BUDAPEST UNIVERSITY OF TECHNOLOGY AND ECONOMICS
FACULTY OF ELECTRICAL ENGINEERING AND INFORMATICS
DEPARTMENT OF AUTOMATION AND APPLIED INFORMATICS

The Gutenberg Dialog Dataset for Neural Conversational Modeling

SCIENTIFIC STUDENTS' ASSOCIATIONS REPORT

Author:
Richárd Krisztián Csáky

Supervised by
Gábor Recski

2019

Kivonat

Létrehozunk egy új, nagy, és jó minőségű adathalmazt neurális dialógus modellezésre, és bemutatjuk előnyeit jelenlegi adathalmazokhoz képest. A konverzációs ágens (chatbot) egy olyan program mely emberekkel képes kommunikálni természetes nyelven, feldolgozva a felhasználó bemenetét és releváns és érdekes válaszokat adva. Míg a múltban a szabály-alapú modellek népszerűek voltak, manapság mély tanulás alapú modellek dominálják a dialógus modellezés területét. A nagy adathalmazokon való neurális háló alapú konverzációs ágensek tanításának paradigmája fontos kérdéseket vet fel, hogy az adatminőség hogyan befolyásolja ezeket a chatbotokat, és milyen evaluációs módszerekkel tudjuk hatékonyan felmérni a betanított modellek teljesítményét. Ez a pályamunka megpróbál pár kérdésre választ adni, a releváns háttér feltárásával és egy új dialogus adathalmaz bemutatásával.

Munkám első részében bemutatásra kerülnek a jelenlegi általános dialógus adathalmazok az irodalomból, majd bemutatjuk hogyan hoztunk létre egy új, nagy és jó minőségű adathalmazt. A dialógusok a Gutenberg Projekt[1] online könyveiből vannak kinyerve. Egy részletes adat elemzést mutatunk be, és megmagyarázzuk a hiperparaméterek és előfeldolgozási lépések mögötti okokat, egy minél jobb minőségű adathalmaz létrehozásának érdekében. Továbbá egy részletes hibaelemzést is adunk, mind mondat, mind dialógus szinten.

Munkám következő részében evaluáljuk az adathalmazt, más nagy adathalmazokhoz hasonlítva transzfer tanulási kísérlet keretében. Amellett érvelünk, hogy a mi adatunkon előtanítva jobb eredményeket lehet elérni kisebb downstream adatokon. Továbbá, az adathalmazunkat felhasználva tovább validáljuk előzőleg bemutatott módszerünket dialógus adathalmazok szűrésére [Csáky et al., 2019]. A jelenlegi neurális háló alapú dialógus modellekből hiányzik a diverzitás, és unalmas válaszokat generálnak nyílt végű bemeneti mondatokra. Szerintünk ez annak köszönhető, hogy az adathalmazokban általában egy bemenetre sok elfogadható válasz létezik, és hasonlóan egy kimenetre sok potenciálisan jó bemenet létezik. A szűrési módszerünk ezt a problémát megpróbálja kezelni azzal, hogy eltávolítja a generikus mondatokat a tanítóadatból egy egyszerű entrópia-alapú módszerrel. Ebben a dolgozatban röviden bemutatjuk ezt a módszert és a hatékonyságát különböző dialógus adatokon beleértve a Gutenberg Dialógus Adathalmazt. Továbbá megtárgyaljuk a jelenlegi evaluációs metrikákkal felvetődő potenciális problémákat, és hogy ezek hogyan befolyásolják az eredményeink helyességét. Zárásul, tovább motiváljuk a transzfer tanulási hatások feltárását, és egy többnyelvű Gutenberg Dialógus Adathalmazt javaslunk.

[1] `http://www.gutenberg.org/`

Abstract

We create a new, large, high-quality dataset for neural dialog modeling, and explore its benefits over existing datasets. A dialog agent (chatbot) is a software that communicates with humans through natural language, by processing the users' utterances and outputting relevant and interesting responses. While in the past simple rule-based models were popular for building chatbots, currently deep learning models dominate the field of dialog modeling. The paradigm of training neural network based conversational agents on large dialog data raises the questions of how data quality affects these agents, and what evaluation methods can we use to effectively gauge the performance of trained models. This work tries to answer some of these questions by exploring relevant background and proposing a new dialog dataset.

First, we look at the current open-domain dialog datasets used in the literature, and we create a new, large, high-quality dataset. Dialogs are extracted from online books from Project Gutenberg[2]. We present a detailed data analysis, and we explain the reasoning behind choices of hyperparameters and pre-processing steps in order to build as high-quality of a dataset as possible. We also give an extensive error analysis both at the utterance and dialog level.

Second, we evaluate our dataset comparing it to other large datasets in the context of transfer learning. We argue that pre-training on our dataset results in better performance on smaller downstream datasets. Moreover, we use our dataset to further validate our previously published method for filtering dialog datasets [Csáky et al., 2019]. Current neural network-based conversational models lack diversity and generate boring responses to open-ended utterances. We believe the reasons for this are tied to the dataset, in that for one input there exist many potentially good outputs, and similarly, for one output there exist many potentially good inputs in the data. Our filtering method tackles this problem by removing generic utterances from training data using a simple entropy-based approach. In this work we briefly present this method and show its effectiveness on different dialog data including the Gutenberg Dialog Dataset. We also discuss potential problems with current evaluation metrics, and how these influence the validity of our results. We conclude by motivating further exploration of the observed transfer learning effects, and proposing a multi-language Gutenberg Dialog Dataset.

[2] http://www.gutenberg.org/

Contents

1 Introduction

A dialog agent (chatbot) is a software that communicates with humans through natural language, by processing the users' utterances and outputting relevant and interesting responses. While there are many different ways of approaching this problem, we specifically focus on open-domain neural network based single turn dialog modeling. Open-domain, in contrast with goal-oriented chatbots, means that the agent should be able to have a conversation about virtually anything, similarly to humans. We focus on neural network based models, since in the last few years these have been achieving state-of-the-art results in open-domain dialog modeling and more generally in many natural language processing (NLP) tasks [Vaswani et al., 2017, Devlin et al., 2018]. Lastly, single-turn means that the neural network takes as input the previous turn from a dialog and its task is to output a relevant response. This approach is also called sequences-to-sequence transduction. For a more in-depth review of the different types of chatbots we refer readers to our prior work [Csaky, 2019].

The paradigm of training neural network based conversational agents has three main hurdles: the training data, the model, and the evaluation method. In this work we mainly focus on the data hurdle and secondarily on evaluation methods. In Section 2 we discuss various dialog datasets used in the literature, the neural network model used in subsequent experiments, and our evaluation methods.

Training on large and good-quality datasets is essential for having good results in any NLP task. Unfortunately, there is a lack of such datasets for open-domain dialog modeling, thus our main contribution is the creation of a new, large, high-quality dataset. In Section 3 we present our method for collecting this dataset, the various preprocessing steps we used to make it high-quality, and also an extensive error analysis.

In Section 4 we evaluate our dataset by comparing it to other large datasets in the context of transfer learning. We argue that pre-training on our dataset results in better performance on smaller downstream tasks. Finally, in section 5 we present our prior work on data filtering [Csáky et al., 2019] and apply it to our current dataset. We conclude (Section 6) by motivating further exploration of the observed transfer learning effects, and proposing a multi-language dataset.

2 Background

2.1 Datasets

Open-domain dialog datasets vary in size, quality, and type, among other properties. We compare current datasets used in the literature in Table 1. Even though our estimate of quality is very subjective we can still observe a general trend of bigger datasets being of poorer quality. This is somewhat unsurprising since usually small datasets are carefully built using Amazon Mechanical Turk, while bigger datasets are scraped from sources likes twitter, reddit, or movie subtitles. The lack of a high-quality, large dataset motivates this work.

Dataset	**Size**	**Type**	**Quality**
DailyDialog [Li et al., 2017c]	90	collected from english textbooks	high
Wizard-of-Wikipedia [Dinan et al., 2019]	100	mechanical turk	high
Document-grounded [Zhou et al., 2018b]	100	mechanical turk	high
Persona-Chat [Zhang et al., 2018b]	150	mechanical turk	high
Self-dialogue [Fainberg et al., 2018]	150	mechanical turk	high
Cornell Corpus [Danescu-Niculescu-Mizil and Lee, 2011]	300	movie scripts	medium
Self-feeding chatbot [Hancock et al., 2019]	500	human-bot conversations	medium
Twitter corpus[3]	5000	twitter post-reply pairs	low
Opensubtitles [Henderson et al., 2019]	300000	movie subtitles	low
Reddit [Henderson et al., 2019]	650000	reddit threads	low

Table 1: Comparison of various open-domain dialog datasets. *Size* is the rough number of utterances in thousands (not exact). *Type* describes how the dataset was collected and *Quality* is a rough, subjective estimate of its quality.

2.2 The Transformer

When applying neural networks to NLP tasks, each word (symbol) has to be transformed into a numerical representation. This is done through word embeddings, which represent each word as a fixed size vector of real numbers. Word embeddings are useful because instead of handling words as huge vectors of the size of the vocabulary, they can be represented in much lower dimensions. Each vector representing a word can be regarded as a set of parameters and these parameters can be jointly learned with the neural network's parameters.

We can distinguish two main types of neural network model setups used for dialog modeling. Retrieval models simply score responses found in training data and choose the most likely response. In contrast generative models like the Transformer [Vaswani et al., 2017] synthesize the response one word at a time by outputting probabilities over the whole vocabulary. In this work we are only concerned with generating replies. For a more in depth discussion of different models we refer readers to [Csaky, 2019].

Generative encoder-decoder networks are trained on input-target pairs, by processing the input, generating an output, and comparing the generated output with the target. The log probability of a

correct target sequence T given the source sequence S is maximized:

$$\frac{1}{\boldsymbol{S}} \sum_{T,S \in \boldsymbol{S}} \log(p(T|S)) \tag{1}$$

Cross-entropy loss is usually used between the generated output and the target in order to make the network more likely to output the target.

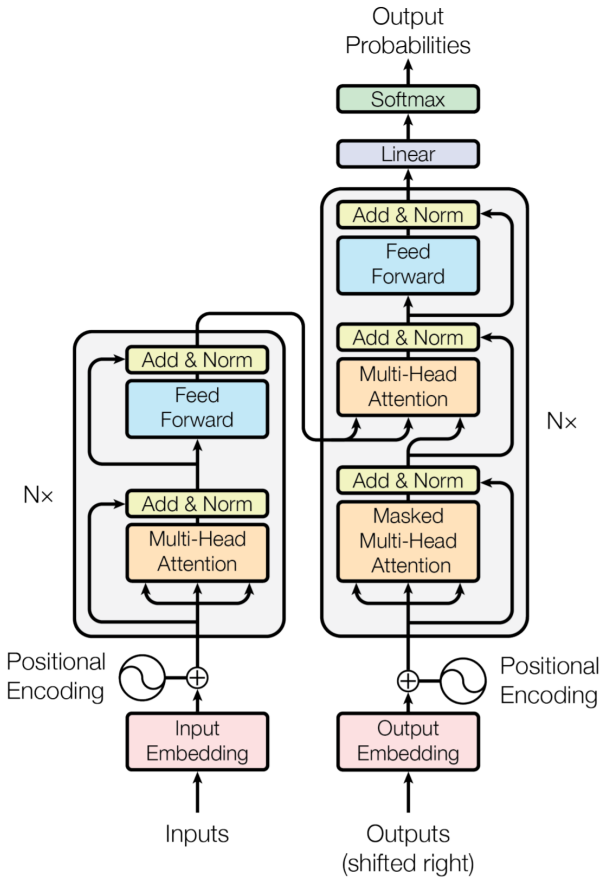

Figure 1: The architecture of the Transformer model [Vaswani et al., 2017]. The encoder network can be seen on the left and the decoder network on the right.

The Transformer model is part of the more general encoder-decoder model family. While previously RNN-based models have been used both in conversational modeling and other sequence-to-sequence NLP tasks [Sutskever et al., 2014], ever since the publication of the Transformer, such models have been shown to perform better in a variety of tasks [Dinan et al., 2019, Mazare et al., 2018, Devlin et al., 2018]. RNN-based models are a special type of neural network which use recurrence to deal with the sequential nature of data. In contrast Transformer-type models use only feed-forward and attention mechanisms [Bahdanau et al., 2015], which make them faster and more parallelizable. The architecture of the Transformer model can be seen in Figure 1, however this is not further detailed since it's not the focus of this work. We refer readers to the original paper [Vaswani et al., 2017] and our prior work [Csaky, 2019] for more details.

2.3 Evaluation Methods

This section describes the automatic metrics that will be used to assess the quality of generated responses throughout the paper, and also some issues that these metrics have. The description largely follows our prior paper [Csáky et al., 2019].

