Answer-based Adversarial Training for Generating Clarification Questions (Supplementary Material)

Anonymous NAACL submission

1 Details of Sequence-to-sequence Model

In this section, we describe some of the details of the attention based sequence-to-sequence model introduced in Section 2.1 of the main paper. In equation 1, $\tilde{h}_t$ is the attentional hidden state of the RNN at time t obtained by concatenating the target hidden state h_t and the source-side context vector $\tilde{c}_t$, and W_s is a linear transformation that maps h_t to an output vocabulary-sized vector. Each attentional hidden state $\tilde{h}_t$ depends on a distinct input context vector $\tilde{c}_t$ computed using a global attention mechanism over the input hidden states as:

$$\tilde{c}_t = \sum_{n=1}^{N} a_{nt} h_n \tag{1}$$

$$a_{nt} = \text{align}(h_n, h_t) \tag{2}$$

$$= \exp\left[h_t^T W_a h_n\right] \Big/ \sum_{n'} \exp\left[h_t^T W_a h_{n'}\right] \tag{3}$$

The attention weights a_{nt} is calculated based on the alignment score between the source hidden state h_n and the current target hidden state h_t.

2 Experimental Details

In this section, we describe the details of our experimental setup.

We preprocess all inputs (context, question and answers) using tokenization and lowercasing. We set the max length of context to be 100, question to be 20 and answer to be 20. We test with context length 150 and 200 and find that the automatic metric results are similar as that of context length 100 but the experiments take much longer. Hence, we set the max context length to be 100 for all our experiments. Similarity, we find that an increased length of question and answer yields similar results with increased experimentation time.

Our sequence-to-sequence model (Section 2.1) operates on word embeddings which are pre-trained on in domain data using Glove (Pennington et al., 2014). As frequently used in previous work on neural network modeling, we use an embeddings of size 200 and a vocabulary with cut off frequency set to 10. During train time, we use teacher forcing (Williams and Zipser, 1989). During test time, we use beam search decoding with beam size 5.

We use a hidden layer of size two for both the encoder and decoder recurrent neural network models with size of hidden unit set to 100. We use a dropout of 0.5 and learning ratio of 0.0001. In the MIXER model, we start with $\Delta = T$ and decrease it by 2 for every epoch (we found decreasing Δ to 0 is ineffective for our task, hence we stop at 2).

3 Analysis of System Outputs on Amazon Dataset

Table 1 shows the system generated questions for three product descriptions in the Amazon dataset.

In the first example, the product is a shower curtain. The Reference question is specific and highly useful. Lucene, on the other hand, picks a moderately specific ("how to clean it?") but useful question. MLE model generates a generic but useful "is it waterproof?". Max-Utility generates comparatively a much longer question but in doing so loses out on relevance. This behavior of generating two unrelated sentences is observed quite a few times in both Max-Utility and GAN-Utility models. This suggests that these models, in trying to be very specific, end up losing out on relevance. In the same example, GAN-Utility also generates a fairly long question which, although awkwardly phrase, is quite specific and useful.

In the second example, the product is a Du-

1

vet Cover Set. Both Reference and Lucene questions here are examples of questions that are pretty much useful only to the person asking the question. We find many such questions in both Reference and Lucene outputs which is the main reason for the comparatively lower usefulness scores for their outputs. All three of our models generate irrelevant questions since the product description explicitly says that the set is full size.

In the last example, the product is a set of mopping clothes. Reference question is quite specific but has low usefulness. Lucene picks an irrelevant question. MLE and Max-Utility generate highly specific and useful questions. GAN-Utility generates an ungrammatical question by repeating the last word many times. We observe this behavior quite a few times in the outputs of both Max-Utility and GAN-Utility models suggesting that our sequence-to-sequence models are not very good at maintaining long range dependencies.

4 Analysis of System Outputs on Stack Exchange Dataset

Table 2 includes system outputs for three posts from the Stack Exchange dataset.

The first example is of a post where someone describes their issue of not being able to recover from their boot. Reference and Lucene questions are useful. MLE generates a generic question that is not very useful. Max-Utility generates a useful question but has slight ungrammaticality in it. GAN-Utility, on the other hand, generates a specific and an useful question.

In the second example, again Reference and Lucene questions are useful. MLE generates a generic question. Max-Utility and GAN-Utility both generate fairly specific question but contain unknown tokens. The Stack Exchange dataset contains several technical terms leading to a long tail in the vocabulary. Owing to this, we find that both Max-Utility and GAN-Utility models generate many instances of questions with unknown tokens.

In the third example, the Reference question is very generic. Lucene asks a relevant question. MLE again generates a generic question. Both Max-Utility and GAN-Utility generate specific and relevant questions.

