Naturalness Evaluation of Natural Language Generation in Task-oriented Dialogues using BERT

Ye Liu[1]**, Wolfgang Maier**[2]**, Wolfgang Minker**[3] and **Stefan Ultes**[4]

[1,2,4]Mercedes-Benz AG, Sindelfingen, Germany
[1,3]Ulm University, Ulm, Germany
[1]`ye.y.liu@daimler.com`
[2,4]`{wolfgang.mw.maier,stefan.ultes}@daimler.com`
[3]`wolfgang.minker@uni-ulm.de`

Abstract

This paper presents an automatic method to evaluate the naturalness of natural language generation in dialogue systems. While this task was previously rendered through expensive and time-consuming human labor, we present this novel task of automatic naturalness evaluation of generated language. By fine-tuning the BERT model, our proposed naturalness evaluation method shows robust results and outperforms the baselines: support vector machines, bi-directional LSTMs, and BLEURT. In addition, the training speed and evaluation performance of naturalness model are improved by transfer learning from quality and informativeness linguistic knowledge.

1 Introduction

With the increasing popularity of virtual assistants such as Alexa or Siri, users have a higher demand for conducting natural conversations. They would like to chat with these assistants more naturally—maybe even like talking to a real human being. One of the key questions arising from this, though, is how to measure the naturalness of the generated language. In the past, native speakers judged the quality of the generated language by answering questions like "is this utterance natural?" or "could it have been produced by a native speaker?" to rate the naturalness (Novikova et al., 2016). However, this approach heavily depends on manual effort and is impractical for broader use. On the other hand, the widely used automatic metrics for evaluating language generation, like BLEU (Papineni et al., 2002) and METEOR (Banerjee and Lavie, 2005), rely on word overlap mechanism and compare the generated sentence to one or more human-created reference sentences (Stent et al., 2005), but cannot directly reflect naturalness information.

Likert scale ratings are generally used in human evaluation. The Figure 1 shows an example of

Sentence: X is a moderately priced restaurant in X.

Figure 1: Example of 6-point Likert scale for evaluating the naturalness of a sentence.

6-point Likert scale rating on naturalness of a sentence. In general, the human evaluation items (naturalness, coherence, etc.) are hard explainable objectively and native speakers purely rely on their underlying criteria to judge the performance of generated language without any specified rules. To alleviate the manual effort previously needed, the goal of this paper is to render the task of assessing the naturalness of generated language as an automatic reference-based classification problem having each Likert scale rating representing one class. To our knowledge, we are the first to propose this task and present an approach evaluating naturalness of generation through fine-tuning a pre-trained Language Model (LM).

The contribution of this paper is two-fold: a pre-trained BERT model is fine-tuned to estimate the naturalness of a generated sentence based on a reference sentence. Three baselines are proposed for robust performance comparison: a support vector machine (SVM) baseline using bag-of-words (BoW) vectors for input representation, a bi-directional Long Short Term Memory (LSTM) neural network, and the pre-trained BLEURT model (Sellam et al., 2020). Second, based on the positive correlation between naturalness and other annotated information: quality and informativeness of the generated sentence, the proposed method is extended to leverage this additional information through transfer learning. And the learning speed and estimation performance of naturalness evaluation model is significantly increased.

The remainder of this paper is structured as fol-

lows: Section 2 introduces the related works. In Section 3, our proposed approach for BERT-based naturalness estimation is presented. In Section 4, the experiment setups are covered, which include fine-tuning BERT, comparison against the baselines, and transfer learning. Section 5 describes the experiments results. The last Section 6 draws conclusions and outlines future work.

2 Related Work

With the development of Natural Language Generation (NLG) applications and their benchmark datasets, evaluation of NLG systems has become increasingly important. Generally, multiple automatic metrics are used in parallel to evaluate the performance of language generation, such as BLEU (Papineni et al., 2002), METEOR (Banerjee and Lavie, 2005) or ROUGE (Lin, 2004). However, Chaganty et al. (2018) demonstrated that the existing automatic metrics have poor instance-level correlation with mean human judgment and that they assign bad scores to many good quality responses.