Currently, there is no well-defined automatic evaluation method [Liu et al., 2016], and while some metrics that correlate more with human judgment have been proposed recently [Li et al., 2017b, Lowe et al., 2017, Tao et al., 2018], they are harder to measure than simpler automatic metrics like perplexity or BLEU [Papineni et al., 2002]. Furthermore, even human evaluation has its downsides, like high variance, high cost, and difficulty of replicating experimental setups [Zhang et al., 2018b, Tao et al., 2018]. Some works resort to human evaluations [Krause et al., 2017, Fang et al., 2018], others use automatic metrics only [Olabiyi et al., 2018, Xing and Fernández, 2018, Kandasamy et al., 2017, Shalyminov et al., 2018, Xu et al., 2018b], and some use both [Shen et al., 2018a, Xu et al., 2018a, Baheti et al., 2018, Ram et al., 2018]. While extensive human evaluation of the methods presented here is left for future work, we do conduct an especially thorough automatic evaluation.

In order to get as complete a picture as possible, we use 17 metrics that have been applied to dialog models over the past years, briefly described below. These metrics assess different aspects of response quality, thus models should be compared on the whole set of metrics.

Response length. Widely used as a simple engagement indicator [Serban et al., 2017b, Tandon et al., 2017, Baheti et al., 2018].

Word and utterance entropy. The per-word entropy $H_w = -\frac{1}{|U|}\sum_{w \in U} \log_2 p(w)$ of responses is measured to determine their non-genericness [Serban et al., 2017b]. Probabilities are calculated based on frequencies observed in the training data. We introduce the bigram version of this metric, to measure diversity at the bigram level as well. Utterance entropy is the product of H_w and $|U|$, also reported at the bigram level.

KL divergence. We use the KL divergence between model and ground truth (GT) response sets to measure how well a model can approximate the GT distribution of words. Specifically, we define distributions p_{gt} and p_m based on each set of responses and calculate the KL divergence $D_{kl} = \frac{1}{|U_{gt}|}\sum_{w \in U_{gt}} \log_2 \frac{p_{gt}(w)}{p_m(w)}$ for each GT response. The bigram version of this metric is also reported.

Embedding metrics. Embedding *average*, *extrema*, and *greedy* are widely used metrics [Liu et al., 2016, Serban et al., 2017b, Zhang et al., 2018c]. *average* measures the cosine similarity between the averages of word vectors of response and target utterances. *extrema* constructs a representation by taking the greatest absolute value for each dimension among the word vectors in the response and target utterances and measures the cosine similarity between them. Finally, *greedy* matches each response token to a target token (and vice versa) based on the cosine similarity between their embeddings and averages the total score across all words.

Coherence. We measure the cosine similarity between pairs of input and response [Xu et al., 2018b]. Although a coherence value of 1 would indicate that input and response are the same, generally a higher value seems better as model responses tend to have lower coherence than targets.

Distinct metrics. *Distinct-1* and *distinct-2* are widely used in the literature [Li et al., 2016a, Shen et al., 2018a, Xu et al., 2018b], measuring the ratio of unique unigrams/bigrams to the total number of unigrams/bigrams in a set of responses. However, they are very sensitive to the test data size, since increasing the number of examples in itself lowers their value. While the number of total words increases linearly, the number of unique words is limited by the vocabulary, and we found that the ratio decreases even in human data (Figure 2). It is therefore important to only compare *distinct* metrics computed on the same test data.

Bleu. Measuring n-gram overlap between response and target is widely used in the machine learning and dialog literature [Shen et al., 2018a, Xu et al., 2018b]. We report BLEU-1, BLUE-2, BLEU-3, and BLEU-4 computed with the 4th smoothing algorithm described in [Chen and Cherry, 2014].

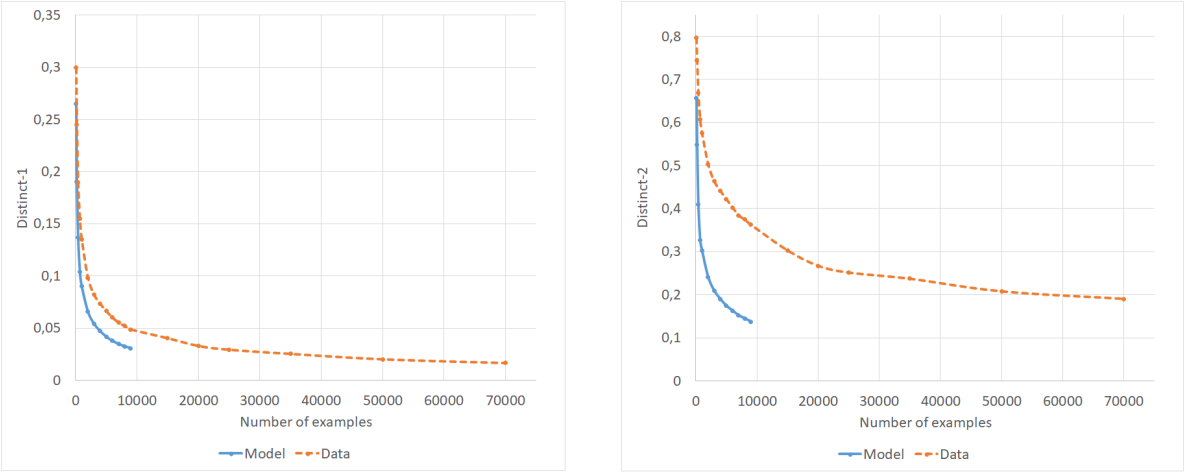

Figure 2: Distinct-1 (left) and Distinct-2 (right) metric with respect to number of test examples, on the DailyDialog dataset. Model responses were evaluated on 9000 examples only, since the rest were training examples.

3 The Gutenberg Dialog Dataset

The lack of large, high-quality dialog datasets motivates the creation of one, since datasets are one of the major hurdles in advancing the field of dialog modeling. Our corpus, the Gutenberg Dialog Dataset[4] is constructed by extracting dialogs from online books. Project Gutenberg[5] is an online library containing 60.000 books, most of which are in the public domain. We used the gutenberg python package[6] to download these books, and query their metadata like language, author, license.

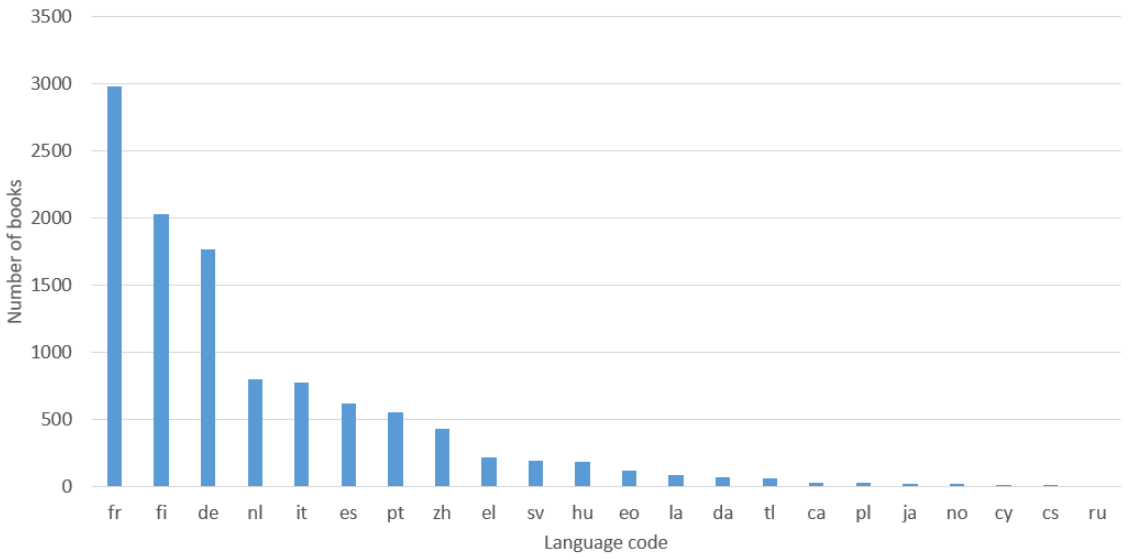

Figure 3: Number of books per language. English has 50.000 books, purposefully left out from the figure.

In Figure 3 we can see the number of books in various languages. English has roughly 50.000 books, while the remaining 10.000 are split among many languages. The top two authors are Shakespeare and Mark Twain, with 323 and 224 books, respectively.

3.1 Pipeline

Project Gutenberg is useful for building a dialog dataset, because books generally contain interesting and authentic dialog, and most books are in the public domain. As discussed in Section 2 most large dialog datasets today are not that high-quality, because they use resources which don't contain natural dialog. One of the largest datasets, Opensubtitles, uses movie subtitles as the dialog dataset. However, subtitles can be of descriptive nature as well, and they don't provide a clear delimitation between turns of speakers, thus this dataset is low-quality. Other large datasets include the Twitter and Reddit dialog datasets, in which we can be certain that nearly all text is of dialogic nature, and turns are clearly delimited by usernames. However, these are more post-reply style dialogs, which are far from natural dialog occuring between two people. Usually there is a post text and then replies are related to that post, and conversations can happen between multiple people

[4]Code and data will be released in the following months

[5]`https://www.gutenberg.org/`

[6]`https://github.com/ageitgey/Gutenberg`

at the same time. While conversations in books are generally natural and high-quality it is far from trivial to extract them without introducing some amount of noise. The main challenges are finding the correct delimitations of utterances and between separate conversations. What follows is a descprition of the pipeline of going from raw books to the final dialog dataset, while discussing in detail all processing steps and parameters. After downloading we remove books which are not in the public domain, and separate them by language using metadata information. This work is only concerned with building an English dataset, however a potential line of future work is to make a multilingual dataset.

3.1.1 Filter old books

The first preprocessing step is meant to filter out noise and also books in which the language is older or very different. In order to do this, we first calculate the distribution of words (vocabulary) across all books. Then we look at the vocabulary of each book individually and compare it with the overall vocabulary of all books. We weight the word counts by the total number of words in the respective books, so the two distributions are comparable. We then simply subtract the two distributions from each other, getting the difference between word counts. We sum this difference across word counts to a single number and then we divide it by the overall word count sum. If this fraction is above a certain threshold, meaning that there are big differences in word counts, we remove the book from our dataset. Because this is a very rough filtering, we don't perform it if a book is short (below 20.000 words).

We empirically set this fraction threshold to 0.45, by trying lower and lower values until we observed that books which shouldn't be removed also started to be removed. 0.45 is the lowest we could go with relatively few false positives.

3.1.2 Dialog extraction

This is the main step of our pipeline consisting of multiple sub-steps. To extract good dialogs from books there are three main challenges that need to be overcome. First we have to somehow identify what text is part of a dialog, and not part of the narrative or any other text. Second within this text we have to find separation between different dialogs. By different dialogs we mean for example if the dialogs are carried out by different persons, or at different times. Basically if the two dialogs don't have anything to do with each other then they are separate. So we have to segment our continuous conversational text into dialogs that are not related to each other. Finally, the segmented dialogs need to be further segmented into separate turns of utterances.

We found that in most English books there is a clear distinction whether a text is part of a dialog. Most often the text is placed between quotation marks, or underscores, which we collectively call delimiters. Thus the first challenge is cleared. Obviously quotation marks can be used for other purposes as well, but such uses are rare. A simple further heuristic that we employ is to check whether in a paragraph the first occurrence of an utterance starts with an upper-case word. Since quotation marks are usually used to highlight specific words and not sentences, this method can filter out such false use-cases. We also employ a simple filtering step here by looking at the total number of delimiters in a book relative to the number of words it contains. If this number is small that either means that the book doesn't contain any dialog and the delimiters are used for different purposes, or it contains very little dialog. Thus, we can filter a lot of books which don't contain

dialog at a small expense of cutting a few dialogs from our dataset. We empirically set this number to 150 delimiters per 10.000 words, by increasing it until we observed that books containing dialog were also removed. We also tried to balance the number of books that we remove, and looking at Figure 4 we can see that 150 corresponds to a clear cutoff point from which further increasing it would throw away a larger number of books. The vertical axis is the size of our final dataset (sample), and we can see that increasing the threshold removes more and more books. Since there are multiple delimiters that can be used for delimiting dialog we had to decide somehow automatically which delimiter a book uses. For this we simply counted how many times the various delimiters appear and choose the highest count.

