Project Laboratory

Csáky Richárd Krisztián

1 Introduction

Current conversational models lack diversity and generate boring responses to open-ended utterances (Li et al., 2015; Wei et al., 2017; Shao et al., 2017). *Priors* provide additional information to dialog models to aid response generation, but annotating a dataset with priors such as persona (Li et al., 2016a), emotion (Zhou et al., 2017), or topic (Xing et al., 2017), is expensive and such annotations are rarely available. In this work a method is presented for improving chatbots' responses to open-ended utterances by removing those utterances from training data using a simple entropy-based approach that does not require human supervision. It is shown that training on this filtered dataset results in better conversational quality as chatbots learn to output better and more diverse responses to these utterances.

A brief background and previous approaches to the issues mentioned are given in Section 2, and Section 3 describes the method in detail. Section 4 presents an analysis of the filtered dataset, dialog systems trained on the new datasets are evaluated in Section 5. Section 6 concludes and presents future work. All code for the experiments in this paper can be found at `https://github.com/ricsinaruto/Seq2seqChatbots`.

2 Background

2.1 Chatbots

A conversational agent (chatbot) is a piece of software that is able to communicate with humans using natural language. Many types of chatbots exist, but in this work the focus is on single-turn neural network based generative agents. Single-turn means that the only information the chatbot has is the previous utterance emitted by the user, and it has to form a reply based on this. Neural networks are widely used for modeling language (Mikolov, 2010), and they have been shown to be capable of modeling dialog (Vinyals and Le, 2015a). Finally, generative means that the chatbot is trying to emit replies that are not retrieved from some given dataset, but rather generated by the neural network model. Refer to (Csáky, 2017) for a more in-depth background on conversational agents.

2.2 Neural Networks

Neural networks are high-dimensional non-linear functions that can be used to model a plethora of tasks. Besides natural language they have been applied to image- and audio-based tasks as well (Krizhevsky and Sutskever, 2012; Van Den Oord et al., 2016). In the chatbot case the neural network takes as input an utterance from a dialog dataset. The string utterance is transformed to a numerical representation using word vectors (Mikolov et al., 2013). The neural network takes these vectors as input and applies some mathematical transformations to produce an output. In the chatbot case the output is the response utterance to the given input. The exact type of mathematical transformations used is given by the architecture of the neural network. For conversational modeling some type of encoder-decoder model is used (Sutskever et al., 2014). Neural network models have a plethora of parameters that can be changed inside the mathematical transformations. Through changing these parameters the right way a neural network can learn to produce better and better outputs, this being called learning or training. Essentially the output of the network is compared to the target output and based on the error, gradient descent is used to find the parameters inside the network that can best approximate the target output through an iterative process. Refer to (Csáky, 2017) for a more in-depth description of neural networks and their application to conversational modeling.

2.3 Issues

Current open-domain NCMs are based on neural architectures developed for machine translation (MT). Conversational data differs greatly from MT data in that targets to the same source sentence may vary not only grammatically but also semantically (Wei et al., 2017; Tandon et al., 2017); consider plausible replies to the question: *What did you do today?*. Dialogue datasets also contain responses that appear after many different inputs, e.g. answers such as *yes*, *no* and *i don't know* appear after a large and diverse set of inputs. Following the approach of modeling conversation as a sequence to sequence (`seq2seq`) (Sutskever et al., 2014) transduction of single dialog turns, these issues can be referred to as the *one-to-many*, and *many-to-one* problem, respectively. Since `seq2seq` architectures are inherently deterministic, meaning that once trained they can't output different sequences to the same input sequence, they are not suited to deal with the ambiguous nature of dialogs.

The focus of this work is the *one-to-many*, and *many-to-one* problem, previous approaches to which can be grouped into three categories. First, the encoding procedure can be modified by feeding more information into the model, like dialog history (Serban et al., 2016), persona information (Li et al., 2016a; Joshi et al., 2017; Zhang et al., 2018), mood/emotion category (Zhou et al., 2017; Li et al., 2017b), topic category (Xing et al., 2017; Liu et al., 2017), etc. Second, some approaches augment the decoding process, with e.g. latent variable sampling (Serban et al., 2017b; Zhao et al., 2017) or beam search (Goyal et al., 2017; Wiseman and Rush, 2016; Shao et al., 2017). Finally, directly modifying the loss function (Wiseman and Rush, 2016) or training procedure of the model, by using reinforcement (Li et al., 2016c; Serban et al., 2017a; Li et al., 2016b; Lipton et al., 2017) or adversarial learning (Li et al., 2017a) are also among the solutions proposed.

3 Methods

In this work the *one-to-many*, *many-to-one* issue is approached from a different perspective: instead of adding more complexity, we try simple data filtering methods to exclude source-target utterance pairs that have high entropy, since we believe that these cause dialog models to output safe but boring responses. Entropy of utterances has also been used before for evaluation purposes (Serban et al., 2017b). The entropy of a source/target utterance is calculated based on the distribution of the target/source utterances that it is paired with in the dataset. In essence, the learning task is formulated in a way for which the maximizing likelihood approach is more suitable. NCMs have been shown to produce better qualitative results after they overfit the training data (Csáky, 2017; Tandon et al., 2017). This also supports the claim that the loss function is not capturing conversational goals, since a neural network model should perform best when the validation loss is minimal. Our experiments suggest that when training NCMs on our filtered datasets, validation loss becomes a better indicator of the model's performance.

Of the 72 000 unique source utterances in the DailyDialog dataset (see Section 4 for details), 60 000 occur with only a single target. For these it seems straightforward to maximize the conditional probability $P(T|S)$, S and T denoting a specific source and target utterance. However, in the case of sources that appear with multiple targets in the dataset, models are forced to learn some "average" of observed responses. This is the *one-to-many* problem. We can similarly formulate the *many-to-one* problem, where a diverse set of source utterances are observed with the same target. This may be a less prominent issue in training NCMs, since the probability of source utterances given some target doesn't appear in standard loss functions (although it is used in some special objective functions (Li et al., 2015)). Still, we shall experiment with excluding such targets (e.g. *I don't know*), since conversational models generate these quite frequently and they are typically uninformative and unengaging (see Section 5.1 on evaluation principles).

