Adversarial Evaluation for Open-Domain Dialogue Generation

Elia Bruni and **Raquel Fernández**
Institute for Logic, Language and Computation
University of Amsterdam
elia.bruni@gmail.com raquel.fernandez@uva.nl

Abstract

We investigate the potential of adversarial evaluation methods for open-domain dialogue generation systems, comparing the performance of a discriminative agent to that of humans on the same task. Our results show that the task is hard, both for automated models and humans, but that a discriminative agent can learn patterns that lead to above-chance performance.

1 Introduction

End-to-end dialogue response generation systems trained to produce a plausible utterance given some limited dialogue context are receiving increased attention (Vinyals and Le, 2015; Sordoni et al., 2015; Serban et al., 2016; Li et al., 2016). However, for systems dealing with chatbot-style open-dialogue, where task completion is not applicable, evaluating the quality of their responses remains a challenge. Most current models are evaluated with measures such as perplexity and overlap-based metrics like BLEU, that compare the generated response to the ground-truth response in an actual dialogue. This kind of measures, however, correlate very weakly or not at all with human judgements on response quality (Liu et al., 2016).

In this paper, we explore a different approach to evaluating open-domain dialogue response generation systems, inspired by the classic Turing Test (Turing, 1950): measuring the quality of the generated responses on their indistinguishability from human output. This approach has been preliminary explored in recent work under the heading of *adversarial evaluation* (Kannan and Vinyals, 2016; Li et al., 2017), drawing a parallel with generative adversarial learning (Goodfellow et al., 2014). Here we concentrate on exploring the potential and the limits of such an adversarial evaluation approach by conducting an in-depth analysis. We implement a discriminative model and train it on the task of distinguishing between actual and "fake" dialogue excerpts and evaluate its performance, as well as the feasibility of the task more generally, by conducting an experiment with human judgements. Results show that the task is hard not only for the discriminative model, but also for human judges. We then implement a simple chatbot agent for dialogue generation and test the discriminator on this data, again comparing its performance to that of humans on this task. We show that both humans and the discriminative model can be fooled by the generator in a significant amount of cases.

2 The Discriminative Agent

Our discriminative agent is a binary classifier which takes as input a sequence of dialogue utterances and predicts whether the dialogue is real or fake. The agent treats as positive examples of coherent dialogue actual dialogue passages and as negative examples passages where the last utterance has been randomly replaced. Random replacement has been used in the past to study discourse coherence (Li and Hovy, 2014).

2.1 Model

The classifier is modelled as an attention-based bidirectional LSTM. LSTMs are indeed very effective to model word sequences, and are especially suited for learning on data with long distance dependencies (Hochreiter and Schmidhuber, 1997) such as multi-turn dialogues. The bidirectional LSTM includes both a forward function ($\overrightarrow{\text{LSTM}}$, which reads the sentence s_i from w_{i1} to w_{iT}) and a backward function ($\overleftarrow{\text{LSTM}}$, which reads the sentence s_i from w_{iT} to w_{i1}):

$$x_{it} = W_e w_{it}, t \in [1, T] \qquad [1]$$

Proceedings of the SIGDIAL 2017 Conference, pages 284–288,
Saarbrücken, Germany, 15-17 August 2017. ©2017 Association for Computational Linguistics

$$\overrightarrow{h}_{it} = \overrightarrow{\text{LSTM}}(x_{it}), t \in [1, T] \qquad [2]$$

$$\overleftarrow{h}_{it} = \overleftarrow{\text{LSTM}}(x_{it}), t \in [T, 1] \qquad [3]$$

The words of a dialogue turn do not always contribute equally to determine coherence. We thus use an attention mechanism to extract words that are important to detect plausibility or coherence of a dialogue passage and parametrize their aggregation accordingly. Having an aggregated vector representation which is adaptive to the content of each time step allows the classifier to assign large weights to the most "discriminative" words. Contemporarily, the attention should also have an advantage in modelling long sequences by considering different word locations in the dialogue in a relatively even manner:

$$u_{it} = \tanh(W_w h_{it} + b_w) \qquad [4]$$

$$\alpha_{it} = \frac{\exp(u_{it}^T u_w)}{\sum_t \exp(u_{it}^T u_w)}, \quad \sum_i \alpha_i h_i \qquad [5]$$

