Topic-relevant Response Generation using Optimal Transport for an Open-domain Dialog System

Shuying Zhang
Kyoto University

Tianyu Zhao
Kyoto University

Tatsuya Kawahara
Kyoto University

{zhang,zhao,kawahara}@sap.ist.i.kyoto-u.ac.jp

Abstract

Conventional neural generative models tend to generate safe and generic responses which have little connection with previous utterances semantically and would disengage users in a dialog system. To generate relevant responses, we propose a method that employs two types of constraints - topical constraint and semantic constraint. Under the hypothesis that a response and its context have higher relevance when they share the same topics, the topical constraint encourages the topics of a response to match its context by conditioning response decoding on topic words' embeddings. The semantic constraint, which encourages a response to be semantically related to its context by regularizing the decoding objective function with semantic distance, is proposed. Optimal transport is applied to compute a weighted semantic distance between the representation of a response and the context. Generated responses are evaluated by automatic metrics, as well as human judgment, showing that the proposed method can generate more topic-relevant and content-rich responses than conventional models.

1 Introduction

In the past decade, personal assistants based on a speech dialog system have shown significant potential and become a trend, which have aroused great interest in dialog systems. Recently, because of the access to large public conversation datasets, neural response generation has been investigated for an open-domain dialog system. Neural response generation is to learn a response generation model under sequence-to-sequence (seq2seq) (Bahdanau et al., 2014) framework. Such a model is trained to optimize the conditional likelihood $P(Y|X)$ of generating a response Y given its context X. This objective function is especially suitable for tasks like machine translation, of which the source and its target translation are semantically corresponding. However, for open-domain dialog response generation, a response can be weakly relevant to its previous utterances. There usually does not exist an obvious alignment between previous utterances and the response. Thus, neural generative models tend to generate safe and generic response, which contains little meaningful information, such as "*really?*", and "*I don't know.*". These generic responses are usually short and vague, and can seldom provide effective information as feedback to the other partner in the dialog. This has been identified as one of major challenges in neural response generation (Li et al., 2016).

To ameliorate the problem of generic responses, we design a model that can generate responses that are relevant to previous utterances. Two constraints are introduced, 1) topical constraint and 2) semantic constraint, to modify the objective function $P(Y|X)$. Responses which are highly related to their context usually have the same topic as the context does. Thus, the topical constraint conditions the model's decoding process on topic words in the context, which are identified by a topic word sequence labeler. Moreover, a relevant response should be semantically related to its context. The semantic constraint regularizes the model's predicted response by minimizing the semantic distance between the generated response and its context. Optimal transport (Peyré et al., 2019) is a method for finding a joint distribution of two distributions which minimizes the Wasserstein distance when converting one distribution

Proceedings of the 28th International Conference on Computational Linguistics, pages 4067–4077
Barcelona, Spain (Online), December 8-13, 2020

to another. It is applied to the computation of semantic distance in the proposed method. We use both automated evaluation and human evaluation to verify the effect of our proposed method.

2 Neural Response Generation Model

Serban et al. (2016) proposed Hierarchical Recurrent Encoder-Decoder (HRED) model, which improves the traditional seq2seq model by applying a hierarchical encoder. While its decoder keeps the same, HRED's encoder uses an utterance-level RNN and a context-level RNN to model the intra- and inter-utterance dependencies, respectively. Given K previous utterance $U = \{U_1, ..., U_K\}$ as input context, an utterance-level RNN first converts each utterance $U_k = \{w_{k,1}, ..., w_{k,N}\}$ into a fixed-length utterance vector $\mathbf{u}_k$, then the context-level RNN takes as input the sequence of utterance vectors $\mathbf{u}_1, ..., \mathbf{u}_K$ and produces a context vector $\mathbf{c}$ from its final hidden state and $\mathbf{c}$ is used to initialize the hidden states of HRED's decoder RNN. Although better performances can be achieved with attention mechanism, the problem of generic responses remains unsolved because the likelihood-based objective function Eq (1) still solely depends on the input context.

$$\begin{aligned}\hat{Y} &= \underset{Y}{\text{argmax}} P(Y|U_1, ..., U_K) \\ &= \underset{y_1,...,y_m}{\text{argmax}} \sum_{i=1}^{m} P(y_i|y_1, ..., y_{i-1}, U_1, ..., U_K)\end{aligned} \tag{1}$$

The HRED model is used as the baseline model in this work.