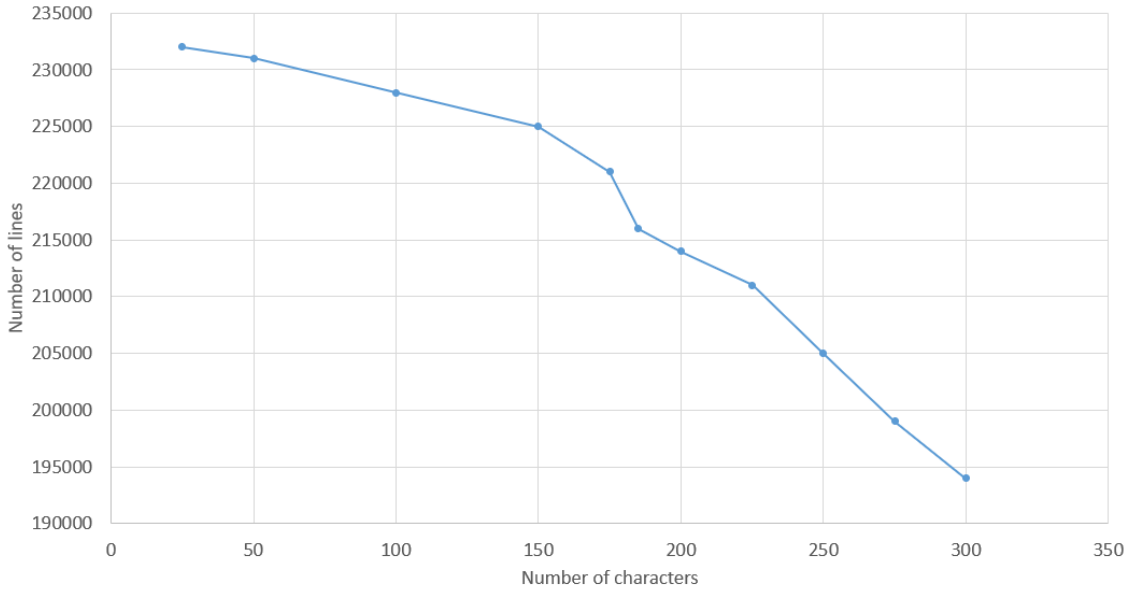

Figure 4: Dataset size with respect to the delimiter threshold (in characters per 10.000 words).

For the second challenge we employed a very simple heuristic. We simply count the number of characters between each two conversational text segments (found earlier with the delimiters), and if this number is above a certain threshold we consider the two segments as parts of different dialogs. The problem with this method is that it will always create bad splits no matter how we set this threshold, since the variability of text length between dialogs is very high. Thus we tuned this parameter by balancing out the number of cases in which the threshold is too low and the number of cases in which it is too high. We call this parameter the dialog gap and set it to 150 characters. In Figure 5 we can see the relationship between this parameter and the average length of the extracted dialogs. Since the relationship is mostly linear this analysis didn't provide any further information as to where this parameter could be optimal (in contrast with the delimiter parameter where there was a clear change in the graph).

For the final challenge we found that in most books separate utterances are in separate paragraphs. Thus a dialog can be segmented into utterances by taking consecutives paragraphs as consecutive turns. In a single paragraph there can be multiple delimited text sections which are all part of the same turn, and we simply join them. A dialog sample which highlights all the different findings that we mentioned can be seen in Figure 6. As can be seen consecutive utterances are in consecutive paragraphs, and sometimes within a single paragraph an utterance might

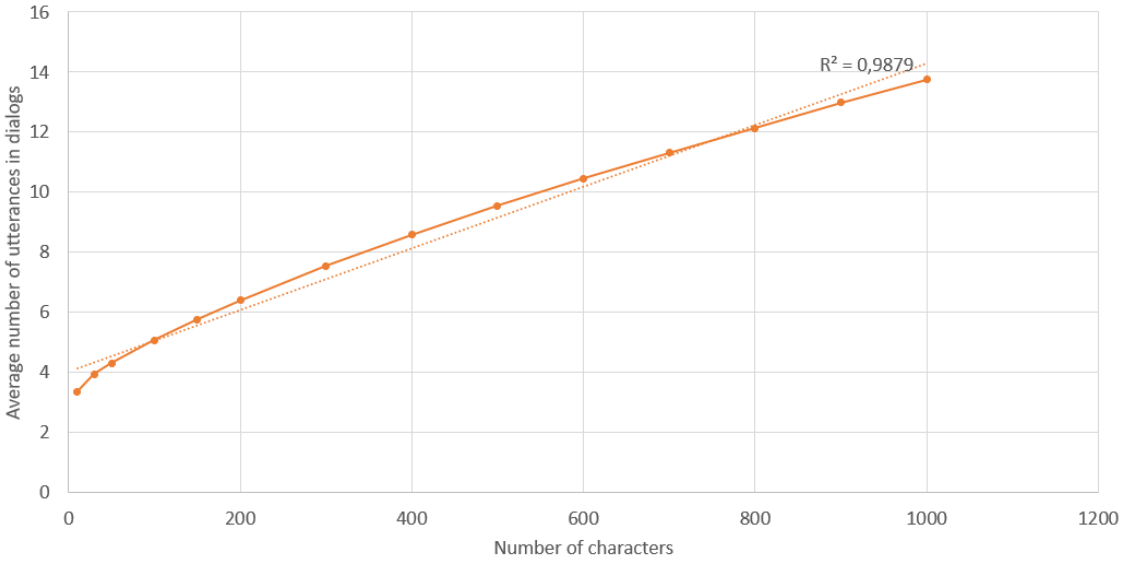

Figure 5: Average number of utterances with respect to dialog gap size in number of characters.

be broken up by non-dialog text which has to be removed. Unfortunately, dialogs do not always appear so cleanly and we will see the different errors that arise from our preprocessing decisions in Section 3.2.

```
11279   "He is a misanthrope!" said Basia.
11280
11281   "Baska," said Zagloba, "imagine to yourself that you had a daughter,
11282   and that you had to give her to some Tartar--"
11283
11284   "Azya is a prince."
11285
11286   "I do not deny that Tugai Bey comes of high blood. Ketling was a noble;
11287   still Krysia would not have married him if he had not been
11288   naturalized."
11289
11290   "Then try to obtain naturalization for Azya."
11291
11292   "Is that an easy thing? Though some one were to admit him to his
11293   escutcheon, the Diet would have to confirm the choice; and for that,
11294   time and protection are necessary."
11295
```

Figure 6: A dialog example showing the different findings.

A final step we do in order to remove some noise and make the dataset more robust is to throw out utterances longer than 100 words. When such an utterance is removed from the middle of a dialog, the dialog is simply cut up into two separate dialogs. This step ensures that the utterances that remain are truly conversational and not descriptive or some other artifacts. Also neural models are known to have trouble processing long sequences, so this makes the task easier [Dai et al., 2019].

3.1.3 Final filtering

Now that we have the extracted dialogs we perform one final filtering based on vocabulary again. We build the vocabulary of all extracted dialogs, and keep the top 100.000 words in a list. Then we examine each dialog individually and if more than 20% of the words in the dialog are not in the top 100.000 word list, we remove the dialog. This again removes a very small part of our dataset, but it helps removing edge cases and noise. After this we split our dialogs into train, validation, and test data, containing 90%, 5%, and 5% of our dataset, respectively.

3.2 Error Analysis

In order to judge the quality of our dialog dataset and the errors that the various preprocessing steps induced we perform an error analysis both at an utterance level and at a dialog level.

3.2.1 Utterance level

We look at 30 random consecutive utterance pairs sampled from our dialog dataset. Based on these we identified 3 error categories presented in Table 2. Out of the 30 pairs, only 8 contained any errors, from which half were *same speaker* errors, and the other half was further divided into *different dialog* and *not dialog* errors. No utterance pair contained more than 1 type of error.

Same speaker means that both utterances were from the same speaker which is obviously a big error. This error mostly occured because our assumption that consecutive paragraphs contained utterances from different speakers was violated.

Different dialog means that the two utterances in the pair should have been split into different dialogs either because different people were talking or because the second utterance in the pair was not related to the first one. Dialog splitting issues will be of more consequence in the dialog level error analysis (next section).

Not dialog means that one of the two utterances was not conversational text. This fortunately only occurred in 2 out of 30 examples, which means that we managed to filter out most of these false positive cases.

Error category	same speaker	different dialog	not dialog
Number of errors	4	2	2

Table 2: Number of errors in the various error categories. Total number of examined utterance pairs is 30.

3.2.2 Dialog level

We examined 30 randomly sampled dialogs from out dataset. We plot the error distribution for each error category in Figure 8. We can see that there were only 2 dialogs in which there were no errors, most dialogs contained 1 error, some contained 2 errors and very few contained 3 or 4 errors. The percentages of the various errors are also visualized in Figure 7. Now we discuss them in decreasing order of frequency.

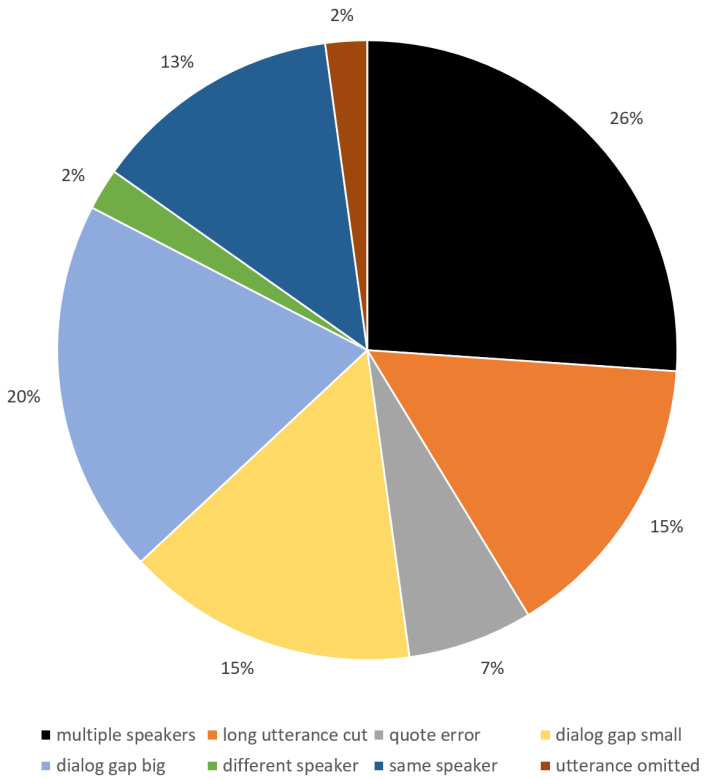

Figure 7: Distribution of the 8 error categories.

Multiple speakers error refers to a dialog containing more than 2 speakers. This was the most frequent error, as books often contain dialog between multiple speakers. However, this is not a really big problem since it's still a coherent dialog.

Dialog gap big/small These two errors are paired, they just mean that the dialog gap was either too big, so multiple dialogs ended up together, or too small, in which case a single dialog got cut up into 2 dialogs. As we discussed before there is no good way to set this dialog gap, but the fact that these two opposing errors occurred with similar frequency means that we at least set it to a good balance. If we look at Figure 8 we can see that there were some cases where even within a single dialog both errors occurred, which points to the high variability of the dialog gap. Future work should focus on better ways to segment conversational text into dialogs.

Long utterance cut means that because of our long utterance filtering (more than 100 words) the dialog was cut in two, however this is not technically an error since we deliberately choose to cut the dialog.

Same speaker means that the same speaker uttered at least 2 consecutive utterances in the same dialog, which is obviously a big error, but as we mentioned in the utterance level error analysis there is no good way to find these errors automatically.

Quote error similarly to utterance level errors means that the quote was used for other purposes than dialog delimitation, which fortunately didn't occur too much.

The last two error categories only occurred once throughout 30 examples so they have little impact on the overall quality. All in all, *same speaker* and *quote errors* are the biggest problems (20% of all errors) as these break up the flow of dialog and the continuity of utterance turns, while

the other errors are generally more minor and wouldn't cause big problems for neural networks.