We first compute the hidden representation of h_{it} through a one-layer MLP u_{it}; we then weight the importance of u_{it} by computing its similarity to a word-level context vector, normalized via a softmax function. The context vector is learned end-to-end by the classifier and is meant to represent a general query about the level of "discriminability" of a word (see, e.g., Sukhbaatar et al. 2015 or Yang et al. 2016). The output of the attention is then fed to a sigmoid function, which returns the probability of the input being real or fake:

$$p = \text{sigmoid}(W_c^v + b_c) \qquad [6]$$

As loss function we then use the negative log likelihood of the correct labels:

$$L = -\sum_d \log p_{dj} \qquad [7]$$

2.2 Training Details

We trained the discriminator with a combination of three different datasets: *MovieTriples*, *SubTle* and *Switchboard*. MovieTriples (Serban et al., 2016) has been created from the Movie-Dic corpus of film transcripts (Banchs, 2012) and contains 3-utterance passages between two interlocutors who alternate in the conversation. SubTle (Ameixa et al., 2014) is made of 2-utterance passages extracted from movie subtitles. To discourage the pairing of utterances coming from different movie scenes, we selected only those pairs with a maximum difference of 1 second between the first and the second turn. Switchboard (Godfrey et al., 1992) is a corpus of transcribed telephone conversations. We ignored utterances that consist only of non-verbal acts such as laughter, and selected sequences of three consecutive utterances. In all cases, we consider the last utterance of a passage the target response, and the previous utterances, the context. For the three datasets, we restrict ourselves to dialogue passages where the context and the response have a length of 3 to 25 tokens each. We concatenated the three datasets, obtaining a total of 3,289,835 dialogue passages (46,499 from MovieTriples, 3,211,899 from SubTle, and 77,936 from Switchboard).

For training, we limit the vocabulary size to the top 25K most frequent words.[1] We used mini-batch stochastic gradient descent, shuffling the batches each epoch. We use a bidirectional layer, with 500 cells, and 500-dimensional embeddings (we tried with more layers and higher number of cells without significant improvements). All model parameters are uniformly initialized in $[-0.1, 0.1]$ and as optimizer we used Adam with an initial learning rate of 0.001. Dropout with probability 0.3 was applied to the LSTMs.

3 Human Evaluation

To assess the performance of our discriminative model, we conduct an experiment with human annotators. To our knowledge, this is the first study of its kind ever conducted. Previous human evaluation experiments of dialogue generation systems have mostly consisted in asking participants to choose the better response between two options generated by different models or to rate a generated dialogue along several dimensions (Vinyals and Le, 2015; Lowe et al., 2017; Li et al., 2017). In contrast, here we present humans with the same task faced by the discriminator: We show them a dialogue passage and ask them to decide whether, given the first one or two utterances of context, the shown continuation is the actual follow-up utterance in the original dialogue or a random response.

The data for this experiment consists of 900 pas-

[1] All remaining words are converted into the universal token <unk>.

data	discriminator							humans							agreement	
		real			random				real			random			Fleiss' π	
	Acc	P	R	F1	P	R	F1	Acc	P	R	F1	P	R	F1	hum	disc
SWB	.583	.549	.933	.691	.778	.233	.359	.670	.650	.714	.690	.695	.604	.647	.299	.068
MOV	.677	.645	.787	.709	.726	.567	.637	.677	.664	.713	.688	.690	.640	.664	.303	.258
SUB	.737	.763	.687	.723	.715	.787	.749	.640	.635	.660	.647	.646	.620	.633	.304	.301

Table 1: Accuracy, Precision, Recall, and F-score of discriminator and humans against ground-truth. Inter-annotator agreement among humans and between the discriminator and the human majority class.

sages: 300 randomly selected per dataset, with 50% real and 50% fake dialogues. We use the CrowdFlower platform to recruit annotators, restricting the pool to English native speakers.[2] Each item is classified as real or random by three different annotators. A total of 137 annotators participated in the experiment, with each of them annotating between 10 and 150 items.

We test the discriminator on the same data and compare its performance to the human judgements. Chance level accuracy for both humans and the discriminator is 50%, namely when real and fake passages are indistinguishable from each other. The results are summarised in Table 1. Let us first consider the performance of humans on the task. We compute inter-annotator agreement using Fleiss π (Fleiss, 1971), suitable for assessing multi-coder annotation tasks. Agreement is low: $\pi = 0.30$ across the 3 corpora, indicating that the task is challenging for humans (there is limited consensus on whether the shown dialogue passages are plausible or not). Looking into the human performance with respect to the ground truth, we see similar accuracy scores for Switchboard and MovieTriples, while accuracy is lower for SubTle, where the context consists of one utterance only. Across the three datasets, we observe slightly higher F-score for positive instances (real) than negative instances (random). For the positive instances, recall is higher than precision, while the opposite is true for negative instances. Arguably, this indicates that humans tend to accommodate responses that in fact are random as possible coherent continuations of a dialogue, and will only flag them as fake if they are utterly surprising.