3 Topical and Semantic Constraints

To generate responses that are more relevant to the context, topical constraint and semantic constraint into the HRED are explored.

If a response and its context have the same topic, then the response is likely to be relevant to the context. Topic words are keywords which have more impact on following conversation, and they can dynamically define the topic of a conversation. To push the topic of a response more relevant to its context, we train a sequence labeler to recognize topic words $\{t_1, ..., t_n\}$ from a context and condition the decoding process on both the context and topic words. Then the objective function is modified as Eq (2):

$$\hat{Y} = \underset{Y}{\text{argmax}} P(Y|U_1, ..., U_K, T) \tag{2}$$

Here $T = \{t_1, ..., t_n\}$ indicates the topic words extracted from the context.

To further strengthen the semantic relevance between a generated response and its context, we try to close the representation of these two in the semantic space (Mikolov et al., 2013). We compute the semantic distance between a response and its context at the sequence level, and regularize the objective function with the distance as in Eq (3).

$$\hat{Y} = \underset{Y}{\text{argmax}} \{P(Y|U_1, ..., U_K) - \Delta(Emb(U), Emb(Y))\} \tag{3}$$

Here Δ is the function computing the semantic distance between the response and the context; $Emb()$ indicates the word embedding of a sequence of tokens.

Theoretically, by adding the topical and semantic constraints into the HRED, we can have more control on selecting generated content.

4 Incorporating Topical Constraint

To encourage topics of a response to be relevant to the topic of its context, we first pre-train a sequence labeler to extract topic words from the context, then jointly train it with the HRED. We use topic words to refer to a topic.

Here topic words are defined as words that 1) contain meaningful information and 2) are shared by the context and the response. In the example in Table 1, words "*sale*" and "*live*" contains useful information, among which "*live*" has a higher correlation with the response "*I have lived here for about twenty years.*" Therefore, it is reasonable to consider "*live*" as the topic word in this context-response pair.

Context	A: Good afternoon. I believe that this house is for **sale**. B: That's right. A: May I have a look at it please? B: Yes, of course. Come in. A: How long have you **lived** here ?
Response	B: I have lived here for about twenty years.

Table 1: An example of identifying topic words in a context.

In order to automatically identify such topic words in a context, a topic word dataset is constructed from a conversation corpus. This task is formulated as a sequence labeling problem, where each word in the context is labeled as either a topic word or a non-topic word.

4.1 Data Construction

Currently, there is no public data with topic word annotations, so the construction of such a dataset is necessary. According to the definition, a topic word can be filtered from context and a response by keywords, because keywords generally carry rich information. Then a keyword can be further filtered by checking whether the same keyword or a semantically related word appears in both the context and the response.

The training data comes from the *DailyDialog* corpus, which includes 13,118 open-domain human-human text conversations. The data is organized as {context, response} pairs. In the first step of keyword filtering, a public keyword extraction toolkit, YAKE (Campos et al., 2020), is used to extract keywords from both context and responses. These keywords are topic word candidates in the next step.

Since topic words are also expected to be shared by both the context and the response, two strategies are investigated to further filter the keywords extracted by YAKE.

- A hard matching strategy: Only keywords that appears in both a context and its corresponding response commonly are labeled as topic words. In this way, the context and response have the same topic words in the dataset.

- A soft matching strategy: Firstly, word embedding-based pairwise cosine similarity is computed between the context's keywords and the response's keywords, then the context-side words in the top 3 pairs are labeled as topic words. In this way, the context and response are expected to have relevant topics, rather than just the same keywords.

Table 2 shows the statistics of the original and the constructed dataset. These are used as ground-truth topic words, which should be detected in the context and used in the following response.

The number of tokens in the dataset constructed following hard structure is much less than the other dataset because of the tough matching requirements.

4.2 BERT-based Sequence Labeler

Sequence labeling is a pattern recognition task that assigns a categorical label to each components in a sequence. In the task of topic word labeling, a model takes as input a sequence of context tokens $X = \{x_1, ..., x_N\}$ and predicts a sequence of topic word labels $Y^{\text{topic}} = \{y_1^{\text{topic}}, ..., y_N^{\text{topic}}\}$, where $y_i^{\text{topic}} \in \{1, 0\}$ denotes whether word x_i is a topic word or not.