For each source utterance s in the dataset we calculate the entropy of the distribution $T|S = s$, i.e. given a dataset D of source-target pairs we define the *target entropy* of an utterance s as

$$H_{\text{tgt}}(s, D) = -\sum_{(s,t_i)\in D} p(t_i|s)\log_2 p(t_i|s)$$

Similarly, *source entropy* of an utterance can be defined as

$$H_{\text{src}}(t, D) = -\sum_{(s_i,t)\in D} p(s_i|t)\log_2 p(s_i|t)$$

The probabilities are calculated based on the observed relative frequency of utterance pairs in the data. After calculating source and target entropies for each utterance in a corpus, we filter the training data using one of 3 strategies. TARGET-BASED, where pairs are filtered if the source utterance has high target entropy. SOURCE-BASED, where we filter based on the source entropy of the target utterance. Finally, the ST-BASED dataset is obtained by filtering pairs based on both entropy values.

4 Experiments

4.1 Dataset

We use the DailyDialog dataset[1] (Li et al., 2017b) in our experiments. With 90 000 utterances in 13 000 dialogs, it is comparable in size with the Cornell Movie-Dialogs Corpus (Danescu-Niculescu-Mizil and Lee, 2011), but contains real-world high quality dialogs, instead of movie conversations, which are "not truthful representations of real-life conversations" (Danescu-Niculescu-Mizil and Lee, 2011). The vocabulary was set to 16384, covering most of the words in the corpus (roughly 19000).

4.2 Models

For dialog modeling we use `transformer` (Vaswani et al., 2017), a novel encoder-decoder architecture. Compared to the standard recurrent neural network (RNN) based `seq2seq` models, it doesn't use recurrent connections and relies only on attention mechanisms (Bahdanau et al., 2015). Consequently, it can be trained much faster, and using less memory (training the `seq2seq` model of (Vinyals and Le, 2015b) was not possible with the 8GB of GPU memory we had access to). We further justify the use of this model with the fact that it achieves state-of-the-art performance in NMT. Since the original `seq2seq` model was adopted from NMT (Cho et al., 2014) to dialog modeling, it is natural to do the same with the `transformer` architecture. To justify its use for the dialog task, we also train a `seq2seq` model (of limited size) on the same dataset for comparison. The `transformer` and `seq2seq` models contain 53M and 317M parameters, respectively. They are both large compared to the dataset thus they easily overfit it, as will be shown in Section 5. In the case of the `transformer` model we also experimented with different dropout (Srivastava et al., 2014) values our findings will also be presented in Section 5.

[1] http://yanran.li/dailydialog.html

We trained randomly initialized word embeddings (of size 512) together with the model parameters. Layer, attention, and relu dropout was set to 0.2, 0.1 and 0.1, respectively for the `transformer` model. At test time we used beam search with a beam size of 10 (Graves, 2012).

4.3 Filtered data

The 90 000 utterance pairs in the DailyDialog dataset contain about 72 000 unique utterances. We plot target entropies of source utterances in Figure 1, ranked from lowest to highest entropy, not showing the majority of utterances which have 0 entropy (i.e. they do not appear with more than one target). Source entropies of target utterances are very similar. In the following experiments we shall discard utterance pairs whose target and/or source entropy is greater than 1. This affects 5.64%, 6.98% and 12.24% of the data, for the TARGET-BASED, SOURCE-BASED and ST-BASED scenario, respectively.

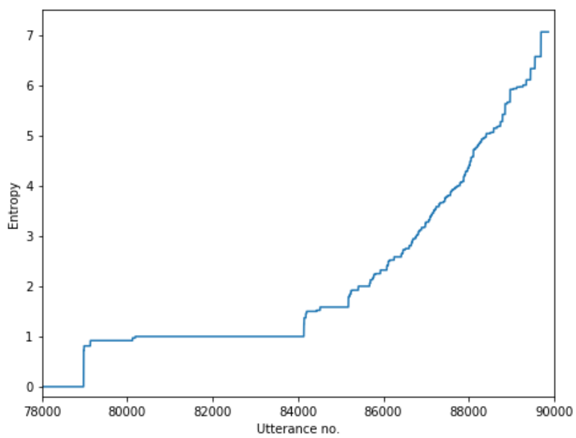

Figure 1: Source utterances by target entropy

Entropy is clearly proportional to utterance frequency (Figure 2), however we found that only 485 utterances overlap in the top 700 utterances (roughly what gets discarded) when ordered by both entropy and frequency, and those that are different in the frequency ordered list are long utterances, that we don't wish to filter out. Entropy offers a more fine-grained measure compared to frequency, and in the case of low frequency pairs, this is especially helpful. For example, all utterances that have a frequency of 3, are in the same category based on frequency, but their entropy can range from 0 to $\log_2 3 \approx 1.58$, which would be over our filtering threshold.

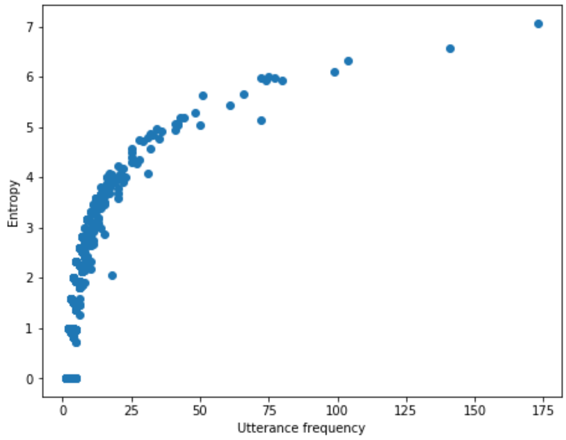

Figure 2: Target entropies of sources with respect to utterance frequency.

After noticing that high-entropy utterances are relatively short, we also examined the relationship between entropy and utterance length (Figure 3). Given the relationship between frequency and entropy it comes as no surprise that longer sentences have lower entropy, although this effect is less pronounced in the range affected by filtering.