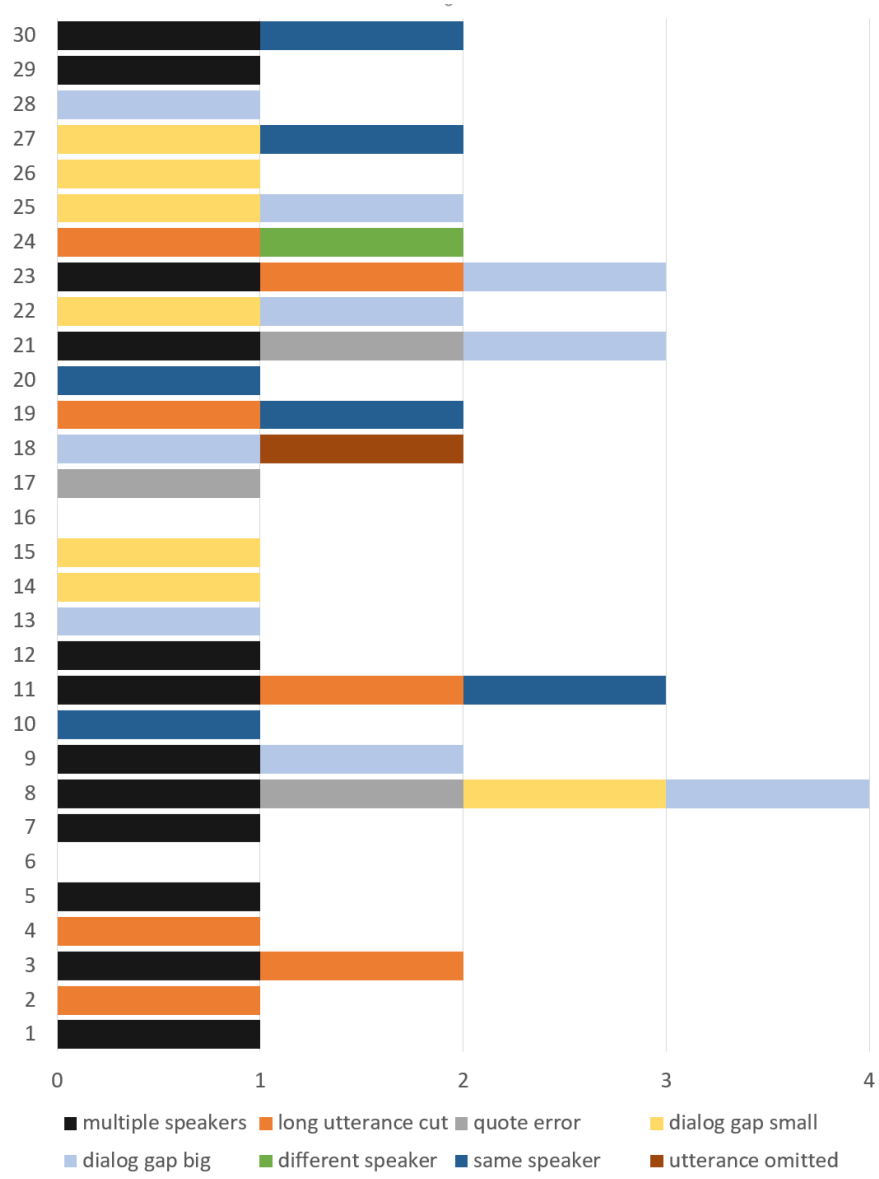

Figure 8: Distribution of the various error categories over the 30 examined dialogs.

4 Transfer Learning Experiments

To evaluate our new dataset we consider how it can be used for transfer learning to smaller downstream tasks. We train a Transformer model on the Gutenberg dialog dataset until validation loss minimum is reached, and we train another Transformer on a subset of the Opensubtitles dataset (which has the same size as our Gutenberg dataset) for the same number of epochs. This ensures that the comparison between these two datasets is fair. Then we look at how the models trained on these two datasets perform on the test set of the DailyDialog dataset. We evaluate zero-shot performance, when the models are not further trained on DailyDialog, and we also compare models after finetuning on DailyDialog until validation loss minimum is reached.

In all training experiments in this work we used the official implementation[7] of Transformer. Word embeddings of size 512 were randomly initialized and we used the Adam optimizer [Kingma and Ba, 2014]. We experimented with various beam sizes [Graves, 2012], but greedy decoding performed better according to all metrics, also observed previously [Asghar et al., 2017, Shao et al., 2017, Tandon et al., 2017]. The hyperparameters of the transformer model can be found in Table 3. In all tables the 17 metrics from left to right are: response length, unigram and bigram entropy, unigram and bigram utterance entropy, unigram and bigram KL divergence, embedding *average*, *extrema* and *greedy*, coherence, *distinct-1* and *distinct-2*, and finally, BLEU-1, BLEU-2, BLEU-3 and BLEU-4 (as detailed in Section 2).

Name	Value
Hidden size	512
Number of hidden layers	6
Label smoothing	0.1
Filter size	2048
Number of attention heads	8
Layer dropout	0.2
Relu dropout	0.1
Attention dropout	0.1
Learning rate	0.2
Learning rate warmup steps	8000

Table 3: Transformer hyperparameters.

4.1 Zero-shot performance

In Table 4 we present the results of the Transformer trained on Gutenberg and on Opensubtitles. We evaluate these models on the test set of DailyDialog in a zero-shot context, meaning no further training was performed on DailyDialog. We can see that the model trained on Gutenberg performs better across nearly all metrics than the one trained on Opensubtitles. Sadly the model doesn't always get better results than randomly selected responses from the training set of DailyDialog, but this can be partially attributed to the nature of the metrics. Also DailyDialog is notorious for

[7] https://github.com/tensorflow/tensor2tensor

containing overlapping examples between train and test splits, making this comparison somewhat unfair.

	$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
GUT	**24.6**	**7.21**	**11.7**	**160**	**237**	2.07	3.63	**.649**	.455	**.615**	**.682**	**.006**	**.03**	**.082**	**.098**	**.105**	**.102**
OPEN	4.81	6.65	10.6	32.2	41.1	2.05	**3.56**	.607	**.466**	.611	.605	.0009	.0017	.075	.068	.063	.056
RAND	13.6	8.41	14.2	118	180	.051	.173	.666	.387	.603	.666	.067	.403	.087	.119	.128	.124

Table 4: Transformer trained on the Gutenberg Dataset (GUT) and Opensubtitles (OPEN) evaluated on the test of DailyDialog. RAND refers to randomly selected responses from the training data of DailyDialog. Best results are highlighted with bold (if significantly better).

4.2 Finetuned performance

In Table 5 we can see the results of the Transformer trained on Gutenberg and Opensubtitles and then further finetuned on DailyDialog until the validation loss reached a minimum. It is clear that the model first trained on Gutenberg achieves much better results than the model trained on Opensubtitles which struggles to finetune on DailyDialog. Furthermore, the model trained on Gutenberg is even better than the model trained solely on DailyDialog according to some metrics. This means that pre-training on Gutenberg offers some benefits over just training on DailyDialog. These benefits will be further explored in future work.

	$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
GUT	8.24	**7.02**	**11.7**	**59.5**	**84.4**	**.593**	**1.09**	**.666**	**.473**	**.655**	**.688**	**.025**	**.127**	**.134**	**.126**	**.121**	.110
OPEN	**8.78**	6.69	10.2	58.8	79.6	2.93	4.16	.646	.466	.626	.643	.002	.0049	.107	.117	.118	.111
DD	9.85	7.12	11.5	71.8	95.1	.895	1.61	.663	.461	.642	.667	.0133	.063	.127	.129	.127	.118

Table 5: Transformer trained on the Gutenberg Dataset (GUT) and Opensubtitles (OPEN) and further finetuned on DailyDialog evaluated on the test of DailyDialog. DD refers to a model trained only on DailyDialog. Best results are highlighted with bold (if significantly better).

5 Data Filtering Experiments

Most of this section was previously published [Csáky et al., 2019]. We introduce our prior work briefly (with some additions which were not in the original paper) since it's also related to creating better quality datasets. We also try our method on the new Gutenberg Dialog Dataset.

Current open-domain neural conversational models (NCM) are trained on pairs of source and target utterances in an effort to maximize the likelihood of each target given the source [Vinyals and Le, 2015]. However, real-world conversations are much more complex, and a plethora of suitable targets (responses) can be adequate for a given input. We propose a data filtering approach where the "most open-ended" inputs - determined by calculating the entropy of the distribution over target utterances - are excluded from the training set. We show that dialog models can be improved using this simple unsupervised method which can be applied to any conversational dataset. We conduct several experiments to uncover how some of the current open-domain dialog evaluation methods behave with respect to overfitting and random data. Our software for filtering dialog data and automatic evaluation using 17 metrics is released on GitHub under an MIT license[8][9].

Most open-domain NCMs are based on neural network architectures developed for machine translation (MT, [Sutskever et al., 2014, Cho et al., 2014, Vaswani et al., 2017]). Conversational data differs from MT data in that targets to the same source 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?*. Dialog datasets also contain generic responses, e.g. *yes*, *no* and *i don't know*, that appear in a large and diverse set of contexts [Mou et al., 2016, Wu et al., 2018]. 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. `seq2seq` architectures are not suited to deal with the ambiguous nature of dialogs since they are inherently deterministic, meaning that once trained they cannot output different sequences to the same input. Consequently they tend to produce boring and generic responses [Li et al., 2016a, Wei et al., 2017, Shao et al., 2017, Zhang et al., 2018a, Wu et al., 2018].

Previous approaches to the *one-to-many*, *many-to-one* problem can be grouped into three categories. One approach involves feeding extra information to the dialog model such as dialog history [Serban et al., 2016, Xing et al., 2018], categorical information like persona [Li et al., 2016b, Joshi et al., 2017, Zhang et al., 2018b], mood/emotion [Zhou et al., 2018a, Li et al., 2017c], and topic [Xing et al., 2017, Liu et al., 2017, Baheti et al., 2018], or through knowledge-bases [Dinan et al., 2019, Ghazvininejad et al., 2018, Zhu et al., 2017, Moghe et al., 2018]. A downside to these approaches is that they require annotated datasets which are not always available, or might be smaller in size. Augmenting the model itself, with e.g. latent variable sampling [Serban et al., 2017b, Zhao et al., 2017, Zhao et al., 2018, Gu et al., 2019, Park et al., 2018, Shen et al., 2018b, Gao et al., 2019], or improving the decoding process [Shao et al., 2017, Kulikov et al., 2018, Mo et al., 2017, Wang et al., 2018] is also a popular approach. Sampling provides a way to generate more diverse responses, however such models are more likely to output ungrammatical or irrelevant responses. Finally, directly modifying the loss function [Li et al., 2016a], or training by reinforcement [Li et al., 2016d, Serban et al., 2017a, Li et al., 2016c, Lipton et al., 2018, Lewis

[8] `https://github.com/ricsinaruto/Seq2seqChatbots`

[9] `https://github.com/ricsinaruto/dialog-eval`

et al., 2017] or adversarial learning [Li et al., 2017b, Ludwig, 2017, Olabiyi et al., 2018, Zhang et al., 2018c] has also been proposed, but this is still an open research problem, as it is far from trivial to construct objective functions that capture conversational goals better than cross-entropy loss.

Improving dataset quality through filtering is frequently used in the machine learning literature [Sedoc et al., 2018, Ghazvininejad et al., 2018, Wojciechowski and Zakrzewicz, 2002] and data distillation methods in general are used both in machine translation and dialog systems [Axelrod et al., 2011, Li et al., 2017a]. [Xu et al., 2018b] introduced coherence for measuring the similarity between contexts and responses, and then filtered out pairs with low coherence. This improves datasets from a different aspect and could be combined with our present approach. However, natural conversations allow many adequate responses that are not similar to the context, thus it is not intuitively clear why filtering these should improve dialog models. Our experiments also further support that cross-entropy is not an adequate loss function (shown qualitatively by [Csaky, 2019] and [Tandon et al., 2017]), by showing that many automatic metrics continue to improve after the validation loss reaches its minimum and starts increasing. However, we found that the metrics steadily improve even after we can be certain that the model overfitted (not just according to the loss function). Further investigation into this issue is presented in Section 5.3.

5.1 Filtering Method

We approach the *one-to-many*, *many-to-one* problem from a relatively new perspective: instead of adding more complexity to NCMs, we reduce the complexity of the dataset by filtering out a fraction of utterance pairs that we assume are primarily responsible for generic/uninteresting responses. Of the 72 000 unique source utterances in the DailyDialog dataset (see Section 5.2 for details), 60 000 occur with a single target only. 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 (*one-to-many*), models are forced to learn some "average" of observed responses [Wu et al., 2018].

The entropy of response distribution of an utterance s is a natural measure of the amount of "confusion" introduced by s. For example, the context *What did you do today?* has high entropy, since it is paired with many different responses in the data, but *What color is the sky?* has low entropy since it's observed with few responses. The *many-to-one* scenario can be similarly formulated, where a diverse set of source utterances are observed with the same target (e.g. *I don't know* has high entropy). While this may be a less prominent issue in training NCMs, we shall still experiment with excluding such generic targets, as dialog models tend to generate them frequently.

We refer with IDENTITY to the following entropy computation method. For each source utterance s in the dataset we calculate the entropy of the conditional distribution $T|S = s$, i.e. given a dataset D of source-target pairs, we define the *target entropy* of s as

$$H_{\text{tgt}}(s, D) = -\sum_{(s,t_i)\in D} p(t_i|s)\log_2 p(t_i|s) \tag{2}$$

Similarly, *source entropy* of a target utterance is

$$H_{\text{src}}(t, D) = -\sum_{(s_i,t)\in D} p(s_i|t)\log_2 p(s_i|t) \tag{3}$$

The probabilities are based on the observed relative frequency of utterance pairs in the data.

