained model.

In this work, we thus introduce a semi-supervised approach to controlling the level of specificity of clarification questions to a given context.

**References**