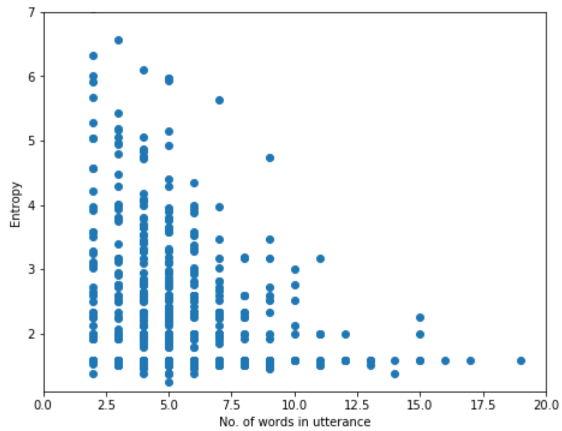

Figure 3: Target entropies of sources with respect to utterance length.

5 Results

5.1 Evaluation Principles

Due to the limited scope of automatic evaluation (Liu et al., 2016), our claims are supported by both qualitative and quantitative evaluation of our models' outputs. We first summarize the principles used for comparison, then present our findings.

We shall evaluate models based on answers they give to a list of input utterances. When comparing responses given to the same question, we first consider their *coherence*, i.e. whether they could have been written by a human speaker. Models that return coherent answers to an input are further compared based on whether the answer is boring/generic (e.g. *I don't know*) or engaging/specific. Generic responses make sense in more dialog histories, however because of this they do not add new information to the conversation. More engaging responses, i.e. those that can further a conversation, are preferred, but only if they are coherent with the source utterance. While these principles, and thus our judgments presented here, are quite subjective, we believe that the differences observed between various trainings are sufficiently pronounced, and our findings are also grounded by the quantitative analysis.

Four metrics are used to quantitatively evaluate our models in Section 5.5. In order to measure the information content of a models' responses, average utterance length $|U|$, word entropy H_w and utterance entropy H_u is computed (Serban et al., 2017b). The entropies are computed with respect to the maximum-likelihood unigram distribution of the training set. Thus: $H_w = -\sum_{w\in U} p(w)\log(p(w))$, and H_u is simply the product of utterance length and word entropy. Additionally the (string-based) average jaccard similarity J between source utterances and model responses is computed, measuring their coherence and relevance.

5.2 `seq2seq`, `transformer`, and overfitting

We first compare overfitted and non-overfitted versions of the `seq2seq` (S2S) and `transformer` (TRF) models trained on unfiltered data. Overfitted model versions, are those that were trained further after the lowest value of validation loss was reached, until training loss converges. We used a list of 34 general input utterances chosen from the ones used in Vinyals and Le (2015b), which we will call the NCM test inputs. We filtered this list by cutting well-known inputs to which each model learns to respond well (eg. *Hello!*, *How are you?*), and also removed inputs where any word was missing from the vocabulary.

For each input we determined the responses which we judged best, based on the principles outlined at the beginning of this section (Table 1). Also, we present the best response counts in Table 2, from which the first col-

Input	S2S-BASE	S2S-BASE-O	TRF-BASE	TRF-BASE-O
what 's your name ?	my name is james . my name is tiang .	**my name is wang lin . my english name is...CONT.**	**john .**	**my name is james .**
when were you born ?	i was born in beijing . i was born .	i was born in beijing but i spent most of...CONT.	i was born in beijing .	i was born in .
what year were you born ?	i was born in beijing . i was born .	i was born in beijing but i spent most of...CONT.	i was born in beijing .	i was born in xinjiang .
where are you from ?	**i 'm from germany . how about you ?**	**i 'm from china . what about you ?**	i am from china .	i 'm from germany .
are you a man or a woman ?	**i 'm not sure . what about you ?**	yes he is . he s climbing the tree .	i am .	he is five years old .
why are we here ?	i do n't know what to say .	i do n't know . the weatherman did n't .	it looks good .	**it 's just near the dong si shi tiao subway...CONT.**
my name is david . what is my name ?	this is mary .	my name is <unk>.	nice to meet you .	**hi david . nice to meet you .**
my name is john . what is my name ?	this is mary .	my name is <unk>.	nice to meet you .	**hi lee . nice to meet you .**
who is bill clinton ?	i do n't know . he 's a liar .	it 's martin scorsese . you liked his film .	that 's john .	that s the superintendent of this building .
is sky blue or black ?	it 's a little tight in the back .	i 'm not sure . what about you ?	**black please .**	**white .**
does a cat have a tail ?	yes of course . it s a good idea .	no he does n't . he 's an engineer .	yes we do .	as a matter of fact we have n't .
does a cat have a wing ?	yes he does . he likes it very much .	yes he does . he is a science fiction writer...CONT.	**no they do n't .**	yes but not necessarily .
can a cat fly ?	of course . that s a good idea .	yes they are . and driving is very good .	yes that 's right .	yes cats are fine . and they do n't need...CONT.
how many legs does a cat have ?	i think it 's a good idea .	three . two boys and a pair of them .	about minutes .	oh come on . we have only one day .
how many legs does a spider have ?	**let me see . . . here it is .**	**15 . and for a hard one only 6 .**	oh no .	only one year .
what is the color of the sky ?	it 's the size of the size .	red . it 's made of fine leather shoes .	**it 's brown .**	it s a brandnew seiko wrist watch .
what is the color of water ?	it 's the best we can do it .	leopard . it s famous for its quality only .	**it 's red .**	**it 's red .**
what is the color of blood ?	it s the most important thing to me .	it is made of wood not plastic .	**it 's white .**	your doctor has requested a urine sample .
what is the usual color of a leaf ?	it 's a <unk>.	it 's a <unk>with a vinyl roof .	**it 's white .**	**it 's blue .**
what is the color of a yellow car ?	it 's a <unk>.	it 's a luminous watch with a <unk>function .	**it 's blue .**	**it 's a blue nissan minivan .**
how much is two plus two ?	**let me see . . . here it is .**	15 . and for a hard one only 6 .	.	altogether .
what is the purpose of life ?	i think it 's the most important thing to me...CONT.	i 'll be studying . i 'm doing an mba...CONT.	i 'm not sure yet .	for sightseeing .
what is the purpose of living ?	i think it 's the most important thing to me...CONT.	i 'll be studying . i 'm doing an mba...CONT.	i 'm a personnel manager .	it 's a small family .
where are you now ?	**on the second side of the campus .**	**i am on route 80 a few miles east .**	i 'm from canada .	**i am on route a few miles east of the...CONT.**
what is the purpose of dying ?	**it s the most important thing to me .**	they have a really good dj and a big dance...CONT.	he is a lawyer .	**it takes a long time .**
what is the purpose of being intelligent ?	the most important thing is the most important thing .	**i have no idea . i 'm impressed .**	i do n't know .	i do n't know .
what is moral ?	**nothing .**	**pardon ?**	she 's a lawyer .	it is largest pop square in the world .
give me some examples of moral actions . . .	what ?	so what ?	what 's wrong ?	anything else ?
what is integrity ?	the most important thing is the most important thing .	you are suspected of hiding illegal drugs .	**the main thing is to people .**	the y . h . a . mun . the...CONT.
be moral !	what ?	what 's the matter ?	what 's wrong ?	and your name ?
what do you like to talk about ?	i do n't know what to say .	**i want to talk about this year s election .**	**i 'd like to talk to you about it .**	i do n't like her . ok .
what do you think about bill gates ?	i think it 's a good idea .	**well i heard people say he has a bad cold...CONT.**	i 'm not sure .	**well he had a lot of nerve telling us this...CONT.**
what is your job ?	i would like to work on my own .	i m a keyboard operator . what s your job...CONT.	i have worked as a personnel manager .	i 'm a bank manager .
what do you do ?	i do n't know how to use it .	**i have my own company that designs computer systems .**	**i 'm a student .**	**i m a podiatrist . what about you ?**