Since automatic metrics still fall short of replicating human decisions (Krahmer and Theune, 2010; Reiter, 2018), many NLG papers include some form of human evaluation. For example, Hashimoto et al. (2019) report that 20 out of 26 generation papers published at ACL2018 presented human evaluation results for showing their robust performance comparison. Celikyilmaz et al. (2020) and Gatt and Krahmer (2018) also highlighted that human evaluation is commonly viewed as the best reliable way to evaluate NLG systems, but come with many issues, such as costly and time consuming and human judgement bias. And more importantly, the evaluation results from human efforts are not always repeatable (Belz and Reiter, 2006). Dusek et al. (2017) previously attempted to predict quality ratings of generated language by using Recurrent Neural Network (RNN) with the help of the Meaning Representations (MRs) and showed promising performance. However, our work is more focused on naturalness evaluation and based on the gold reference. In order to relieve the human labor, we propose a reference-based method for naturalness evaluation by utilizing neural network to learn the complicated linguistic relationship from sentences. And this work can be easily extended to other human evaluation criteria (coherence, quality, etc..), if the corresponding human evaluation data is available.

In recent years, with huge success of pre-trained LMs (Devlin et al., 2019; Radford et al., 2018), many machine learned metrics for evaluating generation are proposed. Especially the pre-trained BERT (Bidirectional Encoder Representations from Transformers) (Devlin et al., 2019) shows its superiority in this field. Shimanaka et al. (2019) presented automatic machine translation evaluation by using BERT and achieved the best performance in segment-level metrics tasks on the WMT17 dataset for all to-English language pairs. Zhang et al. (2019) proposed an automatic evaluation metric for text generation based on pre-trained BERT contextual embeddings: BERTScore. BERTScore computes the similarity of two sentences as a sum of cosine similarities between their tokens' embeddings. And Zhang et al. (2019) showed BERTScore correlates better with human judgments and provides stronger model selection performance than existing metrics. Sellam et al. (2020) presented BLEURT, which continually pretrained BERT on synthetic data and then fine-tuned on task-specific ratings. And Sellam et al. (2020) demonstrated BLEURT can model human assessment with superior accuracy. Given the superiority of BERT, we also apply the pre-trained BERT for our proposed naturalness evaluation on generated language.

3 Naturalness Evaluation Using BERT

The task of estimation the naturalness of a generated sentence is framed as a classification task. Two sentences are used as input: the candidate sentence to be scored and a reference sentence. The naturalness score is derived through fine-tuning a pre-trained BERT (Devlin et al., 2019) model. The main architecture of BERT uses the encoder of a Transformer (Vaswani et al., 2017), which is an advanced encoder-decoder architecture leveraging the attention mechanism. Considering that the input may be sentence pairs in several tasks, technical innovation Next Sentence Prediction (NSP) helps BERT to learn the relationship between sentence pairs by receiving pairs of sentences as input and separating them with [SEP] token to learn predicting if the second sentence is the subsequent sentence in the original document. To do so, a [CLS] token is added at the beginning for every input to learn the meaning of the entire input. Exactly because of these specific characters, eleven NLP tasks

		judge	**naturalness**	quality	informativeness
sys_ref:	*zuni cafe, is expensive.*	1	**6**	5	6
orig_ref:	*how about zuni cafe,*	2	**5**	4	6
	an expensive one?	3	**6**	4	5

Table 1: Example of the pre-processed data set.

naturalness	1	2	3	4	5	6
data size	394	373	670	2,185	3,062	4,669
total				11,353		

Table 2: Distribution of naturalness labels of the data set with the majority label marked red.

in Devlin et al. (2019) obtained new state-of-the-art results by fine-tuning BERT.

Figure 2 shows the fine-tuning structure for naturalness estimation with a candidate sentence (sys_ref) and a reference sentence (orig_ref) as input. Example for a sys_ref and orig_ref is shown in Table 1. In accordance to the NSP task, representing both sentences on the input side is achieved by separating them with the [SEP] token. As can be seen in Figure 2, an additional [CLS] token is inserted at the beginning of the first sentence. The final output of the [CLS] token, which is called pooled_output, forms a representation of the entire input. Then a linear layer with softmax activation is added to the top of pooled_output to predict the probability of sentence-level naturalness label. To our knowledge, we are the first to fine-tune BERT to learn the abstract naturalness linguistic information and demonstrate the robust performance of this method.

4 Experimental Setup

In this section, the experiment procedure including the pre-processing of the used data is introduced. First, BERT is fine-tuned for naturalness estimation and then compared it with the baselines. Furthermore, external knowledge is added demonstrating the impact on naturalness estimation through transfer learning.