Bidirectional Encoder Representation from Transformers (BERT) (Devlin et al., 2019) is a Transformer-based model (Vaswani et al., 2017), which fully relies on attention mechanism. Many

Corpus	*DailyDialog*	*DailyDialog_topic*	
Topic Word Extraction	N/A	hard	soft
# tokens	600k	200k	500k
# topic words	N/A	7k	66k
# topic words/# tokens	N/A	3.5%	12.0%
# {context, response} pairs	60k	22k	52k

Table 2: Comparison of two data construction strategies: a hard matching strategy and a soft matching strategy.

works showed that by finetuning a pretrained BERT model, state-of-the-art results are achieved in sequence labeling tasks (Tsai et al., 2019; Shi and Lin, 2019). Thus, we also build the sequence labeler on the top of BERT.

BERT-based sequence labeler utilizes the output hidden states from BERT as the contextualized encoding of each word and uses a fully connected layer and a sigmoid function to compute the probability of each word being a topic word.

$$(\mathbf{h}_1^{\text{BERT}}, ..., \mathbf{h}_N^{\text{BERT}}) = \text{BERT}(x_1, ..., x_N) \tag{4}$$

$$p(x_i \text{ is topic word}) = \text{sigmoid}(\text{FC}(\mathbf{h}_i^{\text{BERT}})) \tag{5}$$

Cross entropy loss is used as the loss function for this model.

4.3 Response generation with topical constraint

To condition response generation on the extracted topic words as in Eq (2), a simple method is proposed to integrate the information from the HRED's encoder and the information from topic words' embeddings. Let $\mathbf{c}$ denote the context vector produced by the HRED's encoder, p_i denote $p(x_i$ is topic word), the decoder's hidden states are initialized with an updated $\mathbf{c}'$ that combines context information and topic word information. Here p_i is calculated by Eq (5).

$$\mathbf{c}' = \mathbf{c} + \text{FC}(\sum_i^N p_i * \text{Emb}(x_i)), \tag{6}$$

Here FC$(\cdot)$ is a fully-connected layer, Emb$(\cdot)$ is the embedding lookup table.

5 Incorporating Semantic Constraint

Another approach to generate more relevant responses is to consider the semantic distance between a generated response and its context. In some past works, the semantic distance was calculated by the similarity of the sequence-level semantic representations. However, it is possible to lose details in the process. We propose a novel method to calculate the semantic distance from a word-level representation.

Many word alignments between a response and the context may not be meaningful, such as an alignment of two irrelevant stop words. It is reasonable that we should value the alignments with strong semantic connection. Thus, a semantic distance considering the weights of different alignments is needed.

5.1 Optimal Transport

Assuming X and Y are two references belonging to the same space, an optimal transport (OT) problem can be described as following.

$$\begin{cases} P^* = \underset{P \in \mathcal{P}}{\text{argmin}} \sum_{i=1}^n \sum_{j=1}^m c(x_i, y_j) \cdot p_{ij} = \underset{P \in \mathcal{P}}{\text{argmin}} \langle P, C \rangle \\ \mathcal{P} = \{P \in \mathbb{R}_+^{n \times m} \mid P\mathbf{1}_m = X, P^T\mathbf{1}_n = Y\} \end{cases} \tag{7}$$

where $\mathcal{P}$ is a set of joint distributions, with marginal distribution X and Y, and each P is called a transport plan; $c(x_i, y_j)$ is a cost function for moving from x_i to y_j, C is the cost matrix given by $c(x_i, y_j)$; $\langle P, C\rangle$ represents a dot product. Therefore, an OT problem can be regarded as to find a transport plan which minimizes the OT distance - the weighted cost of moving from X to Y.

5.2 Decoding with Optimal Transport

In the OT problem formulated in Eq (7), the important components are transport plans and a cost matrix. We compute the OT distance of textual conversation data to provide sequence-level guidance to the decoding process as shown in Eq (3).

Many popular word embedding representations can be regarded as distributions over a semantic space. When applied to textual conversation data, values of the cost matrix can be the pairwise cosine distance between word embeddings of the context and the predicted response. For word embeddings which are close in semantic space, the cosine distance is correspondingly small, and the cost of moving from one embedding to another is small.