Table 1: Comparison between the two models (seq2seq and transformer) trained on unfiltered data, and between overfitted and non-overfitted variants. The input utterance is in the left-most column, the other columns contain responses by the various models. S2S and TRF represent seq2seq and transformer respectively, and the *O* notation in the model name means that it is an overfitted version. In each row we highlighted the best responses.

Unfiltered trainings	NCM test set	High entropy test set	NCM test set (2)
S2S-BASE	7	-	-
S2S-BASE-O	9	-	-
TRF-BASE	***11***	12	15
TRF-BASE-O	**12**	-	-
Filtered trainings			
TRF-ST-BASED	-	**23**	15
TRF-ST-BASED-O	-	7	13
TRF-TARGET-BASED	-	-	11
TRF-SOURCE-BASED	-	11	**16**

Table 2: Qualitative best response counts based on the different test sets. The test sets are the same as in the qualitative results. Since we evaluated separately the filtered and unfiltered trainings qualitatively there are two NCM test set columns. First, trainings which were trained on the normal (unfiltered) dataset are presented, and then trainings run on the filtered datasets. TRF refers to the `transformer` model, and S2S refers to the `seq2seq` model. The type of filtering is also noted (ST-BASED, SOURCE-BASED, TARGET-BASED), and the O notation means that it is an overfitted version of the model. Results of best non-overfitted models are in italic boldface, while best results overall are noted by simple boldface.

umn is of relevance to this section. Generally, the `transformer` performed better than the `seq2seq` model, achieving 11 best responses (among the 4 models), compared to only 7 for `seq2seq`. It managed to output colours when asked about the color of objects, while the `seq2seq`'s replies were irrelevant.

Overfitted models performed at least as well (in human evaluation) as non-overfitted models, strengthening the points raised in Csáky (2017); Tandon et al. (2017): the loss function does not adequately represent the quality of a chatbot model. The overfitted `seq2seq` model achieved 9 best responses (2 more than the non-overfitted version), and in the case of the `transformer` the overfitted version achieved 12 best responses. We note that overfitted models tend to generate longer responses, which is generally good, but in some cases we obtain too specific and probably memorized responses to unrelated inputs (eg. *they have a really good dj and a good dance.* to the input *what is the purpose of dying*).

Since our models overfit quickly, we also experimented with dropout. With a high dropout rate (0.5) we can essentially force the validation loss to stay at its minimum for longer, before starting to overfit. However, the minimum does not go lower compared to low dropout (0.2) trainings, and the replies were generally the same even after training more with high dropout, further consolidating the observation that the validation loss minimum does not represent the best state of the model.

5.3 High Entropy Inputs

We evaluate the `transformer` model trained on unfiltered and filtered datasets (according to the 3 filtering types discussed in Section 3) on the 45 highest entropy source utterances (Table 3). These are the most challenging utterances (eg. *yes; thank you; why?; sure; no; what's that?; here you are*), where dialog models tend to fail, because of the high diversity observed in the dataset. The TARGET-BASED filtering variant is excluded from this evaluation, because as we will see in Section 5.4 it performs poorly on the NCM test set, and also according to the automatic metrics (Section 5.5).

Counter-intuitively there is clear improvement in the performance on these utterances which the filtered models didn't see during training. The second column of Table 2 summarizes the best response counts for this section. The ST-BASED training gives the best response in 23 cases, while the unfiltered training only in 12 cases. Solely filtering the target side, gives slightly worse results, achieving only 11 best responses, however its responses can be often selected as the second-best after the ST-BASED. A closer look at the ST-BASED replies shows two main enhancements. First, the model was able to generate more diverse responses, while also keeping them general enough (eg. *I have a bad headache.*, or *I'm glad to hear that.*), which is probably mostly due the source-side filtering. Second, where the unfiltered model often choose to output the same safe re-