Dataset	Threshold	TARGET	BOTH
DailyDialog	1	6.07%	-
Cornell	4	7.39%	14.1%
Twitter	0.5	1.82%	9.96%
Gutenberg	4	5.87%	-

Table 6: Entropy threshold and amount of data filtered for all datasets in 2 filtering scenarios.

Entropy values obtained with this method were used to filter dialog data in three ways. The SOURCE approach filters utterance pairs in which the source utterance has high entropy, TARGET filters those with a high entropy target, and finally the BOTH strategy filters all utterance pairs that are filtered by either SOURCE or TARGET. The amount of data we ended up filtering for the different filtering ways and various datasets can be seen in Table 6. In preliminary experiments we observed that SOURCE filtering performs much worse than the other two, thus it's removed from further experiments.

5.2 DailyDialog Dataset

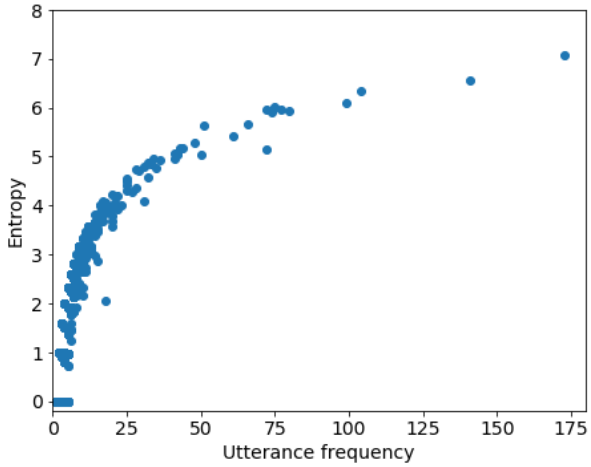

Figure 9: Entropy of source utterances with respect to utterance frequency.

With 90 000 utterances in 13 000 dialogs, DailyDialog [Li et al., 2017c], is comparable in size with the Cornell Movie-Dialogs Corpus [Danescu-Niculescu-Mizil and Lee, 2011], but contains real-world conversations. Using the IDENTITY approach, about 87% of utterances have 0 entropy (i.e. they do not appear with more than one target), 5% have an entropy of 1 (e.g. they appear twice, with different targets), remaining values rise sharply to 7. This distribution is similar for source and target utterances. In Table 7 we can see the 20 highest entropy utterances, which are exactly what we wanted to find, open-ended and without much context.

Entropy is clearly proportional to utterance frequency (Figure 9), but has a wide range of values among utterances of equal frequency. For example, utterances with a frequency of 3 can have

Utterance	Frequency	Entropy
yes .	173	7.06
thank you .	141	6.57
why ?	104	6.33
here you are .	99	6.10
ok .	75	6.00
what do you mean ?	77	5.97
may i help you ?	72	5.96
can i help you ?	80	5.93
really ?	74	5.91
sure .	66	5.66
what can i do for you ?	51	5.63
why not ?	61	5.42
what ?	48	5.27
what happened ?	44	5.18
anything else ?	43	5.17
thank you very much .	72	5.14
what is it ?	41	5.06
i see .	42	5.05
no .	42	5.04
thanks .	50	5.03

Table 7: Top 20 source utterances (from DailyDialog) sorted by entropy.

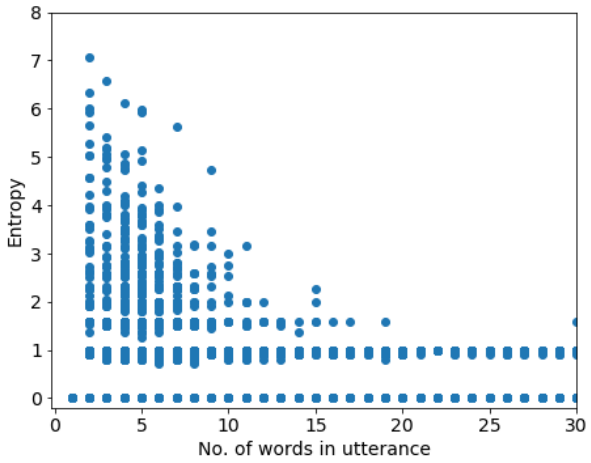

Figure 10: Entropy of source utterances with respect to utterance length.

entropies ranging from 0 to $\log_2 3 \approx 1.58$, the latter of which would be over our filtering threshold of 1. Since high-entropy utterances are relatively short, we also examined the relationship between entropy and utterance length (Figure 10). Given the relationship between frequency and entropy, it comes as no surprise that longer utterances have lower entropy.

5.3 Evaluation Issues

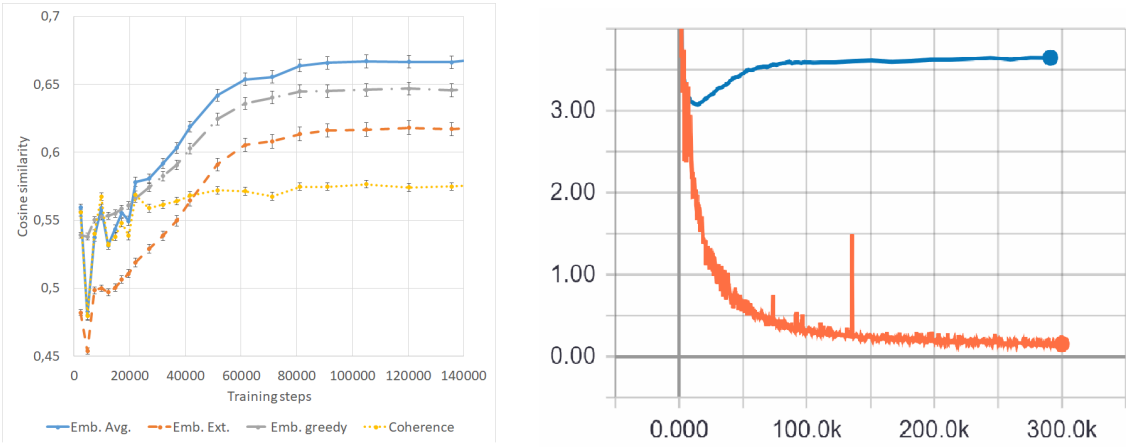

Figure 11: Embedding metrics and coherence (on validation data) as a function of the training evolution of `transformer` on unfiltered DailyDialog data on the left. The evolution of the validation loss (blue) and train loss (orange) on the right for the same training.

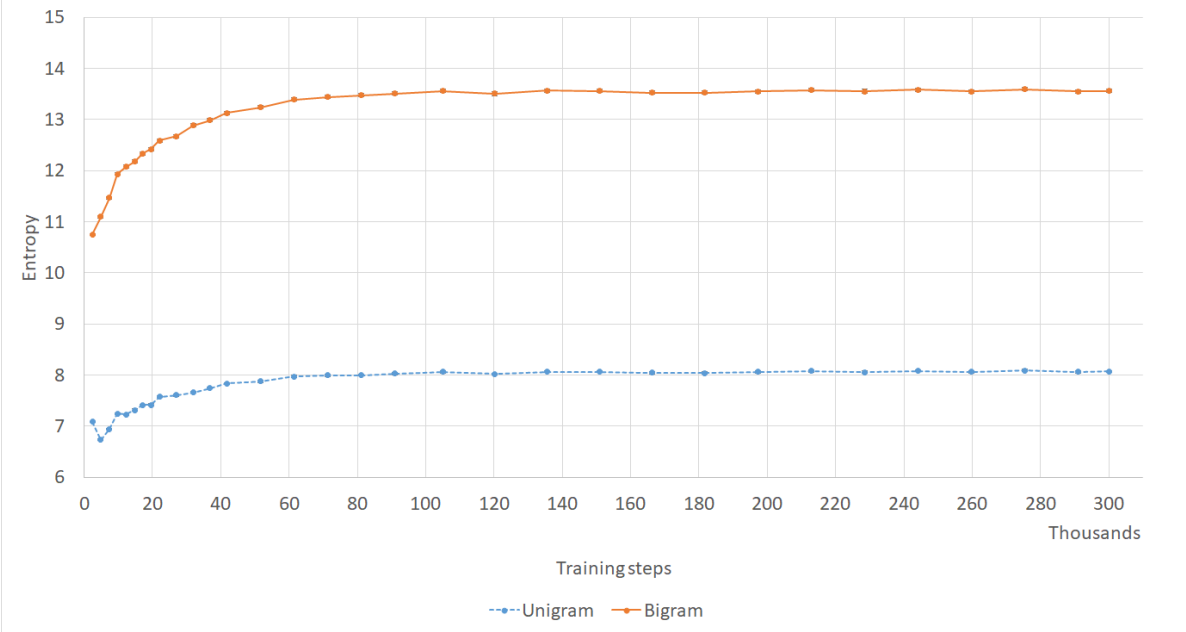

Figure 12: Word entropy of responses (computed on the validation set) with respect to the number of training steps of the `transformer` trained on unfiltered data (DailyDialog).

Normally metrics are computed at the validation loss minimum of a model, however in the case of chatbot models loss may not be a good indicator of response quality (Section 2), thus we also looked at how our metrics progress during training on the DailyDialog dataset. Figure 11 shows how coherence and the 3 embedding metrics saturate after about 80-100k steps (left graph), and

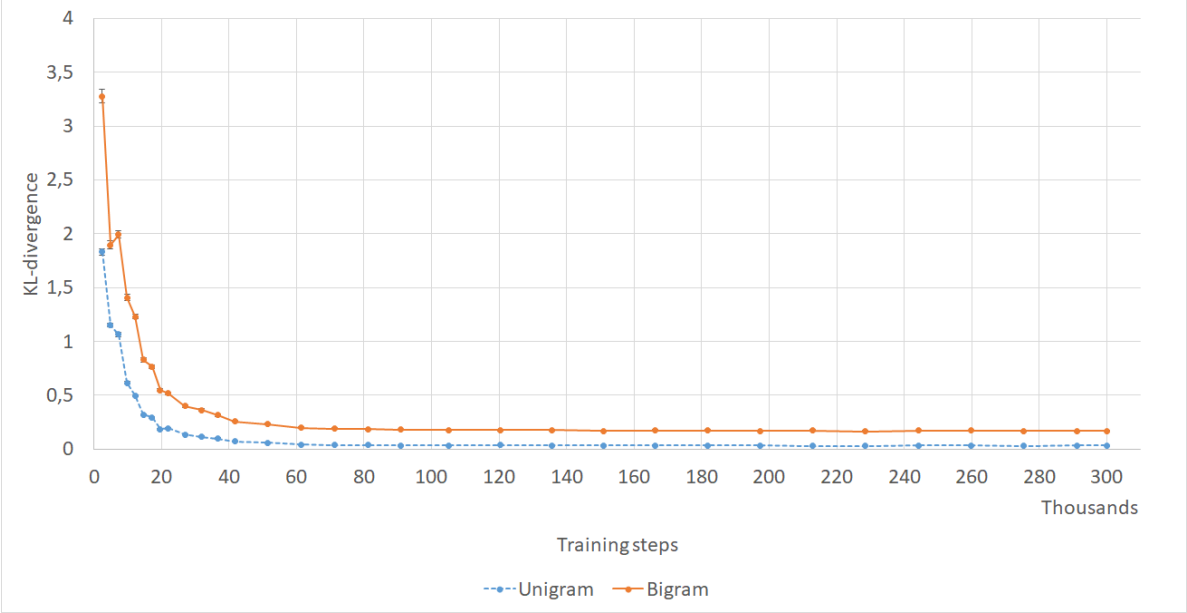

Figure 13: KL divergence of responses (computed on the validation set) with respect to the number of training steps of the `transformer` trained on unfiltered data (DailyDialog).