A transport plan, which is a joint distribution of the context word embedding and the predicted response word embedding, represents the global semantic matching between the context tokens and the response tokens. To minimize the OT distance, an optimal transport plan should assign a large weight to the token pairs, in which the cost is smaller to move from one word embedding representation to another.

Therefore, the dot product of an optimal transport plan P showing the degree of alignments and the cost matrix C showing the initial semantic distance can be regarded as a weighted semantic distance, where strongly relevant alignments are emphasized. The weighted semantic distance is calculated as follows:

$$\Delta(Emb(X), Emb(Y)) = \min_{P\in\mathcal{P}} \langle P, C\rangle \tag{8}$$

When matching words in a response with words in the context with optimal transport, as shown in Figure 1, optimal transport matching can not only capture the semantic relationship among the tokens but also the degree of the relationship.

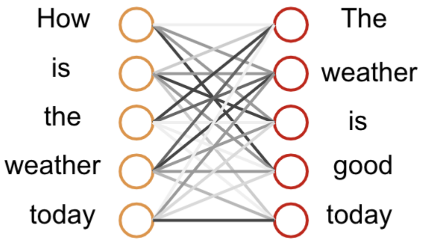

Figure 1: Match words using optimal transport between a response and its context. Darker lines indicate word alignments with stronger semantic connection.

By adding the factor OT distance into the objective function Eq (3), the model are expected to generate responses that are semantically similar to the context.

We solve the OT problem by approximating it with entropy-regularized OT as follows:

$$\mathcal{D}(X, Y) = \min_{P\in\mathcal{P}} \langle P, C\rangle - \epsilon H(P) \tag{9}$$

where $\mathcal{D}$ is the OT distance; entropy $H(P)$ is computed as follows:

$$H(P) = -\sum_{ij} P_{ij}(\log(P_{ij}) - 1) \tag{10}$$

The entropy-regularized OT is a strongly convex function, thus it has a unique solution. Theoretically, the solution of the approximated OT, P_ϵ, convergence to the optimal solution of OT as ϵ approaches to 0 (Genevay, 2019).

Sinkhorn iteration (Cuturi, 2013) is an efficient algorithm for entropy-regularized OT. The algorithm is as follows.

```
Algorithm 1: Sinkhorn interation for entropy-regularized OT
Data: cost matrix $C$, regularization coefficient $\epsilon$, maximum iteration number max_iter
Result: $P_\epsilon$
initialize $\mu$, $\nu$ as uniform distribution
kernal matrix $K \leftarrow C$
while iteration number < max_iter do
    $u^{k+1} = \frac{\mu}{Kv^{(k)}}$
    $v^{k+1} = \frac{\nu}{Ku^{(k+1)}}$
$P_\epsilon$ = diag($u$)$K$diag($v$)
```

6 Experimental Evaluations

To investigate the impact of the datasets constructed in Section 4 on the topic word labeler, the sequence labeler is trained and tested on the two different datasets. To confirm the effect of the proposed methods in improving the response quality, experiments are conducted using the *DailyDialog* corpus.

6.1 Experiments on Bert-based Sequence Labeler

The performance of the sequence labeler is evaluated by automatic metrics of precision, recall and F1 score. Taking the imbalance of data into consideration, these metric scores are calculated under different conditions as follows:

- Binary: Calculating scores regarding topic words.
- Macro: Calculating scores regarding topic words and non-topic words separately, then calculating the unweighted average of the scores of two labels.
- Coverage: Calculating scores regarding topic words without repeated tokens.

The results for sequence labeler trained on two different datasets is shown in Table 3. It shows that precision and F1 scores are higher when the model is trained with *DailyDialog_topic_soft*, and it is reasonable since the ratio of topic words is much higher in *DailyDialog_topic_soft* than that of *DailyDialog_topic_hard* as in Table 2.

Data construction method	Condition	Precision	Recall	F1
DailyDialog_topic_hard (hard matching)	Binary	0.32	0.71	0.45
	Macro	0.66	0.84	0.71
	Coverage	0.34	0.71	0.44
DailyDialog_topic_soft (soft matching)	Binary	0.60	0.60	0.60
	Macro	0.79	0.79	0.79
	Coverage	0.62	0.63	0.62

Table 3: Evaluation of the topic word labeler regarding different data construction methods.