We compute the agreement of the discriminator's predictions and the human majority class over 3 annotators. For Switchboard, agreement is at chance level ($\pi = .07$), while for the other two datasets it is on a par with agreement among humans. As for the discriminator's performance with respect to the ground truth, not surprisingly we obtain low accuracy on Switchboard, but slightly higher accuracy than humans in the other datasets, in particular SubTle, possibly due to the larger amount of training data from this corpus. In what follows, we investigate what information the discriminator may be exploiting to make its predictions.

4 Analysis

To inspect the discriminator's internal representation of the dialogue turns, at testing time we run two extra forward passes, inputting context and target separately, and compute the cosine similarity between the respective LSTM hidden states. We find some clear patterns: The context and response of the dialogue passages classified as coherent by the discriminator (true and false positives) have significantly higher cosine similarity than the passages classified as fake (true and false negatives). This holds across the 3 datasets ($p < .001$ on a two-sample Wilcoxon rank sum test) and indicates that the discriminator is exploiting this information to make its predictions. We also observe that, while there is a tendency to higher cosine similarity in the ground-truth positive instances than in the negative ones in Switchboard ($p = .05$) and MovieTriples ($p = .03$), the effect is highly significant in SubTle ($p < .001$), which is in line with the higher performance of the discriminator on this corpus. Since accuracy is higher than humans in this case, presumably the discriminator is sensitive to patterns that may not be apparent to humans. Whether this capacity is useful for developing generative models that interact with humans, however, is an open question.

We find another interesting pattern within the attention mass distribution between context and target: For true and false positives, higher attention is concentrated on the response ($\approx$ 90%),

[2] We use strict quality controls, only accepting annotators considered "highly trusted" by CrowdFlower (www.crowdflower.com) and requiring 90% accuracy on so-called "test questions". Annotators are paid $4 cents per item.

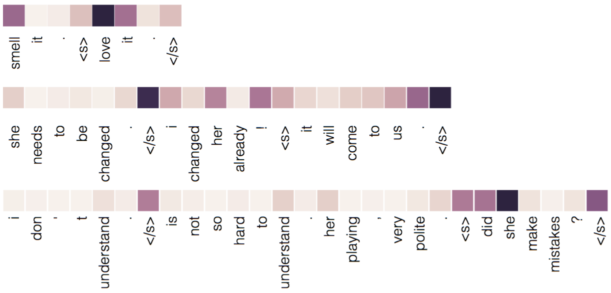

Figure 1: Attention visualization.

while for true and false negatives the attention is more balanced between the two ($\approx 50\%$). Figure 1 shows three sample dialogue passages with word-level attention weights displayed in different color intensities. The token <s> separates the context from the target response. The sample at the top is a passage from SubTle that humans judged to be incoherent, but that was rightly classified by the discriminator as a positive instance (the passage is real). The sample in the middle (a passage from MovieTriples where the target is random) illustrates how attention weights are more balanced in negative instances. Finally, the sample at the bottom shows a passage from MovieTriples rightly classified as coherent by human annotators and by the discriminative agent. As can be seen, attention is more prominent on the target response, with particular focus on the pronoun 'she' whose antecedent 'her' in the context also receives some attention mass. In all cases the token </s> receives high attention, suggesting that the discriminative agent is keeping track of turn alternations.

5 Discriminating Generated Responses

We implement a baseline generative agent to test the extent to which the discriminator's ability to distinguish between generated and actual responses is comparable to humans.

5.1 The Generator Agent

The generator directly models the conditional probability $p(y|x)$ of outputting the subsequent dialogue turn $y_1, ..., y_m$ given some previous context $x_1, ..., x_n$. The model consists of a SEQ2SEQ model, divided into two components: an *encoder* which computes a representation for the dialogue context and a *decoder* which generates the subsequent dialogue turn one word at a time. A natural choice for implementing both the *encoder* and the *decoder* is to use an LSTM (see Section 2). The *decoder* is also equipped with an attention system.

We train the generator to predict the next dialogue turn given the preceding dialogue history on the OpenSubtitles dataset (Tiedemann, 2009). We considered each line in the dataset as a target to be predicted by the model and the concatenation of the two foregoing lines as the source context. We opt for OpenSubtitles rather than for the cleaner datasets used for training the discriminative agent, because the SEQ2SEQ model requires a very large amount of data to converge, and with more than 80 million triples, OpenSubtitles is one of the largest dialogue dataset available.

During training, we filtered out passages with context or target longer than 25 words. We used mini-batch stochastic gradient descent, shuffling the batches each epoch. We use stacking LSTM with 2 bidirectional layers, each with 2048 cells, and 500-dimensional embeddings. All model parameters are uniformly initialized in $[-0.1, 0.1]$; we train using SGD, with a start learning rate of 1, and after 5 epochs we start halving the learning rate at each epoch; the mini-batch size is set to 64 and we rescale the normalized gradients whenever the norm exceeds 5. We also apply dropout with probability 0.3 on the LSTMs.