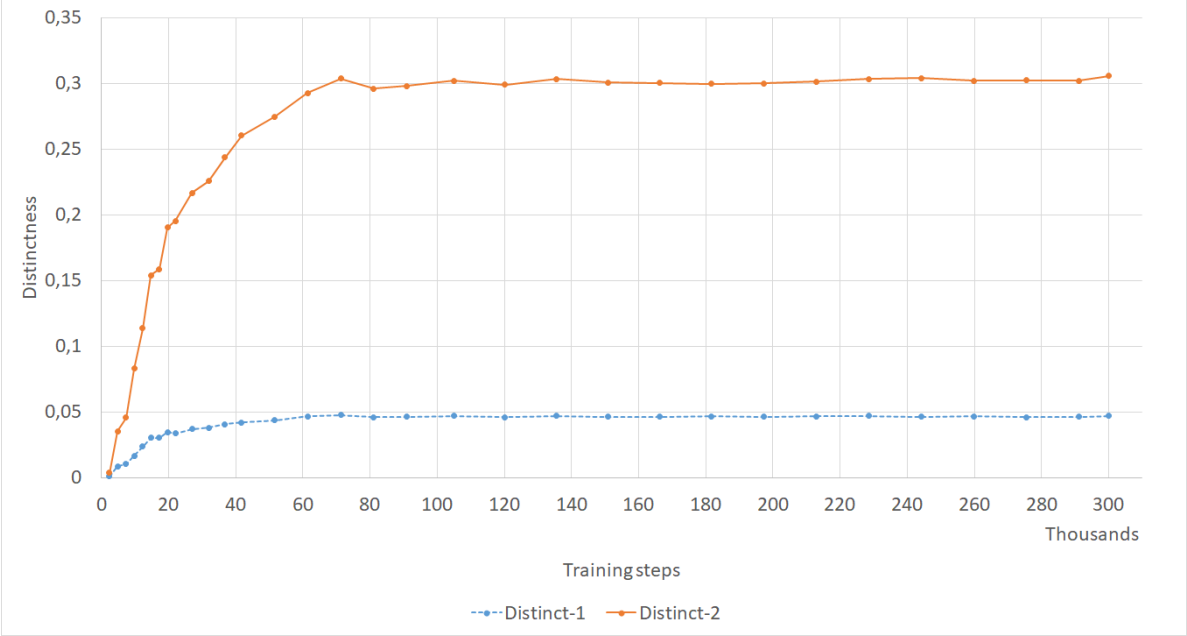

Figure 14: Distinct-1 and distinct-2 metrics (computed on the validation set) with respect to the number of training steps of the `transformer` trained on unfiltered data (DailyDialog).

never decrease (we ran the training for 300k steps, roughly 640 epochs). Most metrics show a similar trend of increasing until 100k steps, and then stagnating (see Figure 12, 13, and 14).

In contrast, validation loss for the same training reaches its minimum after about 10-20k steps (Figure 11 right graph). This again suggests the inadequacy of the loss function, but it also ques-

tions the validity of these metrics, as they seem to favor a model that overfitted the training data, which we can assume after 640 epochs. Most interesting are embedding metrics and BLEU scores (Section 5.4), since they show that even after overfitting responses do not get farther from targets. This is in line with other findings reporting that qualitatively responses are better after overfitting [Csaky, 2019, Tandon et al., 2017], however occasionally they also tend to be too specific and irrelevant.

Since our original findings we found out that most of this issue is due to the fact that the DailyDialog official train and test splits overlap significantly. 20% of the examples in the test set appear in the training set as well. Because of this if the model overfits it will perform extremely well on that 20% which has a bigger impact than performing worse on the other examples. Thus, we constructed the DailyDialog Curated dataset which has the same validation and test splits, however we remove examples in the training data which appear in the validation or test sets. Thus, the test set is untouched, but it provides a more fair comparison, since models are now trained on the curated training set.

5.4 DailyDialog Results

	$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF	8.6	7.30	12.2	63.6	93	.330	.85	.540	.497	.552	.538	**.0290**	.149	.142	.135	.130	.119
BOTH	9.8	7.44	12.3	71.9	105	.315	.77	.559	**.506**	*.555*	.572	.0247	.138	.157	.151	.147	.136
TARGET	10.9	**7.67**	**12.7**	**83.2**	**121**	**.286**	**.72**	**.570**	**.507**	.554	**.584**	.0266	**.150**	**.161**	**.159**	**.156**	**.146**
TRF-O	11.5	7.98	**13.4**	95	142	.0360	.182	.655	**.607**	**.640**	.567	**.0465**	**.297**	**.333**	.333	.328	.315
BOTH	**13.1**	**8.08**	**13.6**	**107**	**162**	.0473	.210	**.668**	**.608**	**.638**	**.598**	.0410	.275	**.334**	**.340**	**.339**	**.328**
TARGET	12.2	8.04	**13.6**	100	150	**.0335**	**.181**	**.665**	**.610**	**.640**	.589	.0438	.289	**.338**	**.341**	**.339**	**.328**
RT	13.5	8.40	14.2	116	177	.0300	.151	.531	.452	.481	.530	.0577	.379	.090	.121	.130	.125
GT	14.1	8.39	13.9	122	165	0	0	1	1	1	.602	.0488	.362	1	1	1	1

Table 8: First three rows are metrics computed at the minimum of the validation loss on the unfiltered test set (DailyDialog), while the second three rows are the same models but overfitted. TRF refers to `transformer`, TARGET-side, and filtering BOTH are compared. Best results are highlighted with bold (if significantly better). Best results are given separately for the first three and second three rows. **GT** refers to ground truth responses and **RT** refers to randomly selected responses from the training set.

We compute metrics on the unfiltered test set to show that filtered trainings perform better even on utterances that would have been filtered from the training data. TRF, the baseline `transformer` model trained on unfiltered data is compared to the filtered trainings. In all tables the 17 metrics from left to right are: response length, unigram and bigram entropy, unigram and bigram utterance entropy, unigram and bigram KL divergence, embedding *average*, *extrema* and *greedy*, coherence, *distinct-1* and *distinct-2*, and finally, BLEU-1, BLEU-2, BLEU-3 and BLEU-4 (see Section 2).

On the original DailyDialog dataset evaluating at the minimum validation loss (first three rows of Table 8) clearly shows that models trained on data filtered by IDENTITY surpass the baseline

Input	**Unfiltered**	**Unfiltered-Overfitted**	**TARGET**	**TARGET-overfitted**
some high speed chase scene i think .	i 'm not sure .	but many things are went wrong !	i 'm so sorry .	enough !
how long is this bus ride ?	about twelve hours .	it 's about fortyfive minutes to an hour .	it 's about 20 minutes .	it 's only about fortyfive minutes to an hour .
is that yellow bus it ?	no it 's not far . it 's a bit cold .	yeah . there s a new one in america .	no it 's a yellow one .	no it 's not very expensive these days .
hi mary ! what do you like to do in your spare time ?	i like playing basketball .	well i like to play the piano .	i like playing basketball .	i like playing chess .
let me see . we can visit the country . we have n't done that for such a long time .	i hope so .	yes there are many famous paintings on show . i hope it 's the nicest city in the united states .	i 'm glad to hear that .	where are your parents ?
thirty u . s . dollars . . . and it 's only been used for six months .	that 's fine .	i see . is it a long time to come to the station ?	how many months did you get there ?	what 're two weeks ?
five hundred dollars altogether .	here you are .	here you are .	here is your bill .	here you are . thank you byebye .
look at this bridge . we will soon be crossing the thames .	yes i agree .	oh yeah .	how far is it ?	well there 's a spanish teacher . so i 'll often go back to the spanish museum .
is the furniture crafted well ?	yes it is .	yes it 's fully furnished .	yes the furniture is right over there .	yes it 's a furnished eatin kitchen . the furniture and the icebox are all pretty outdated .
where do they go ?	they 're on the bottom floor .	i have no idea .	they re in the cabinet next to the library .	go to them department please .
let me pay for it .	ok . here you are .	cash credit card or debit card ?	ok . here is the money .	cash credit card or debit card ?
he never does and in fact he seldom communicates with him .	he is a hard job . he can also learn a lot from the government .	that 's right . he has no children and his wife experience too .	he is a very kind of man .	yeah . he 's a very easy teacher .
so you must have had a busy and enjoyable day yesterday .	yes i am . i really enjoyed it .	yes . i am very sorry i can t make it .	i will . thanks for coming .	yes . let 's go .
yes . the reservation under mr . watson .	ok .	could you spell that please ?	i 'm afraid there 's no seat in the middle of the table .	ok . would you please fill out the table ?
there are so many to choose from . i did n't know you could do all of this using a card .	i see .	what 's the best part ?	i 'm not sure . i 'll go to the doctor and get my card right away .	you mean you do n't have to tell me the truth ?
it was great ! how was your year ?	it was great . i had a great time .	my mother had a birthday party for me .	it was great . i 'm glad i did n't like it .	it was ok . there was a lot of turbulence though . the worst thing was n't too good !

Table 9: Responses to randomly selected test inputs which we made sure were not in the training data (DailyDialog). *Unfiltered* is the model trained on unfiltered data, and TARGET is the model trained on IDENTITY, TARGET filtered data. *Overfitted* means that the respective model is evaluated at an overfitted point.

on all but the *distinct-1* metric. It seems like filtering high entropy targets is more important than filtering both sides. By removing example responses that are boring and generic from the dataset the model learns to improve response quality. Finding such utterances is useful for a number of purposes, but earlier it has been done mainly manually [Li et al., 2016d, Shen et al., 2017], whereas we provide an automatic, unsupervised method of detecting them based on entropy. Every value is higher after 150 epochs of training than at the validation loss minimum (second three rows of Table 8). The most striking change is in the unigram KL divergence, which is now an order of magnitude lower. IDENTITY still performs best, falling behind the baseline on only the two *distinct* metrics. In some cases the best performing model gets quite close to the ground truth performance. On metrics that evaluate utterances without context (i.e. entropy, divergence, *distinct*), randomly selected responses achieve similar values as the ground truth, which is expected. However, on embedding metrics, coherence, and BLEU, random responses are significantly worse than those of any model evaluated.

	$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF	9.85	7.12	11.5	71.8	95.1	.895	1.61	.520	.479	.543	.527	**.0133**	.063	.127	.129	.127	.118
TARGET	**11.4**	**7.25**	**12.1**	**85.6**	**120**	**.55**	**1.12**	**.549**	**.495**	.541	**.574**	.013	**.0788**	**.14**	**.143**	**.143**	**.134**
TRF-O	11.1	7.72	12.9	88	130	.0908	.291	.578	.522	.562	.565	.0424	.262	.187	.186	.183	.171

Table 10: First two rows are the baseline and the filtered model evaluated at the validation loss minimum, while last row is an overfitted model. Best results are highlighted with bold and best results separately for each entropy computing method are in italic (and those within a 95% confidence interval).

In order to see if our method improves over DailyDialog even when it's curated we trained models on the previously described DailyDialog Curated dataset (Table 10). Our TARGET filtering improves across the baseline on almost all metrics in this case as well. However, looking at the overfitted results (last row) we can see that they are much worse than on the original DailyDialog dataset, and thus the gap between validation loss minimum and overfitted results is smaller. This proves that curating did indeed solve part of the problem with metrics being higher of overfitted models.

5.5 Cornell and Twitter Results

To further solidify our claims we tested the two best performing variants of BOTH and TARGET filtering on the Cornell Movie-Dialogs Corpus and on a subset of 220k examples from the Twitter corpus[10]. Entropy thresholds were selected to be similar to the DailyDialog experiments (Table 6). Evaluation results at the validation loss minimum on the Cornell corpus and the Twitter dataset are presented in Table 11 and Table 12, respectively. On these noisier datasets our simple IDENTITY method still managed to improve over the baseline, but the impact is not as pronounced and in contrast to DailyDialog, BOTH and TARGET perform best on nearly the same number of metrics.

[10] https://github.com/Marsan-Ma/chat_corpus/

		$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF		8.1	6.55	10.4	54	75	2.29	3.40	**.667**	.451	.635	**.671**	4.7e-4	1.0e-3	**.108**	.120	.120	.112
ID	B	7.4	**6.67**	**10.8**	50	69	**1.96**	**2.91**	.627	**.455**	.633	.637	**2.1e-3**	**7.7e-3**	.106	.113	.111	.103
	T	**12.0**	6.44	10.4	**74**	**106**	2.53	3.79	.646	**.456**	**.637**	.651	9.8e-4	3.2e-3	**.108**	**.123**	**.125**	**.118**
RT		13.4	8.26	14.2	113	170	.03	.12	.623	.386	.601	.622	4.6e-2	3.2e-1	.079	.102	.109	.105
GT		13.1	8.18	13.8	110	149	0	0	1	1	1	.655	4.0e-2	3.1e-1	1	1	1	1

Table 11: Metrics on the unfiltered test set (Cornell) at the validation loss minimum. TRF refers to `transformer`, **ID** to IDENTITY. TARGET-side, and filtering BOTH sides are denoted by initials. Best results are highlighted with bold. **GT** refers to ground truth responses and **RT** refers to randomly selected responses from the training set.

		$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF		20.6	6.89	**11.4**	121	177	**2.28**	**3.40**	.643	.395	.591	.659	**2.1e-3**	**6.2e-3**	.0519	.0666	.0715	.0693
ID	B	20.3	**6.95**	**11.4**	119	171	2.36	**3.41**	**.657**	.394	.595	**.673**	1.2e-3	3.4e-3	**.0563**	**.0736**	.0795	.0774
	T	**29.0**	6.48	10.7	**157**	**226**	2.68	3.69	.644	**.403**	**.602**	.660	1.4e-3	4.6e-3	**.0550**	**.0740**	**.0819**	**.0810**
RT		14.0	9.81	15.9	136	171	.05	.19	.681	.334	.543	.695	8.5e-2	5.4e-1	.0444	.0751	.0852	.0840
GT		14.0	9.78	15.8	135	167	0	0	1	1	1	.734	8.1e-2	5.3e-1	1	1	1	1

Table 12: Metrics on the unfiltered test set (Twitter) at the validation loss minimum. TRF refers to `transformer`, **ID** to IDENTITY. TARGET-side, and filtering BOTH sides are denoted by initials. Best results are highlighted with bold. **GT** refers to ground truth responses and **RT** refers to randomly selected responses from the training set.