4.1 Dataset Preprocessing

The dataset[1] (Novikova et al., 2017) comprises textual dialog responses produced by three data-driven NLG systems over three different domains. The three NLG systems are respectively **RNNLG**[2], **TGen**[3] and **LOLS**[4]. The applied domains are **SF Hotel** and **SF Restaurant** (Wen et al., 2015),

---

[1] https://researchportal.hw.ac.uk/en/datasets/human-ratings-of-natural-language-generation-outputs
[2] https://github.com/shawnwun/RNNLG
[3] https://github.com/UFAL-DSG/tgen
[4] https://github.com/glampouras/JLOLS

which provide information about hotels and restaurants in San Francisco, and **BAGEL** (Mairesse et al., 2010) that provides information about restaurants in Cambridge. Human annotations on naturalness, quality, and informativeness were collected for each NLG-produced text on a 6-point Likert scale by asking the annotator "could the utterance have been produced by a native speaker?", "How do you judge the overall quality of the utterance in terms of its grammatical correctness and fluency?", "Does the utterance provide all the useful information from the meaning representation?", respectively.

Table 1 shows an annotation example of the data. The sentence pair input comprises sys_ref and orig_ref. The sys_ref presents the output of each of the above-mentioned three NLG systems, one at a time, while orig_ref denotes human written references from the original data. The judges (1, 2, 3) represent the three human raters. The naturalness score is our target label. In addition to the naturalness score, the informativeness score and the quality score are utilised in transfer learning experiments to improve naturalness estimation as introduced in Section 4.4. To derive a single label for each sentence pair, the median of the individual annotations is used. The Table 2 shows the distribution of the final processed data, which includes $11,353$ samples. And we randomly split the data into train/dev/test with 80%/10%/10%.

4.2 Fine-tuning BERT

During fine-tuning, the entire pre-trained BERT model is optimised end-to-end. The output of token [CLS]: the pooled_output, is further fed to a linear layer with softmax activation function with parameters $W \in \mathbb{R}^{K \times H}$, where H is the dimension of the hidden state vectors and K is the number of classes. In this paper, we applied English uncased BERT-Base model [5], which has 12 layers, 768 hidden states and 12 heads, for the naturalness classification. So H is 768 and K is 6 in our experiment.

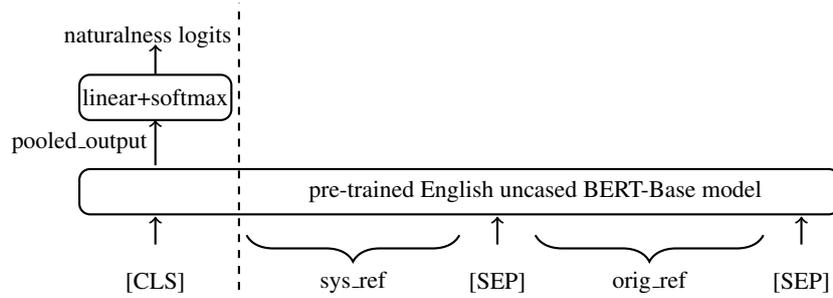[5] https://tfhub.dev/google/bert_uncased_L-12_H-768_A-12/1

Figure 2: Fine-tuning BERT architecture for naturalness estimation

All hyper-parameters are tuned to our dataset. The batch size is 256 and the number of epochs is 25. Adam (Kingma and Ba, 2014) is used for optimization to minimize the cross-entropy with an initial learning rate of 5e-3.

4.3 Baselines

As this is a novel task, there is no available existing baseline for us to compare. Hence, we apply the following three baselines for performance comparison in order to show the robustness of our proposed method.

BoW + SVM: We firstly introduce a SVM classifier using BoW representation as baseline. The SVM (Suykens and Vandewalle, 1999) is a discriminative classifier formally defined by a separating hyperplane and is widely used for classification task because of significant accuracy with less computation power. The BoW model (Zhang et al., 2010) is a text representation that counts the occurrence of words within a document. The BoW approach is very simple and flexible, and can be used for extracting word features from documents. These numerical BoW vector features are used as input to a SVM with linear kernel having hyper-parameters $C = 1.0$ and γ set to 'auto'.

Bi-LSTM: We also introduce the bidirectional LSTM (Bi-LSTM) with one layer for naturalness evaluation as baseline. The Bi-LSTM layer has the same hidden size as fine-tuning BERT, i.e., 768, and the output is forward to one linear layer with softmax function for naturalness classification. We almost remain the same hyper-parameters setting as BERT for Bi-LSTM training, i.e., 256 batch size, 25 epochs and 5e-3 initial learning rate. We restrict the number of LSTM-layers to one because multiple layers resulted in very slow training speed, as the LSTM cannot be trained in parallel, and worse performance.