6.2 Experiments on Response Generation

6.2.1 Competing models

The following models are assessed by automatic metrics and human evaluation.

HRED. The HRED model with a hierarchical context encoder and a response decoder. This HRED model uses a multi-head attention layer with the head size of 4.

HRED-Topic. The HRED model with a sequence labeler on topic words. There are two sequence labelers trained on different datasets, which are introduced in Section 4 and evaluated in Section 6.1. It has two variations.

- **HRED-Topic-Hard.** This is trained with *DailyDialog_topic_hard*.
- **HRED-Topic-Soft.** This is trained with *DailyDialog_topic_soft*.

HRED-OT. The HRED model using the weighted semantic distance between the context and the response to regularize the objective function, where the semantic distance is calculated by optimal transport. The regularization coefficient of OT is 0.5, and the number of the maximal iteration is 50.

HRED-Topic+OT. The HRED model with a sequence labeler and an optimal transport layer. The sequence labeler is trained on *DailyDialog_topic_hard*.

6.2.2 Automatic Evaluation

The models are assessed by the following metrics.

BLEU-2 calculates the unigram and bigram overlap of the predicted response and the ground truth response, then takes the average of them (Papineni et al., 2002).

Embedding similarity calculates the cosine similarity between a predicted response and the ground truth response. There are different ways of choosing representation: average, extrema, greedy, denoted by Emb-A, Emb-E, Emb-G (Foltz et al., 1998; Forgues et al., 2014; Rus and Lintean, 2012).

Distinct calculates the ratio of unique unigram and bigram entries in a predicted response, denoted by Dist-1 and Dist-2. Distinct scores are calculated regarding both the response itself and all responses in the corpus, denoted by Intra Dist and Inter Dist (Li et al., 2016). Distinct scores discover the diversity of the predicted response.

Table 4 shows the automatic evaluation results of all models. All proposed models outperform the baseline model HRED in terms of the BLEU-2 score, and HRED-Topic-Hard even surpasses the baseline model by one point in the BLEU-2 score. When compared with embedding similarity, HRED-OT performs slightly better than the others, showing that OT has a positive impact on reducing the semantic distance of a response and its context. HRED-Topic-Hard outperforms HRED-Topic-Soft in the experiment. One possible reason is that by following the soft construction strategy, data pairs where only weak connection between a response and its context exists are remained in the dataset. In the later parts, HRED-Topic-Soft is not used here after.

Model	BLEU-2	Emb-A	Emb-E	Emb-G	Inter Dist-1	Inter Dist-2	Intra Dist-1	Intra Dist-2
HRED (baseline)	5.82	0.92	0.52	0.77	0.029	0.187	0.905	0.971
HRED-Topic								
HRED-Topic-Hard	6.83	0.92	0.52	0.77	0.028	0.180	0.907	0.973
HRED-Topic-Soft	6.52	0.91	0.52	0.76	0.026	0.166	0.898	0.954
HRED-OT	6.03	0.92	0.54	0.77	0.030	0.195	0.907	0.961
HRED-Topic+OT	5.93	0.91	0.51	0.76	0.025	0.114	0.888	0.951

Table 4: Automated evaluation results.

6.2.3 Human Evaluation

The proposed methods are expected to generate topic-relevant responses, which are not evaluated only with automatic evaluation. Hence, human evaluation is also conducted to assess the proposed methods. One hundred conversation samples are prepared for human evaluation. Each sample includes responses from all models, which is evaluated by 3 annotators from Amazon Mechanical Turk. A topic-relevant response needs to address two aspects of readability and relevance. Hence, responses from each model are scored on two categories.

- Grammar: $1 \sim 3$ (from unreadable to fluent).
- Relevance: $1 \sim 3$ (from irrelevant to closely relevant).

It is worthy investigating the proportion of acceptable responses and high-quality responses. Two additional criteria are defined as follows.

- Acceptance rate: the proportion of responses whose grammar and relevance are scored no less than 2.
- High quality rate: the proportion of responses whose grammar and relevance are scored as 3.