Compared to DailyDialog there are some important distinctions that also underline that these datasets are of lesser quality. The COHERENCE metric is worse on the ground truth responses than on model responses (Table 11), and some embedding metrics and BLEU scores are better on randomly selected responses than on model responses (Table 12).

5.6 Gutenberg results

	$\|U\|$	H_w^u	H_w^b	H_u^u	H_u^b	D_{kl}^u	D_{kl}^b	AVG	EXT	GRE	COH	d1	d2	b1	b2	b3	b4
TRF	**17.4**	7.05	**11.9**	119	182	**1.67**	**3.17**	.665	**.472**	**.641**	**.712**	**.0005**	**.0048**	.080	.086	.089	.085
TARG	16.9	7.04	11.4	**121**	183	3.18	5.32	.666	.463	.634	.676	.00005	.0002	.078	**.092**	**.097**	**.093**
RT	24.3	8.79	15.7	217	367	.008	.053	.704	.397	.615	.704	.006	.107	.075	.101	.114	.113
GT	24.3	8.79	15.5	216	357	0	0	1	1	1	.740	.005	.108	1	1	1	1

Table 13: The Transformer trained on the Gutenberg dataset (TRF) and on TARGET filtered data, evaluated on test data. **GT** refers to ground truth responses and **RT** refers to randomly selected responses from the training set. Best results are highlighted with bold (only if significantly better.)

Input	Model Response	Target
i will ! i will ! i will ! i will !	you will not ! you will not ! you will not ! you will not ! you will not ! you will not ! you will n	no , no ! no , no !
oh , please go away ! i - - i do n't want any help . i 'm only looking for crabs .	i 'm going to help you ,	mais , mademoiselle permettez - moi , je vous prie , ah , pauvre petite ! pauvre petite !
your nephew ? your ladyship never told me you had a nephew .	i have never heard of him ,	i must have had it on the tip of my tongue to tell you , over and over again , but we have had so ma
you ai n't going to hit my mother , grandfather !	i 'm going to hit your grandfather ,	no , no , lad i learned my manners better than that ! whatever i was going to say , i was thinking o
and here is the chance ,	i do n't know , i do n't know , i do n't know .	the chance is here ,
it is pleasant when a character we admire is consistent .	it is a book that will be read with interest .	margaret is very consistent .
mammy , of course ,	i 'm glad you 're going to be a good girl ,	and whats your name ?
for looking out for myself . for my part , i had thought it praiseworthy , but he says none of the r	i dont know , i dont know . i dont know . i dont know . i dont know . i dont know . i dont know . i	rotifer , as before .
did england belong to the english before the jews came ?	no , sir .	how do you mean ? of course it did
certainly not . i am quite independent .	i am glad to hear it .	so i should have supposed . then why not come ?
you were a thousand miles away . i passed you twice and you did n't even see me . if it had n't been	i 'm glad you did n't , i 'm glad you did n't .	i 'm very glad you did n't . i really wanted to talk to you .
no , no , certainly not , how could you live alone in that great , empty house ?	i do n't know , i do n't know , i do n't know . i do n't know . i do n't know . i do n't know . i do	i am not afraid of being alone , but let it be as you like , mr . monk - - at any rate , until you g
that 's it , that 's it , wo n't you risk a look - in at the exposition du lis with us , monsieur ge	i do n't know , i do n't know .	well , no , i want to walk . i shall go with monsieur l'abbe froment to the chamber .
why , yes i think i can .	you can ?	i am not jesting i am suffering , rosalind . you will not leave that girl alone with him ?
what is it , john ?	i do n't know ,	good news , wife , such news as i never hoped for or even dreamed of .
that 's your own statement ! fosdick never had brains enough to figure a thing out like that . all h	i do n't know what you mean ,	so you see , crime does not pay . the net has closed over your head . you erred a score of times . y
kate morton ! why , i have n't seen her for ten years !	i do n't know what you mean ,	was it a hopeless affection , then ? are you only true to her memory ?
twenty - five dollars , well , it 's fortunate that i have them . and who are you ? not one of campb	i am a poor man , sir , and i have no money .	i am a confidential messenger , i carry messages and execute commissions that require more or less d
now , edna , i am amazed to hear you speak of the peruvian government , it would be nothing less tha	i am not a criminal , i am a criminal , and i am not a criminal . i am a criminal , and i am not a c	what do you think , captain ?
you mean that , m'sieu ' ?	yes , i mean it .	i swear it .

Table 14: Model responses and targets to 20 randomly selected test inputs. All utterances are truncated to 100 characters.

In this section we explore how the Transformer trained on the Gutenberg dataset performes on its own test set, and also how applying data filtering affects results. In Table 13 we compare the model trained on unfiltered data and the one trained on filtered data. Interestingly the filtered training only helps with bleu scores, while on other metrics the baseline model performs better. Because of resource and time constraints we could not train further models with the other filtering ways, which could work better on this dataset. Conducting these experiments and uncovering the reasons behind why filtering is not that beneficial on Gutenberg is left for future work.

Another finding is that randomly selecting responses from the training set achieve better results across almost all metrics than the trained model. This is partly because these metrics are not adequate, and because the test set is so huge that there is bound to be some inputs where random responses match really well. Table 14 shows some model responses to randomly selected inputs, and it's clear that the model doesn't perform as bad as the automatic metric results would suggest. Obviously, selecting random responses for these inputs would be much worse for us humans, however these metrics don't really measure what we would consider good responses, which is an active problem in dialog modeling and we've also discussed this in previous sections (Section 2).

6 Conclusion

We created a new, large, high-quality dataset for neural dialog modeling. We presented a detailed preprocessing pipeline and error analysis for this dataset. We showed how in the context of transfer learning it performs better than other large dialog datasets.

We also presented a simple unsupervised entropy-based approach that can be applied to any conversational dataset for filtering generic sources/targets that cause "confusion" during the training of open-domain dialog models. We showed the effectiveness of this method on 4 different datasets. Some limitations of current automatic metrics and the loss function have also been shown, by examining their behavior on random data and with overfitting.

In the future, we plan to explore several directions. We wish to improve the quality of the Gutenberg dialog dataset by choosing smarter ways of finding dialogs (e.g. learning classifiers). We also wish to explore how the multilingual nature of the Gutenberg project can be used to create a multilingual Gutenberg Dialog Dataset. Finally, we wish to extend the transfer learning experiments to compare with other big datasets and on other smaller downstream datasets.

Work partially supported by Project FIEK 16-1-2016-0007, financed by the FIEK_16 funding scheme of the Hungarian National Research, Development and Innovation Office (NKFIH). Work partially supported by the ÚNKP-19-2 New National Excellence Program of the Ministry for Innovation and Technology.

References

[Asghar et al., 2017] Asghar, N., Poupart, P., Jiang, X., and Li, H. (2017). Deep active learning for dialogue generation. In *Proceedings of the 6th Joint Conference on Lexical and Computational Semantics (*SEM 2017)*, pages 78–83. Association for Computational Linguistics.

[Axelrod et al., 2011] Axelrod, A., He, X., and Gao, J. (2011). Domain adaptation via pseudo in-domain data selection. In *Proceedings of the 2011 Conference on Empirical Methods in Natural Language Processing*, pages 355–362, Edinburgh, Scotland, UK. Association for Computational Linguistics.

[Bahdanau et al., 2015] Bahdanau, D., Cho, K., and Bengio, Y. (2015). Neural machine translation by jointly learning to align and translate. In *International Conference on Learning Representations (ICLR 2015)*.

[Baheti et al., 2018] Baheti, A., Ritter, A., Li, J., and Dolan, B. (2018). Generating more interesting responses in neural conversation models with distributional constraints. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3970–3980. Association for Computational Linguistics.

[Chen and Cherry, 2014] Chen, B. and Cherry, C. (2014). A systematic comparison of smoothing techniques for sentence-level BLEU. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 362–367, Baltimore, Maryland, USA. Association for Computational Linguistics.

[Cho et al., 2014] Cho, K., van Merriënboer, B., Gulcehre, C., Bougares, F., Schwenk, H., and Bengio, Y. (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.

[Csaky, 2019] Csaky, R. (2019). Deep learning based chatbot models. National Scientific Students' Associations Conference. https://arxiv.org/abs/1908.08835.

[Csáky et al., 2019] Csáky, R., Purgai, P., and Recski, G. (2019). Improving neural conversational models with entropy-based data filtering. In *Proceedings of the 57th Annual Meeting of the Association for Computational Linguistics*, pages 5650–5669, Florence, Italy. Association for Computational Linguistics.

[Dai et al., 2019] Dai, Z., Yang, Z., Yang, Y., Cohen, W. W., Carbonell, J., Le, Q. V., and Salakhutdinov, R. (2019). Transformer-xl: Attentive language models beyond a fixed-length context. *arXiv preprint arXiv:1901.02860*.

[Danescu-Niculescu-Mizil and Lee, 2011] Danescu-Niculescu-Mizil, C. and Lee, L. (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.

[Devlin et al., 2018] Devlin, J., Chang, M.-W., Lee, K., and Toutanova, K. (2018). Bert: Pre-training of deep bidirectional transformers for language understanding.

[Dinan et al., 2019] Dinan, E., Roller, S., Shuster, K., Fan, A., Auli, M., and Weston, J. (2019). Wizard of wikipedia: Knowledge-powered conversational agents. In *International Conference on Learning Representations*.

[Fainberg et al., 2018] Fainberg, J., Krause, B., Dobre, M., Damonte, M., Kahembwe, E., Duma, D., Webber, B., and Fancellu, F. (2018). Talking to myself: self-dialogues as data for conversational agents. *arXiv preprint arXiv:1809.06641*.

[Fang et al., 2018] Fang, H., Cheng, H., Sap, M., Clark, E., Holtzman, A., Choi, Y., Smith, N. A., and Ostendorf, M. (2018). Sounding board: A user-centric and content-driven social chatbot. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Demonstrations*, pages 96–100. Association for Computational Linguistics.

[Gao et al., 2019] Gao, X., Lee, S., Zhang, Y., Brockett, C., Galley, M., Gao, J., and Dolan, B. (2019). Jointly optimizing diversity and relevance in neural response generation. *arXiv preprint arXiv:1902.11205*.

[Ghazvininejad et al., 2018] Ghazvininejad, M., Brockett, C., Chang, M.-W., Dolan, B., Gao, J., Yih, W.-t., and Galley, M. (2018). A knowledge-grounded neural conversation model. In *Thirty-Second AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

[Graves, 2012] Graves, A. (2012). Sequence transduction with recurrent neural networks. In *Representation Learning Workshop, ICML 2012*, Edinburgh, Scotland.

[Gu et al., 2019] Gu, X., Cho, K., Ha, J.-W., and Kim, S. (2019). DialogWAE: Multimodal response generation with conditional wasserstein auto-encoder. In *International Conference on Learning Representations*.

[Hancock et al., 2019] Hancock, B., Bordes, A., Mazare, P.-E., and Weston, J. (2019). Learning from dialogue after deployment: Feed yourself, chatbot! *arXiv preprint arXiv:1901.05415*.

[Henderson et al., 2019] Henderson, M., Budzianowski, P., Casanueva, I., Coope, S., Gerz, D., Kumar, G., Mrkšić, N., Spithourakis, G., Su, P.-H., Vulić, I., et al. (2019). A repository of conversational datasets. *arXiv preprint arXiv:1904.06472*.

[Joshi et al., 2017] Joshi, C. K., Mi, F., and Faltings, B. (2017). Personalization in goal-oriented dialog. *arXiv preprint arXiv:1706.07503*.

[Kandasamy et al., 2017] Kandasamy, K., Bachrach, Y., Tomioka, R., Tarlow, D., and Carter, D. (2017). Batch policy gradient methods for improving neural conversation models. *arXiv preprint arXiv:1702.03334*.

[Kingma and Ba, 2014] Kingma, D. P. and Ba, J. (2014). Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

[Krause et al., 2017] Krause, B., Damonte, M., Dobre, M., Duma, D., Fainberg, J., Fancellu, F., Kahembwe, E., Cheng, J., and Webber, B. (2017). Edina: Building an open domain socialbot with self-dialogues. In *1st Proceedings of Alexa Prize (Alexa Prize 2017)*.