5.2 Results

We test our discriminative agent on the task of distinguishing passages with real responses versus generated responses and, as before, compare its performance to human performance. For this evaluation, we selected a random sample of 30 generated instances per corpus, avoiding repeated generated responses and responses with <unk> tokens since these would make the human judgements trivial. A summary of results is shown in Table 2. We can see that human accuracy is at chance level, while the discriminator's is above chance, again suggesting that the discriminator may pick up on patterns that are not discernible to humans. The higher performance on SubTle may again be explained by the larger amount of training data from this dataset. We also observe very low inter-annotator agreement, with even negative π for the discriminator with respect to humans in the case of Switchboard.

6 Conclusions

In this paper, we investigated the use of an adversarial setting for open domain dialogue eval-

data	discriminator							humans							agreement	
		real			generated				real			generated			Fleiss' π	
	Acc	P	R	F1	P	R	F1	Acc	P	R	F1	P	R	F1	hum	disc
SWB	.567	.538	.933	.683	.750	.200	.316	.517	.511	.755	.610	.532	.277	.365	.194	−.130
MOV	.633	.618	.700	.656	.654	.567	.607	.467	.478	.733	.579	.428	.200	.273	.177	.062
SUB	.700	.773	.567	.654	.658	.833	.736	.511	.508	.678	.581	.517	.344	.413	.258	.129

Table 2: Performance of discriminator and humans against ground-truth for generator experiment. Inter-annotator agreement among humans and between the discriminator and the human majority class.

uation, providing novel results on human performance that are informative of the difficulty of the task and the strategies employed to tackle it. We found that there is limited consensus among human annotators on what counts as a coherent dialogue passages when only 1 or 2 utterances of context are provided, but that nevertheless a discriminative model is able to learn patterns that lead to above-chance performance.

References

David Ameixa, Luisa Coheur, Pedro Fialho, and Paulo Quaresma. 2014. *Luke, I am your father*: Dealing with out-of-domain requests by using movies subtitles. In *International Conference on Intelligent Virtual Agents*. Springer, pages 13–21.

Rafael E Banchs. 2012. Movie-DiC: a movie dialogue corpus for research and development. In *Proceedings ACL-2012: Short Papers-Volume 2*. pages 203–207.

Joseph L. Fleiss. 1971. Measuring nominal scale agreement among many raters. *Psychological Bulletin* 76(5):378–382.

John J Godfrey, Edward C Holliman, and Jane McDaniel. 1992. Switchboard: Telephone speech corpus for research and development. In *Proceedings of the IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-92)*. volume 1, pages 517–520.

Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative adversarial nets. In *Advances in neural information processing systems*. pages 2672–2680.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation* 9(8):1735–1780.

Anjuli Kannan and Oriol Vinyals. 2016. Adversarial evaluation of dialogue models. In *NIPS Workshop on Adversarial Training*.

Jiwei Li and Eduard H. Hovy. 2014. A model of coherence based on distributed sentence representation. In *Proceedings of EMNLP*.

Jiwei Li, Will Monroe, Alan Ritter, Michel Galley, Jianfeng Gao, and Dan Jurafsky. 2016. Deep reinforcement learning for dialogue generation. In *Proceedings of EMNLP*.

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

Chia-Wei Liu, Ryan Lowe, Iulian Serban, Mike 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. In *Proceedings of the 2016 Conference on EMNLP*.

Ryan Lowe, Michael Noseworthy, Iulian Serban, Nicolas Angelard-Gontier, Yoshua Bengio, and Joelle Pineau. 2017. Towards an automatic Turing test: Learning to evaluate dialogue responses. In *Proceedings of ACL*.

Iulian V Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*.

Alessandro Sordoni, Michel Galley, Michael Auli, Chris Brockett, Yangfeng Ji, Margaret Mitchell, Jian-Yun Nie, Jianfeng Gao, and Bill Dolan. 2015. A neural network approach to context-sensitive generation of conversational responses. In *Proceedings of NAACL-HLT*. pages 196–205.

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, et al. 2015. End-to-end memory networks. In *Advances in neural information processing systems*.

Jörg Tiedemann. 2009. News from opus-a collection of multilingual parallel corpora with tools and interfaces. In *Recent advances in natural language processing*. volume 5, pages 237–248.

Alan M Turing. 1950. Computing machinery and intelligence. *Mind* 59(236):433–460.

Oriol Vinyals and Quoc V. Le. 2015. A neural conversational model. In *ICML Deep Learning Workshop*.

Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of NAACL-HLT*. pages 1480–1489.