Model	**Grammar**	**Relevance**	**Acceptance Rate**	**High Quality Rate**
HRED	2.87	2.26	0.89	0.32
HRED-Topic	2.87	2.18	0.89	0.28
HRED-OT	2.82	2.14	0.90	0.20
HRED-Topic+OT	2.91	2.27	0.94	0.30

Table 5: Human evaluation results of all response samples.

From Table 5, we observe that the proposed model incorporating both constraints surpass the others in both grammar and relevance scores. This integrated model generates more acceptable response but less high quality response, compared to the baseline model.

Generic responses can be reasonable regarding a lot of contexts, and they're less likely to make grammar mistakes. Looking through human evaluation examples, it is found that some of generic responses are also scored as high quality response. Hence, it is necessary to calculate the scores within non-generic and content-rich responses. A generic response is typically short, resulting in higher grammar scores. It is reasonable to think that a longer response is more content-rich than a short response, and it is less likely to be a generic response. We define content-rich responses as responses that are longer than the average length of human responses in the corpus, that is 13. The ratio of content-rich responses of all models are shown in Table 6.

Model	Ratio of content-rich responses
HRED	26%
HRED-Topic	35%
HRED-OT	26%
HRED-Topic+OT	30%

Table 6: Ratio of content-rich responses (that are no less than 13 words)

Human evaluation results are shown in Table 7 by focusing on these content-rich responses, we observe that the integrated model has increment in both acceptance rate and high quality rate, while the baseline model decrement in both. The results shows that the model integrating both proposed constraints can generate more relevant and content-rich responses.

7 Related Work

Our proposed work is related to some past works. To generate more content-rich responses, a seq2seq model with an additional topic and semantic constraint that are based on a topic model (Griffiths et al., 2005) and cosine similarity (Arora et al., 2016) has been proposed, showing promising results in the content richness of generated responses (Baheti et al., 2018). Xing et al. also used topic information as side information and showed good results, where the topics are obtained from a pre-trained LDA

Model	Grammar	Relevance	Acceptance Rate	High Quality Rate
HRED	2.77	2.23	0.88	0.31
HRED-Topic	2.69	2.11	0.83	0.26
HRED-OT	2.65	2.23	0.92	0.19
HRED-Topic+OT	2.90	2.40	0.97	0.37

Table 7: Human evaluation results of content-rich response samples.

model (Xing et al., 2017). To improve the model structure, Mou el al. proposed a model that generates responses backward and forward starting from a keyword, which is a noun with the highest pointwise mutual information (PMI) score as a hard constraint, outperforming traditional Sequence-to-Sequence model (Mou et al., 2016). Zhang et al. built their system for ”relevant, contentful and context-consistent responses”(Zhang et al., 2019); Adiwardana et al. built a multi-turn open-domain chatbot and proposed a metric called Sensibleness and Specificity Average (SSA) for human evaluation (Adiwardana et al., 2020); Smith et al. wanted their chatbot to be engaging, knowledgeable, and empathetic and proposed a new dataset called BlendedSkillTalk for analysis (Smith et al., 2020); Roller et al. evaluated their multi-turn dialogue model for engagingness and humanness (Roller et al., 2020).

8 Conclusion

A topic-relevant response should have a common or similar topic with its context, and it should be semantically related to the context. To generate topic-relevant responses and avoid generic responses like “I don’t know”, two constraints placed onto the decoding process are proposed.

The first constraint is a topical constraint. To extract topic information from the context, a sequence labeler is trained on two differently constructed dataset with topic word annotations. The topic word information predicted by the sequence labeler is then integrated into the context vector generated by the hierarchical encoder to guide the decoding process.

The second constraint is a semantic constraint. A model is designed to generate responses semantically related to context by adding the semantic distance is added into the global loss of the model. In this work, semantic distance is calculated by optimal transport, which gives the optimal alignments between context and its response in a semantic space. The calculated semantic distance is used as a regularization for the objective function, giving sequence-level guidance on the decoding process.

The experimental evaluations demonstrated that combining these constraints lead to responses that are content-rich and with higher relevance to the context while maintaining good performance in the grammar for long sentences. This work has some distinct features in 1) using topic words to dynamically refer to topics; 2) providing with semantic sequence-level guidance, but with access to word-level similarity values.