Input	TRF-BASE	TRF-FILTERED (ST-BASED)	TRF-FILTERED-O (ST-BASED)	TRF-FILTERED (SOURCE-BASED)
yes .	how long have you been here ?	do you have any other ideas ?	what would you like for a drink ?	what 's your name ?
certainly .	thank you very much .	**i 'm glad to hear that .**	ok so you are already a gold card holder with...CONT.	how about the lyrics ?
yes please .	**how do you like it ?**	**how about this one ?**	here is your laundry . i am sorry to tell...CONT.	here is your laundry sir .
yes ?	can you tell me how to use the machine ?	would you like to go with me ?	i was just thinking about someone i went to school...CONT.	what 's your name ?
yes i do .	how do you like it ?	do you like it ?	what would you like for a drink ?	what 's your name ?
here you are .	thank you .	do you have a receipt with you ?	would you please open the suitcase ? what are these...CONT.	here is your receipt .
here you go .	thank you .	how much do you want to pay ?	do you have a credit card with us ?	here is your receipt .
yes it is .	**how much is it ?**	**can you tell me how to use it ?**	we 're very sorry for the delay . i 'll...CONT.	what 's your name ?
here it is .	thank you .	**how much is it ?**	**how much altogether ?**	here is your receipt .
that 's right .	thank you .	**do you have any other ideas ?**	it must be very convenient and pretty quick .	how about the lyrics ?
ok .	thank you .	**do you want to go with me ?**	and when we get back from the doctor 's you...CONT.	how about the lyrics ?
okay .	thank you .	**do you want to go with me ?**	**so what were you calling me about ?**	how about the lyrics ?
sure .	thank you .	**do you have any other ideas ?**	ok . . . i 'm just going to lay...CONT.	how about the lyrics ?
i see .	thank you .	**would you like to go with me ?**	**which one interests you sir ?**	how about the lyrics ?
all right .	thank you .	**do you have any other ideas ?**	the hair stylist cuts the customer 's hair . almost...CONT.	how about some fish ?
no problem .	**thank you very much .**	i 'm sorry to hear that .	what is the price of the suite ?	how about the lyrics ?
of course .	thank you very much .	**do you have any other ideas ?**	how about the lyrics ?	how about the lyrics ?
what do you mean ?	you are suspected of hiding illegal drugs .	**i mean you 're right .**	when you take something out of the stores you need...CONT.	we 're stranded here tom .
is that right ?	**yes that 's right .**	**yes that 's right .**	it must be very convenient...CONT.	yes that 's it .
may i help you ?	yes i 'm looking for a pair of shoes .	i 'm looking for a sweater .	yes i 'm looking for a sweater .	yes i 'm looking for a sweater .
how may i help you ?	i 'm looking for a pair of shoes .	i 'm looking for a sweater .	i need to open a second account .	i need to withdraw some money .
what can i do for you ?	i 'm looking for a pair of shoes .	i 'm looking for a sweater .	well i 'm looking for a job .	i need to withdraw some money .
is there anything i can do for you ?	**yes i 'm looking for a pair of shoes .**	yes i am . thank you very much .	**yes . i was looking for some help .**	**yes i 'm looking for a sweater .**
anything else ?	**no that 's all .**	**no nothing else .**	i have taken some courses in college such as translation...CONT.	i 'd like to have some sweetandsour sauce and pepper .
how much is it ?	yuan .	**59 . 95 .**	**55 dollars .**	**180 yuan .**
how about this one ?	it looks good .	that 's fine . i 'll take it .	the style is all right but i 'm not sure...CONT.	it looks nice . how much is it ?
really ?	yes .	yes it is .	yes and not expensive either .	yes it is .
are you sure ?	yes i am .	yes i am .	yeah ! why not ? they are trendy and look...CONT.	yes i am sure i can .
what ?	you are suspected of hiding illegal drugs .	**i 'm a vegan .**	a vegan . i do n't eat or use any...CONT.	what 's the matter with you ?
why is that ?	i do n't know .	i do n't know .	because you can see the stars perfectly .	it 's just a waste of money .
what 's that ?	it 's filthy !	**it 's a brick !**	well you know i have some students studying for their...CONT.	**it 's a kind of job .**
what happened ?	i did n't know .	**it was nothing .**	i told you to get this consignment out last week...CONT.	**i went to the library .**
what is it ?	**it 's a polo .**	**it 's a new car .**	it only has 4 beds and 4 writing desks but...CONT.	**it 's a polo . it has the logo on the back .**
what 's the matter ?	i do n't know .	**i have a bad headache .**	**i 'm a little worried about my host family ....CONT.**	**my motherinlaw just went into the hospital in...CONT.**
what seems to be the problem ?	i do n't know what to do .	i have been working on a business trip .	my roommate always brings a lot of friends from <unk>...CONT.	i have a problem with my work .
why ?	i do n't know .	i have no idea .	well our public schools here are not very good and...CONT.	**it 's just a waste of money .**
how come ?	**it does n't matter .**	not too bad .	my boss gave me a big project . i had...CONT.	**it 's too old .**
why not ?	**it does n't matter .**	i do n't know .	video tape breaks down after fifteen years . so if...CONT.	**it 's just a waste of money .**
like what ?	can you tell me how to use the internet ?	i like it very much .	i love you too .	how about the mongolian hot pot ?
thank you very much .	**you 're welcome .**	**have a great day !**	please hold the line . i ll see if there...CONT.	**it 's my pleasure .**
thanks a lot .	you 're welcome .	you 're welcome .	you are welcome	do n't mention it .
thanks .	**you 're welcome .**	**you 're welcome .**	look ! here it is !	**you 're welcome . have a nice day .**
no .	how about this one ?	how much is it ?	what did you do last night ?	how about the lyrics ?
hello ?	yes what can i do for you ?	hi tom . how are you ?	no response silence	hello . this is john .
good morning .	**good morning . can i help you ?**	**good morning ! can you help me please ?**	**good morning jamie . what time is it now ?**	may i help you ?

Table 3: Responses to high entropy inputs. Training name explanation: TRF refers to the `transformer` model, and the O notation means that it is an overfitted version. BASE and FILTERED mean that the model is trained on unfiltered and filtered data, respectively. The filtering strategies are noted in parenthesis.