BLEURT: In order to establish robust performance comparison, we also apply a pre-trained model for naturalness evaluation as the third baseline: BLEURT (Sellam et al., 2020), which is a machine learned automatic metric for text generation. BLEURT continually pre-trained BERT (Devlin et al., 2019) with a large number of synthetic reference-candidate pairs on several lexical-and semantic-level supervision tasks and then fine-tuned on multiple human ratings. Sellam et al. (2020) published multiple versions of BLEURT in the official repository[6], which includes BLEURT-tiny, BLEURT-base and BLEURT-large. More information could be found in the link[7]. (Sellam et al., 2020) demonstrated that BLEURT is much closer to human annotation and also recommended to fine-tune the pre-trained BLEURT for custom applications. Hence, we apply the BLEURT-tiny in this work for naturalness classification by adding one additional linear layer with softmax function. We also tried the recommended BLEURT-base model for naturalness evaluation, however, it directly resulted in worse performance in our case.

4.4 Fine-tuning BERT plus Transfer Learning

Analysing the correlation of the naturalness scores with the respective quality and informative scores using Spearman rank correlation coefficient (Hauke and Kossowski, 2011) shows the positive results. The spearman's correlation between naturalness and quality is $\rho = 0.60$ and between naturalness and informativeness is $\rho = 0.45$. Hence, this positive correlation is further leveraged through transfer learning (Pan and Yang, 2009) using the same BERT-based setup. The procedure is as follows: first, the BERT model is fine-tuned to quality (or in-

---

[6]https://github.com/google-research/bleurt
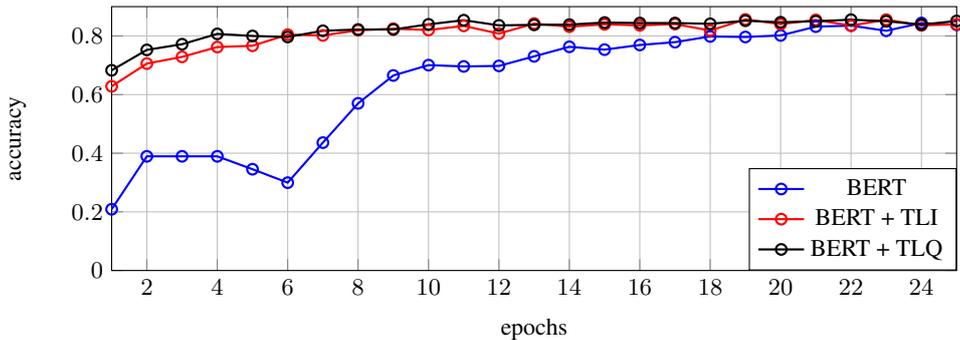[7]https://github.com/google-research/bleurt/blob/master/checkpoints.md

Figure 3: Comparison of the different approaches to naturalness estimation with respect to the training epochs. With additional transfer learning using the quality score (BERT + TLQ), the training speed is increased the most.

formativeness, respectively), and then this already fine-tuned model is continually fine-tuned once more using the naturalness score as target.

5 Results and Discussion

The results of our proposed approach (BERT) for estimating the naturalness of a generated sentence given an additional reference are depicted in Table 3. In addition to the baselines: BoW + SVM and Bi-LSTM and BLEURT, the majority class accuracy is also shown. It is calculated as the proportion of the majority class (naturalness score 6) resulting in $4669/11353 = 0.41$. BERT + TLI and BERT + TLQ represent the Transfer Learning results from Informativeness (TLI) and Quality (TLQ) respectively.

The results in Table 3 show that all BERT-based approaches outperform the baselines for classifying the naturalness of a generated sentence achieving a higher overall accuracy.

Even though the data set is imbalanced as shown in Table 2, Madabushi et al. (2019) indicate that BERT is capable of handling imbalanced classes with no additional data augmentation, which is also confirmed in our work. The Table 4 shows the accuracy of different naturalness score on our proposed model BERT with test data. Even if the data is seriously imbalanced, every naturalness class has comparative accuracy.

Given the imbalanced data set, the macro F1_score, recall and precision are also computed to show the robustness of our proposed approach. Moreover, through the transfer of quality (or informativeness) knowledge to naturalness training, the performance of naturalness estimation is further improved and training speed has also been greatly promoted. Figure 3 shows that the naturalness training based on transfer learning is faster and tends

to be stable after only 5 epochs. Table 3 shows that transferring knowledge from quality results in the highest improvement on naturalness estimation. This is also consistent with the Spearman correlation of the naturalness scores and quality scores which is higher than the correlation of the naturalness scores and informativeness scores.