[Kulikov et al., 2018] Kulikov, I., Miller, A. H., Cho, K., and Weston, J. (2018). Importance of a search strategy in neural dialogue modelling. *arXiv preprint arXiv:1811.00907*.

[Lewis et al., 2017] Lewis, M., Yarats, D., Dauphin, Y. N., Parikh, D., and Batra, D. (2017). Deal or no deal? end-to-end learning for negotiation dialogues. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2443–2453. Association for Computational Linguistics.

[Li et al., 2016a] Li, J., Galley, M., Brockett, C., Gao, J., and Dolan, B. (2016a). A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT 2016*, pages 110–119. Association for Computational Linguistics.

[Li et al., 2016b] Li, J., Galley, M., Brockett, C., Spithourakis, G. P., Gao, J., and Dolan, B. (2016b). A persona-based neural conversation model. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics*, pages 994–1003. Association for Computational Linguistics.

[Li et al., 2016c] Li, J., Miller, A. H., Chopra, S., Ranzato, M., and Weston, J. (2016c). Dialogue learning with human-in-the-loop. *arXiv preprint arXiv:1611.09823*.

[Li et al., 2017a] Li, J., Monroe, W., and Jurafsky, D. (2017a). Data distillation for controlling specificity in dialogue generation. *arXiv preprint arXiv:1702.06703*.

[Li et al., 2016d] Li, J., Monroe, W., Ritter, A., Galley, M., Gao, J., and Jurafsky, D. (2016d). Deep reinforcement learning for dialogue generation. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1192–1202. Association for Computational Linguistics.

[Li et al., 2017b] Li, J., Monroe, W., Shi, T., Ritter, A., and Jurafsky, D. (2017b). Adversarial learning for neural dialogue generation. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2157–2169. Association for Computational Linguistics.

[Li et al., 2017c] Li, Y., Su, H., Shen, X., Li, W., Cao, Z., and Niu, S. (2017c). Dailydialog: A manually labelled multi-turn dialogue dataset. In *Proceedings of the The 8th International Joint Conference on Natural Language Processing*, pages 986–995. AFNLP.

[Lipton et al., 2018] Lipton, Z., Li, X., Gao, J., Li, L., Ahmed, F., and Deng, L. (2018). Bbq-networks: Efficient exploration in deep reinforcement learning for task-oriented dialogue systems. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence.

[Liu et al., 2016] Liu, C.-W., Lowe, R., Serban, I., Noseworthy, M., Charlin, L., and Pineau, J. (2016). How not to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation. In *Proceedings of the 2016 Conference on*

Empirical Methods in Natural Language Processing, pages 2122–2132. Association for Computational Linguistics.

[Liu et al., 2017] Liu, H., Lin, T., Sun, H., Lin, W., Chang, C.-W., Zhong, T., and Rudnicky, A. (2017). Rubystar: A non-task-oriented mixture model dialog system. In *1st Proceedings of Alexa Prize (Alexa Prize 2017)*.

[Lowe et al., 2017] Lowe, R., Noseworthy, M., Serban, I. V., Angelard-Gontier, N., Bengio, Y., and Pineau, J. (2017). Towards an automatic turing test: Learning to evaluate dialogue responses. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 1116–1126. Association for Computational Linguistics.

[Ludwig, 2017] Ludwig, O. (2017). End-to-end adversarial learning for generative conversational agents. *arXiv preprint arXiv:1711.10122*.

[Mazare et al., 2018] Mazare, P.-E., Humeau, S., Raison, M., and Bordes, A. (2018). Training millions of personalized dialogue agents. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2775–2779. Association for Computational Linguistics.

[Mo et al., 2017] Mo, K., Zhang, Y., Yang, Q., and Fung, P. (2017). Fine grained knowledge transfer for personalized task-oriented dialogue systems. *arXiv preprint arXiv:1711.04079*.

[Moghe et al., 2018] Moghe, N., Arora, S., Banerjee, S., and Khapra, M. M. (2018). Towards exploiting background knowledge for building conversation systems. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 2322–2332, Brussels, Belgium. Association for Computational Linguistics.

[Mou et al., 2016] Mou, L., Song, Y., Yan, R., Li, G., Zhang, L., and Jin, Z. (2016). Sequence to backward and forward sequences: A content-introducing approach to generative short-text conversation. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*, pages 3349–3358. The COLING 2016 Organizing Committee.

[Olabiyi et al., 2018] Olabiyi, O., Salimov, A., Khazane, A., and Mueller, E. (2018). Multi-turn dialogue response generation in an adversarial learning framework. *arXiv preprint arXiv:1805.11752*.

[Papineni et al., 2002] Papineni, K., Roukos, S., Ward, T., and Zhu, W.-J. (2002). Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics*, Philadelphia.

[Park et al., 2018] Park, Y., Cho, J., and Kim, G. (2018). A hierarchical latent structure for variational conversation modeling. In *Proceedings of NAACL-HLT 2018*, pages 1792–1801. Association for Computational Linguistics.

[Ram et al., 2018] Ram, A., Prasad, R., Khatri, C., Venkatesh, A., Gabriel, R., Liu, Q., Nunn, J., Hedayatnia, B., Cheng, M., Nagar, A., et al. (2018). Conversational ai: The science behind the alexa prize. *arXiv preprint arXiv:1801.03604*.

[Sedoc et al., 2018] Sedoc, J., Ippolito, D., Kirubarajan, A., Thirani, J., Ungar, L., and Callison-Burch, C. (2018). Chateval: A tool for the systematic evaluation of chatbots. In *Proceedings of the Workshop on Intelligent Interactive Systems and Language Generation (2IS&NLG)*, pages 42–44. Association for Computational Linguistics.

[Serban et al., 2017a] Serban, I. V., Sankar, C., Germain, M., Zhang, S., Lin, Z., Subramanian, S., Kim, T., Pieper, M., Chandar, S., Ke, N. R., et al. (2017a). A deep reinforcement learning chatbot. *arXiv preprint arXiv:1709.02349*.

[Serban et al., 2016] Serban, I. V., Sordoni, A., Bengio, Y., Courville, A. C., and Pineau, J. (2016). Building end-to-end dialogue systems using generative hierarchical neural network models. In *AAAI*, pages 3776–3784.

[Serban et al., 2017b] Serban, I. V., Sordoni, A., Lowe, R., Charlin, L., Pineau, J., Courville, A. C., and Bengio, Y. (2017b). A hierarchical latent variable encoder-decoder model for generating dialogues. In *Thirty-First AAAI Conference on Artificial Intelligence*. Association for the Advancement of Artificial Intelligence.

[Shalyminov et al., 2018] Shalyminov, I., Dušek, O., and Lemon, O. (2018). Neural response ranking for social conversation: A data-efficient approach. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI*, pages 1–8. Association for Computational Linguistics.

[Shao et al., 2017] Shao, Y., Gouws, S., Britz, D., Goldie, A., Strope, B., and Kurzweil, R. (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 2210–2219. Association for Computational Linguistics.

[Shen et al., 2018a] Shen, X., Su, H., Li, W., and Klakow, D. (2018a). 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. Association for Computational Linguistics.

[Shen et al., 2017] Shen, X., Su, H., Li, Y., Li, W., Niu, S., Zhao, Y., Aizawa, A., and Long, G. (2017). A conditional variational framework for dialog generation. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 504–509. Association for Computational Linguistics.

[Shen et al., 2018b] Shen, X., Su, H., Niu, S., and Demberg, V. (2018b). Improving variational encoder-decoders in dialogue generation. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence.

[Sutskever et al., 2014] Sutskever, I., Vinyals, O., and Le, Q. V. (2014). Sequence to sequence learning with neural networks. In *Proc. NIPS*, pages 3104–3112, Montreal, CA.

[Tandon et al., 2017] Tandon, S., Bauer, R., et al. (2017). A dual encoder sequence to sequence model for open-domain dialogue modeling. *arXiv preprint arXiv:1710.10520*.

[Tao et al., 2018] Tao, C., Mou, L., Zhao, D., and Yan, R. (2018). Ruber: An unsupervised method for automatic evaluation of open-domain dialog systems. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence.

[Vaswani et al., 2017] Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., Kaiser, L. u., and Polosukhin, I. (2017). Attention is all you need. In Guyon, I., Luxburg, U. V., Bengio, S., Wallach, H., Fergus, R., Vishwanathan, S., and Garnett, R., editors, *Advances in Neural Information Processing Systems 30*, pages 5998–6008. Curran Associates, Inc.

[Vinyals and Le, 2015] Vinyals, O. and Le, Q. V. (2015). A neural conversational model. In *Proceedings of the 31st International Conference on Machine Learning*.

[Wang et al., 2018] Wang, Y., Liu, C., Huang, M., and Nie, L. (2018). Learning to ask questions in open-domain conversational systems with typed decoders. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2193–2203. Association for Computational Linguistics.

[Wei et al., 2017] Wei, B., Lu, S., Mou, L., Zhou, H., Poupart, P., Li, G., and Jin, Z. (2017). Why do neural dialog systems generate short and meaningless replies? a comparison between dialog and translation. *arXiv preprint arXiv:1712.02250*.

[Wojciechowski and Zakrzewicz, 2002] Wojciechowski, M. and Zakrzewicz, M. (2002). Dataset filtering techniques in constraint-based frequent pattern mining. In *Pattern detection and discovery*, pages 77–91. Springer.

[Wu et al., 2018] Wu, B., Jiang, N., Gao, Z., Li, S., Rong, W., and Wang, B. (2018). Why do neural response generation models prefer universal replies? *arXiv preprint arXiv:1808.09187*.

[Xing et al., 2017] Xing, C., Wu, W., Wu, Y., Liu, J., Huang, Y., Zhou, M., and Ma, W.-Y. (2017). Topic aware neural response generation. In *Proceedings of the Thirty-First AAAI Conference on Artificial Intelligence (AAAI-17)*. Association for the Advancement of Artificial Intelligence.

[Xing et al., 2018] Xing, C., Wu, Y., Wu, W., Huang, Y., and Zhou, M. (2018). Hierarchical recurrent attention network for response generation. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence.

[Xing and Fernández, 2018] Xing, Y. and Fernández, R. (2018). Automatic evaluation of neural personality-based chatbots. In *Proceedings of The 11th International Natural Language Generation Conference*, pages 189–194. Association for Computational Linguistics.

[Xu et al., 2018a] Xu, C., Wu, W., and Wu, Y. (2018a). Towards explainable and controllable open domain dialogue generation with dialogue acts. *arXiv preprint arXiv:1807.07255*.

[Xu et al., 2018b] Xu, X., Dušek, O., Konstas, I., and Rieser, V. (2018b). Better conversations by modeling, filtering, and optimizing for coherence and diversity. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*, pages 3981–3991. Association for Computational Linguistics.

[Zhang et al., 2018a] Zhang, H., Lan, Y., Guo, J., Xu, J., and Cheng, X. (2018a). Reinforcing coherence for sequence to sequence model in dialogue generation. In *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, pages 4567–4573.

[Zhang et al., 2018b] Zhang, S., Dinan, E., Urbanek, J., Szlam, A., Kiela, D., and Weston, J. (2018b). Personalizing dialogue agents: I have a dog, do you have pets too? In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 2204–2213. Association for Computational Linguistics.

[Zhang et al., 2018c] Zhang, Y., Galley, M., Gao, J., Gan, Z., Li, X., Brockett, C., and Dolan, B. (2018c). Generating informative and diverse conversational responses via adversarial information maximization. In *32nd Conference on Neural Information Processing Systems (NeurIPS 2018)*.

[Zhao et al., 2018] Zhao, T., Lee, K., and Eskenazi, M. (2018). Unsupervised discrete sentence representation learning for interpretable neural dialog generation. In *Proceedings of the 56th Annual Meeting of the Association for Computational Linguistics (Long Papers)*, pages 1098–1107. Association for Computational Linguistics.

[Zhao et al., 2017] Zhao, T., Zhao, R., and Eskenazi, M. (2017). Learning discourse-level diversity for neural dialog models using conditional variational autoencoders. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 654–664. Association for Computational Linguistics.

[Zhou et al., 2018a] Zhou, H., Huang, M., Zhang, T., Zhu, X., and Liu, B. (2018a). Emotional chatting machine: Emotional conversation generation with internal and external memory. In *The Thirty-Second AAAI Conference on Artificial Intelligence (AAAI-18)*. Association for the Advancement of Artificial Intelligence.

[Zhou et al., 2018b] Zhou, K., Prabhumoye, S., and Black, A. W. (2018b). A dataset for document grounded conversations. *arXiv preprint arXiv:1809.07358*.

[Zhu et al., 2017] Zhu, W., Mo, K., Zhang, Y., Zhu, Z., Peng, X., and Yang, Q. (2017). Flexible end-to-end dialogue system for knowledge grounded conversation. *arXiv preprint arXiv:1709.04264*.