Input	TRF-BASE	TRF-FILTERED (ST-BASED)	TRF-FILTERED-O (ST-BASED)	TRF-FILTERED (TARGET-BASED)	TRF-FILTERED (SOURCE-BASED)
what 's your name ?	**john .**	**it 's marcia gleason .**	**anderson .**	**my name is mona white .**	**my name is judy lin .**
when were you born ?	i was born in beijing .	i was born in beijing .	i was born in beijing in <unk>.	i was born and raised in beijing .	i was born in xinjiang .
what year were you born ?	i was born in beijing .	i was born in beijing .	i was born in xinjiang .	i was born and raised in beijing .	i was born in xinjiang .
where are you from ?	**i am from china .**	**i 'm from new york .**	i am still in xyz company this meeting took longer...CONT.	**i 'm from new york .**	**i 'm from china .**
are you a man or a woman ?	i am .	**i am not sure .**	that sounds right .	i do n't think so .	i m in the middle of something .
why are we here ?	it looks good .	i 'm not sure .	**it 's 12**	i do n't know .	**the view is breathtaking .**
my name is david . what is my name ?	nice to meet you .	**nice to meet you david .**	james .	nice to meet you .	**how do you spell your last name ?**
my name is john . what is my name ?	nice to meet you .	nice to meet you .	this is john white .	nice to meet you .	nice to meet you .
who is bill clinton ?	**that 's john .**	she 's out of town .	he is .	**tom .**	he 's still sick .
is sky blue or black ?	**black please .**	i do n't like it very much .	**it does n't matter .**	fine thank you .	smoking is fine .
does a cat have a tail ?	**yes we do .**	**yes we do .**	i 'm sorry . we have no filet mignon ....CONT.	**yes we do .**	no we do n't .
does a cat have a wing ?	**no they do n't .**	no i didn t .	i am sorry .	**no ma 'am .**	**no not yet .**
can a cat fly ?	yes that 's right .	yes they do .	no kidding .	sure .	no problem .
how many legs does a cat have ?	about minutes .	**three .**	they 're rice .	**about five .**	**two .**
how many legs does a spider have ?	oh no .	it 's ok .	seriously but they are getting close .	**oh about five .**	**six .**
what is the color of the sky ?	it 's brown .	it 's very old .	**it 's blue .**	it 's brown .	it 's a lovely day .
what is the color of water ?	**it 's red .**	**it 's blond .**	**it s the same color .**	**it 's red .**	**it 's red .**
what is the color of blood ?	**it 's white .**	**it 's blond .**	it 's here .	the shoulder .	**it 's black .**
what is the usual color of a leaf ?	**it 's white .**	i do n't like it .	to tell you the truth it 's really hot .	i do n't think so .	you can always do something to drink .
what is the color of a yellow car ?	**it 's blue .**	**it 's blue .**	**it 's a blue nissan .**	**it 's a red one .**	**it 's a red one .**
how much is two plus two ?	.	**15 .**	**it 's 150 .**	150 a month .	that 's 10 yuan .
what is the purpose of life ?	i 'm not sure yet .	i do n't know .	i do n't know . he has an olympic expert...CONT.	my mother is a lawyer .	i 'm in a mechanized farm .
what is the purpose of living ?	i 'm a personnel manager .	i 'm a sales manager .	you know ? i ca n't help myself .	he 's a famous american musician .	i 'm in charge of marketing .
where are you now ?	i 'm from canada .	i 'm from new york .	**i am on route 80 a few miles east of...CONT.**	i 'm from new york .	**i 'm going to the railway station .**
what is the purpose of dying ?	he is a lawyer .	it s plenty of time .	business .	china .	it s a passport .
what is the purpose of being intelligent ?	**i do n't know .**	it 's <unk>.	you know who it is .	**i have no idea .**	**there are several things you want .**
what is moral ?	she 's a lawyer .	it is .	tomb sweeping day .	that 's right .	**just a moment please .**
give me some examples of moral actions . . .	what 's wrong ?	**no problem .**	here are your passports and tickets .	that 's good .	ok . thank you .
what is integrity ?	**the main thing is to people .**	it 's the powell orchestra .	you 'll find the bread .	what kind of things do you need ?	we need a lot of things .
be moral !	what 's wrong ?	that 's too bad .	**for papa ?**	thank you .	**just a moment please .**
what do you like to talk about ?	**i 'd like to talk to you about it .**	**i want to talk to her about it .**	**i do n't think we 'll talk about it .**	i do n't like it .	she 's a teacher .
what do you think about bill gates ?	**i 'm not sure .**	**well i 'm not sure .**	**well i 'm not really sure .**	well they were playing cards .	well i 'm just thinking of buying them .
what is your job ?	i have worked as a personnel manager .	**i 'm a bank manager .**	**i 'm a bank manager .**	**i m a senior manager in a publishing company .**	**i 'm a sales manager .**
what do you do ?	**i 'm a student .**	**i 'm a clerk in a shop .**	**i 'm a clerk in a shop .**	i do n't know .	**i work in a publishing house . how about you ?**

Table 4: Results on the NCM test inputs. TRF-BASE refers to the non-filtered `transformer` training, and the others are `transformer` trainings on the filtered dataset. We note the filtering strategies in parentheses. The *O* notation in the training name means that it is an overfitted version.

Unfiltered trainings	$\|U\|$	H_w	H_u	J
TRF-BASE	4.93 (1.96)	0.491 (0.178)	2.68 (2.12)	0.0909 (0.0970)
TRF-BASE-O	9.82 (8.25)	0.795 (0.555)	12.1 (24.9)	0.0986 (0.0850)
S2S-BASE	4.35 (5.41)	0.462 (0.485)	4.54 (54.7)	0.0889 (0.0944)
S2S-BASE-O	7.09 (6.0)	0.628 (0.418)	6.73 (26.9)	0.0979 (0.0974)
Filtered trainings				
TRF-ST-BASED	6.31 (1.97)	0.586 (0.211)	4.0 (2.65)	0.0988 (0.0977)
TRF-ST-BASED-O	**10.42** (7.74)	**0.838** (0.522)	**12.4** (23.2)	**0.101** (0.0830)
TRF-TARGET-BASED	5.25 (2.82)	0.525 (0.480)	3.75 (51.5)	0.0961 (0.0980)
TRF-SOURCE-BASED	***6.81*** (2.90)	***0.61*** (0.336)	***4.78*** (20.5)	***0.0995*** (0.0946)
Targets	14.1 (10.9)	1.03 (0.713)	21.8 (58.3)	0.105 (0.0830)