References

Daniel Adiwardana, Minh-Thang Luong, David R So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, et al. 2020. Towards a human-like open-domain chatbot. *arXiv preprint arXiv:2001.09977*.

Sanjeev Arora, Yingyu Liang, and Tengyu Ma. 2016. A simple but tough-to-beat baseline for sentence embeddings.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Ashutosh Baheti, Alan Ritter, Jiwei Li, and Bill Dolan. 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, Brussels, Belgium, October-November. Association for Computational Linguistics.

Ricardo Campos, Vítor Mangaravite, Arian Pasquali, Alípio Jorge, Célia Nunes, and Adam Jatowt. 2020. Yake! keyword extraction from single documents using multiple local features. *Information Sciences*, 509:257–289.

Marco Cuturi. 2013. Sinkhorn distances: Lightspeed computation of optimal transport. In C. J. C. Burges, L. Bottou, M. Welling, Z. Ghahramani, and K. Q. Weinberger, editors, *Advances in Neural Information Processing Systems 26*, pages 2292–2300. Curran Associates, Inc.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *Proceedings of the 2019 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long and Short Papers)*, pages 4171–4186.

Peter W Foltz, Walter Kintsch, and Thomas K Landauer. 1998. The measurement of textual coherence with latent semantic analysis. *Discourse processes*, 25(2-3):285–307.

Gabriel Forgues, Joelle Pineau, Jean-Marie Larchevêque, and Réal Tremblay. 2014. Bootstrapping dialog systems with word embeddings. In *Nips, modern machine learning and natural language processing workshop*, volume 2.

Aude Genevay. 2019. *Entropy-regularized optimal transport for machine learning*. Ph.D. thesis.

Thomas L Griffiths, Mark Steyvers, David M Blei, and Joshua B Tenenbaum. 2005. Integrating topics and syntax. In *Advances in neural information processing systems*, pages 537–544.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016. A diversity-promoting objective function for neural conversation models. In *Proceedings of NAACL-HLT*, pages 110–119.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in neural information processing systems*, pages 3111–3119.

Lili Mou, Yiping Song, Rui Yan, Ge Li, Lu Zhang, and Zhi Jin. 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.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: a method for automatic evaluation of machine translation. In *Proceedings of the 40th annual meeting on association for computational linguistics*, pages 311–318. Association for Computational Linguistics.

Gabriel Peyré, Marco Cuturi, et al. 2019. Computational optimal transport. *Foundations and Trends® in Machine Learning*, 11(5-6):355–607.

Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M Smith, et al. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637*.

Vasile Rus and Mihai Lintean. 2012. An optimal assessment of natural language student input using word-to-word similarity metrics. In *International Conference on Intelligent Tutoring Systems*, pages 675–676. Springer.

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 *Thirtieth AAAI Conference on Artificial Intelligence*.

Peng Shi and Jimmy Lin. 2019. Simple bert models for relation extraction and semantic role labeling. *arXiv preprint arXiv:1904.05255*.

Eric Michael Smith, Mary Williamson, Kurt Shuster, Jason Weston, and Y-Lan Boureau. 2020. Can you put it all together: Evaluating conversational agents' ability to blend skills. *arXiv preprint arXiv:2004.08449*.

Henry Tsai, Jason Riesa, Melvin Johnson, Naveen Arivazhagan, Xin Li, and Amelia Archer. 2019. Small and practical bert models for sequence labeling. In *Proceedings of the 2019 Conference on Empirical Methods in Natural Language Processing and the 9th International Joint Conference on Natural Language Processing (EMNLP-IJCNLP)*, pages 3623–3627.

Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N Gomez, Łukasz Kaiser, and Illia Polosukhin. 2017. Attention is all you need. In *Advances in neural information processing systems*, pages 5998–6008.

Chen Xing, Wei Wu, Yu Wu, Jie Liu, Yalou Huang, Ming Zhou, and Wei-Ying Ma. 2017. Topic aware neural response generation. In *Thirty-First AAAI Conference on Artificial Intelligence*.

Yizhe Zhang, Siqi Sun, Michel Galley, Yen-Chun Chen, Chris Brockett, Xiang Gao, Jianfeng Gao, Jingjing Liu, and Bill Dolan. 2019. Dialogpt: Large-scale generative pre-training for conversational response generation. *arXiv preprint arXiv:1911.00536*.