References

Jeffrey Pennington, Richard Socher, and Christopher Manning. 2014. Glove: Global vectors for word representation. In *Proceedings of the 2014 conference on empirical methods in natural language processing (EMNLP)*, pages 1532–1543.

Ronald J Williams and David Zipser. 1989. A learning algorithm for continually running fully recurrent neural networks. *Neural computation*, 1(2):270–280.

Title	Extra Wide **Shower Curtain** for a Clawfoot Tub/opaque with Magnets		
Product Description	Approximately 180" Wide 70" Long. Extra Wide to keep the water and heat in. Opaque in color to let the light in. Can be used as a liner or the curtain itself.		
		Specificity	Usefulness
Reference	does this curtain come with hooks?	4	4
Lucene	how to clean it ? i had it over 10 months now . and some mildew grow on it	3	3
MLE	is it waterproof ?	2	1
Max-Utility	what is the diameter of the bottom of the shower ? i have a kohler shower curtain .	N/A	N/A
GAN-Utility	does this curtain have suction feet on the bottom of the shower curtain ?	4	3

Title	Chezmoi Collection 7-piece Chic Ruched White **Duvet Cover Set**, Full Size (with Throw Pillows)		
Product Description	Create a world of romance with the elegantly-styled, ruching,and luxurious all white duvet cover set.		
		Usefulness [1-4]	Specificity [1-4]
Reference	How long will it take to ship this to my pin code?	1	1
Lucene	and can you use the duvet as is ? if not what shall i purchase to put in it for winter or fall ?	1	4
MLE	what are the dimensions of the king size ?	N/A	N/A
Max-Utility	what are the dimensions of the king size ?	N/A	N/A
GAN-Utility	does the king size come with a duvet cover or do you have to buy a king duvet ?	N/A	N/A

Title	Microfiber 3-Pack, **Pro-Clean Mopping Cloths** for Braava Floor Mopping Robot		
Product Description	Braavas textured Pro-Clean microfiber mopping cloths remove dirt and hair from your floors. The cloths can be washed and used hundreds of times. They are compatible with all Braava models, including the Pro-Clean Reservoir Pad. Each cloth is easy to attach and remove from the magnetic cleaning pad.		
		Usefulness [1-4]	Specificity [1-4]
Reference	do i have to use a new cloth every time i want to clean my floor? $5/$6 seems expensive per clean	2	4
Lucene	do they remove pet odor ?	N/A	N/A
MLE	will these work with the scooba ?	3	3
Max-Utility	do these cloths work on hardwood floors ?	3	4
GAN-Utility	will this work with the scooba mop mop mop mop mop mop mop	N/A	N/A

Table 1: Example outputs from each of the systems for three product descriptions from the Home & Kitchen category of the Amazon dataset.

3

Title	how can i recover my boot ?
Post	since last week i am trying to recover my boot . after the last update for ubuntu 12.04 i lost it and nobody could help me , i used boot-repair but there are problems with dependences , which ca n't be fix . i will be very grateful if somebody could help me .
Reference	what happens when you try to boot ?
Lucene	can you get into bios ?
MLE	how much ram do you have ?
Max-Utility	do you have a swap partition partition ?
GAN-Utility	what happens when you try to boot into safe mode ?

Title	packages have unmet dependencies when trying to install anything
Post	i 'm running ubuntu 14.04 lts . ive recently run into this problem with several applications , it seems to happen to anything i need to update or install . i just gave up the first 2 or 3 times this happened after trying a few solutions to no avail , but now i 'm having the same issue with steam trying to update , which i use quite a lot . ive looked through dozens of posts about similar issues and tried a lot of solutions and nothing seems to work.
Reference	sudo dpkg -reconfigure all ? ?
Lucene	if you use the graphical package manager , does n't add the required packages automatically ?
MLE	how long did you wait ?
Max-Utility	can you post the output of 'apt-cache policy UNK ?
GAN-Utility	can you post a screenshot of the output of 'sudo apt-get install UNK

Title	full lubuntu installation on usb (uefi capable)
Post	i want to do a full lubuntu installation on a usb stick that can be booted in uefi mode. i do not want persistent live usb but a full lubuntu installation (which happens to live on a usb stick) and that can boot fromanyuefi-capable computer ...
Reference	hello and welcome on askubuntu . could you please clarify what you want ?
Lucene	so , ubuntu was installed to the pen drive ?
MLE	which version of ubuntu ?
Max-Utility	do you have a live cd or usb stick ?
GAN-Utility	what is the model of the usb stick ?

Table 2: Example outputs from each of the systems for three posts of the Stack Exchange dataset.