Table 5: Quantitative metrics computed based on the test set. First, trainings which were trained on the normal (unfiltered) dataset are presented, and then trainings run on the filtered datasets. TRF refers to the `transformer` model, and S2S refers to the `seq2seq` model. The type of filtering is also noted (ST-BASED, SOURCE-BASED, TARGET-BASED), and the O notation means that it is an overfitted version of the model. Results of best non-overfitted models are in italic boldface, while best results overall are noted by simple boldface. Numbers in parentheses are the respective standard deviations.

sponse (*thank you.*), the filtered model responds with engaging questions to further the conversation. This is clearly due to the target side filtering, since the model was forced to not learn to output generic responses. The conclusion is further reinforced by the SOURCE-BASED training, where the model answers with questions more frequently. However, the SOURCE-BASED training is still not diverse enough, combining the two methods seems the most advantageous. We also experimented with an overfitted variant of the ST-BASED training, which performed a lot worse, and was too specific in many cases (giving the best response only in 7 cases). Overall it appears that with our filtered dataset the model performs better at the validation loss minimum.

5.4 NCM test inputs

We also evaluate the `transformer` model trained on unfiltered and filtered datasets on the NCM test inputs (Table 4). The best response counts from the third column of Table 2 are related to this section. The ST-BASED and SOURCE-BASED trainings are on par with the unfiltered training (15, 16, 15 best responses, respectively), followed by the TARGET-BASED training (11 best responses). These results prove that the model is still capable to output good responses to the general NCM test inputs, even when trained on the filtered dataset. Filtering the source side alone gives worse results than filtering the target side alone, demonstrating that discarding generic responses adds more to conversational quality.

Finally, the overfitted version of the ST-BASED training performs slightly worse (getting best response in only 13 cases), somewhat alleviating the problems discussed in Section 5.2. As with the high entropy inputs, this indicates that filtering a dataset based on entropy, makes the learning problem more aligned with the loss function.

5.5 Quantitative Analysis

In Table 5 all metrics are computed based on responses given to a separate test set, containing 10% of the utterances from DailyDialog. Looking only at the unfiltered trainings we can see that the `transformer` performed better than the `seq2seq` model across all metrics except the utterance entropy. Furthermore, the `seq2seq` models' results have much higher variance, which means that the quality of the responses is more unreliable, especially in the case of utterance entropy, showing that perhaps the higher mean value doesn't actually equate to an increase in quality. In contrast to the manual evaluations however, on automatic metrics all examined overfitted models performed much better than their non-overfitted counterparts, but should be noted that they all have high variance, meaning more unreliable responses.

The results of the filtered trainings are also presented in Table 5. It is clear that all types of filtering show significant improvement across nearly all metrics. Interestingly, in contrast to the manual evaluations the SOURCE-BASED filtering achieves the best results, ST-BASED being the second best, and aligned with the manual evaluations TARGET-

BASED is the last. Using SOURCE-BASED filtering alone, and thus filtering boring and generic responses is more important than TARGET-BASED filtering, and combining the two types is not beneficial according to these metrics. It should be noted however that the SOURCE-BASED training results have much higher variance, meaning that perhaps the ST-BASED responses are actually better, because of being more reliable despite the lower average metric values.

Also, the overfitted version of the ST-BASED training achieves the best performance, improving on the unfiltered training variant (but still having high variance). Thus, while training on filtered datasets generally improves performance, a non-overfitted model still can't be competitive with an overfitted variant. However the performance gap between them gets somewhat smaller than in the case of unfiltered trainings. This also shows the limitation of these metrics that value diversity, since as seen in the manual evaluation, overfitted models tend to be too specific, by outputting learned responses.

6 Future Work

We showed how with a simple entropy-based approach we can find generic and safe sources/targets that usual dialog models have problems with. We compared the various trainings in an extensive qualitative and quantitative evaluation. The unfiltered and the filtered trainings were compared on two different test sets, and on several automatic metrics used in the literature. We showed how the model trained on the filtered dataset outputs more engaging and interesting responses to inputs that it has never seen. Moreover, the `transformer` was shown to be at least as good for dialog modeling as the `seq2seq`, and evaluating these models trained on unfiltered data at an overfitted point results in better conversational quality, while training on filtered data somewhat alleviates this issue.

For future work we wish to explore two main objectives. First, we want to test our methods using various experiments. This includes experimenting with different datasets like the Cornell Movie-Dialogs dataset (Danescu-Niculescu-Mizil and Lee, 2011) or the Persona-Chat dataset (Zhang et al., 2018). We would also like to test our method using a different, more state-of-the-art model for dialog modeling, the VHCR (Park et al., 2018). This model can handle more previous utterances so it would be interesting to see how our method can help in this case. Also we would like to test our method with a popular augmentation to dialog models, the persona. The persona is simply a unique token representing each persona in a dialog dataset. This helps dialog models to ground responses based on the persona of the input utterance. We would also like to perform stop word filtering of our dataset before using entropy-based filtering. Stop word filtering can help keep the focus of the entropy calculation on words that are truly relevant.

Second we want to give a better qualitative evaluation of our method. For this we would like to use Amazon Mechanical Turk[2] (MTurk), a widely used service in the dialog modeling literature. With MTurk we can let other people judge the quality of the responses from the various experiments, giving us an unbiased evaluation. Finally, we would also like to increase the scope of the quantitative evaluation. For this we would add most of the metrics used in (Shen et al., 2018), since these offer a complete view of the response quality and are also widely used.

References

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2015. Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR 2015)*.

Kyunghyun Cho, Bart van Merriëenboer, Caglar Gulcehre, Fethi Bougares, Holger Schwenk, and Yoshua Bengio. 2014. Learning phrase representations using rnn encoder-decoder for statistical machine translation. In *Proceedings of the 2014 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, pages 1724–1734, Doha, Qatar.

Richárd Csáky. 2017. Deep learning based chatbot models. Technical report, Budapest University of Technology and Economics, Budapest.

Cristian Danescu-Niculescu-Mizil and Lillian Lee. 2011. Chameleons in imagined conversations: A new approach to understanding coordination of linguistic style in dialogs. In *Proceedings of the 2nd Workshop on Cognitive Modeling and Computational Linguistics*, pages 76–87. Association for Computational Linguistics.

Kartik Goyal, Graham Neubig, Chris Dyer, and Taylor Berg-Kirkpatrick. 2017. A continuous relaxation of beam search for end-to-end training of neural sequence models. *arXiv preprint arXiv:1708.00111*.

[2]https://www.mturk.com/

Alex Graves. 2012. Sequence transduction with recurrent neural networks. In *Representation Learning Workshop, ICML 2012*, Edinburgh, Scotland.

Chaitanya K Joshi, Fei Mi, and Boi Faltings. 2017. Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

A. Krizhevsky and G. Sutskever, I.and Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *NIPS'2012*.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2015. A diversity-promoting objective function for neural conversation models. *arXiv preprint arXiv:1510.03055*.

Jiwei Li, Michel Galley, Chris Brockett, Georgios P Spithourakis, Jianfeng Gao, and Bill Dolan. 2016a. A persona-based neural conversation model. *arXiv preprint arXiv:1603.06155*.

Jiwei Li, Alexander H Miller, Sumit Chopra, Marc'Aurelio Ranzato, and Jason Weston. 2016b. Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016c. Deep reinforcement learning for dialogue generation. *arXiv preprint arXiv:1606.01541*.

Jiwei Li, Will Monroe, Tianlin Shi, Alan Ritter, and Dan Jurafsky. 2017a. Adversarial learning for neural dialogue generation. *arXiv preprint arXiv:1701.06547*.

Yanran Li, Hui Su, Xiaoyu Shen, Wenjie Li, Ziqiang Cao, and Shuzi Niu. 2017b. Dailydialog: A manually labelled multi-turn dialogue dataset. *arXiv preprint arXiv:1710.03957*.

Zachary Lipton, Xiujun Li, Jianfeng Gao, Lihong Li, Faisal Ahmed, and Li Deng. 2017. Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. *arXiv preprint arXiv:1711.05715*.

Chia-Wei Liu, Ryan Lowe, Iulian V Serban, Michael Noseworthy, Laurent Charlin, and Joelle Pineau. 2016. How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. *arXiv preprint arXiv:1603.08023*.

Huiting Liu, Tao Lin, Hanfei Sun, Weijian Lin, Chih-Wei Chang, Teng Zhong, and Alexander Rudnicky. 2017. Rubystar: A non-task-oriented mixture model dialog system. *arXiv preprint arXiv:1711.02781*.

Tomáš Mikolov. 2010. Recurrent neural network based language model. Presentation at Google.

Tomas Mikolov, Kai Chen, Greg Corrado, and Jeffrey Dean. 2013. Efficient estimation of word representations in vector space. International Conference on Learning Representations (ICLR 2013).

Yookoon Park, Jaemin Cho, and Gunhee Kim. 2018. A hierarchical latent structure for variational conversation modeling. *arXiv preprint arXiv:1804.03424*.

Iulian V Serban, Chinnadhurai Sankar, Mathieu Germain, Saizheng Zhang, Zhouhan Lin, Sandeep Subramanian, Taesup Kim, Michael Pieper, Sarath Chandar, Nan Rosemary Ke, et al. 2017a. A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

Iulian Vlad Serban, Alessandro Sordoni, Yoshua Bengio, Aaron C Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

Iulian Vlad Serban, Alessandro Sordoni, Ryan Lowe, Laurent Charlin, Joelle Pineau, Aaron C Courville, and Yoshua Bengio. 2017b. A hierarchical latent variable encoder-decoder model for generating dialogues. In *AAAI*, pages 3295–3301.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2200–2209.

Xiaoyu Shen, Hui Su, Wenjie Li, and Dietrich Klakow. 2018. Nexus network: Connecting the preceding and the following in dialogue generation. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 4316–4327.

Nitish Srivastava, Geoffrey E Hinton, Alex Krizhevsky, Ilya Sutskever, and Ruslan Salakhutdinov. 2014. Dropout: a simple way to prevent neural networks from overfitting. *Journal of Machine Learning Research*, 15(1):1929–1958.

I. Sutskever, O. Vinyals, and Le. Q. V. 2014. Sequence to sequence learning with neural networks. In *Advances in Neural Information Processing Systems*, pages 3104–3112.

Shubhangi Tandon, Ryan Bauer, et al. 2017. A dual encoder sequence to sequence model for open-domain dialogue modeling. *arXiv preprint arXiv:1710.10520*.

Aäron Van Den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew W Senior, and Koray Kavukcuoglu. 2016. Wavenet: A generative model for raw audio. In *SSW*, page 125.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Ł ukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In I. Guyon, U. V. Luxburg, S. Bengio,

H. Wallach, R. Fergus, S. Vishwanathan, and R. Garnett, editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

Oriol Vinyals and Quoc Le. 2015a. A neural conversational model.

Oriol Vinyals and Quoc Le. 2015b. A neural conversational model. *arXiv preprint arXiv:1506.05869*.

Bolin Wei, Shuai Lu, Lili Mou, Hao Zhou, Pascal Poupart, Ge Li, and Zhi Jin. 2017. Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. *arXiv preprint arXiv:1712.02250*.

Sam Wiseman and Alexander M Rush. 2016. Sequence-to-sequence learning as beam-search optimization. *arXiv preprint arXiv:1606.02960*.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *AAAI*, pages 3351–3357.

Saizheng Zhang, Emily Dinan, Jack Urbanek, Arthur Szlam, Douwe Kiela, and Jason Weston. 2018. Personalizing dialogue agents: I have a dog, do you have pets too? *arXiv preprint arXiv:1801.07243*.

Tiancheng Zhao, Ran Zhao, and Maxine Eskenazi. 2017. Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. *arXiv preprint arXiv:1703.10960*.

Hao Zhou, Minlie Huang, Tianyang Zhang, Xiaoyan Zhu, and Bing Liu. 2017. Emotional chatting machine: Emotional conversation generation with internal and external memory. *arXiv preprint arXiv:1704.01074*.