The Table 3 shows that using the BLEURT model for naturalness evaluation results in the worst performance even though BLEURT was already pre-trained on multiple tasks. The possible reason is that the BLEURT was pre-trained with multiple automatic metrics, hence, it has no superiority in our naturalness classification task.

6 Conclusion and Future Work

In this paper, we proposed a novel task of automatically estimating the naturalness for task-oriented generation based on a human reference. We proposed a robust estimation approach by fine-tuning the pre-trained BERT model which outperforms an SVM classifier, Bi-LSTM, fine-tuned BLEURT and majority class accuracy. Taking advantage of the positive correlation of naturalness on quality (or informativeness), we successfully improved the naturalness training speed and estimation performance through transfer learning.

This work sheds light on research towards naturalness evaluation by neural network learning. The final goal of our work is to relieve the human labors from naturalness evaluation task and realize the automatic naturalness evaluation for dialogue generation. Hence, we will firstly collect more human evaluation data for future work. Because the human evaluation data, which is already shared and public on the internet, is very limited. With more collected human evaluation data, we are also interested in the performance of our proposed method on other

	majority class	BLEURT	BoW + SVM	Bi-LSTM	BERT	BERT + TLI	BERT + TLQ
F1_score	-	0.13	0.66	0.69	0.83	0.84	0.86
recall	-	0.22	0.66	0.66	0.83	0.83	0.84
precision	-	0.18	0.67	0.73	0.82	0.85	0.89
accuracy	0.41	0.42	0.68	0.74	0.85	0.87	0.88

Table 3: Performance comparison of different methods shows the superiority of our proposed BERT on naturalness evaluation.

naturalness	1	2	3	4	5	6
test size	41	35	65	227	305	462
prediction size	37	27	46	180	240	441
accuracy	0.90	0.77	0.71	0.79	0.79	0.95

Table 4: Accuracy of different naturalness scores on BERT with test data.

evaluation criteria, such as quality, coherence etc. And we will further verify the performance of our proposed method on chit-chat and open domain dialogue generation.

References

Satanjeev Banerjee and Alon Lavie. 2005. Meteor: An automatic metric for mt evaluation with improved correlation with human judgments. *Intrinsic and Extrinsic Evaluation Measures for Machine Translation and/or Summarization*, page 65.

Anja Belz and Ehud Reiter. 2006. Comparing automatic and human evaluation of nlg systems. In *11th Conference of the European Chapter of the Association for Computational Linguistics*.

Asli Celikyilmaz, Elizabeth Clark, and Jianfeng Gao. 2020. Evaluation of text generation: A survey. *arXiv preprint arXiv:2006.14799*.

Arun Tejasvi Chaganty, Stephen Mussmann, and Percy Liang. 2018. The price of debiasing automatic metrics in natural language evalaution. In *ACL (1)*.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, and Kristina Toutanova. 2019. Bert: Pre-training of deep bidirectional transformers for language understanding. In *NAACL-HLT (1)*.

Ondrej Dusek, Jekaterina Novikova, and Verena Rieser. 2017. Referenceless quality estimation for natural language generation. In *1st Workshop on Learning to Generate Natural Language*. ICML.

Albert Gatt and Emiel Krahmer. 2018. Survey of the state of the art in natural language generation: Core tasks, applications and evaluation. *Journal of Artificial Intelligence Research*, 61:65–170.

Tatsunori Hashimoto, Hugh Zhang, and Percy Liang. 2019. Unifying human and statistical evaluation for natural language generation. 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 1689–1701.

Jan Hauke and Tomasz Kossowski. 2011. Comparison of values of pearson's and spearman's correlation coefficients on the same sets of data. *Quaestiones geographicae*, 30(2):87–93.

Diederik P Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*.

Emiel Krahmer and Mariët Theune. 2010. *Empirical methods in natural language generation: Data-oriented methods and empirical evaluation*, volume 5790. Springer.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text summarization branches out*, pages 74–81.

Harish Tayyar Madabushi, Elena Kochkina, and Michael Castelle. 2019. Cost-sensitive bert for generalisable sentence classification on imbalanced data. In *Proceedings of the Second Workshop on Natural Language Processing for Internet Freedom: Censorship, Disinformation, and Propaganda*, pages 125–134.

François Mairesse, Milica Gašić, Filip Jurčíček, Simon Keizer, Blaise Thomson, Kai Yu, and Steve Young. 2010. Phrase-based statistical language generation using graphical models and active learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*, pages 1552–1561. Association for Computational Linguistics.

Jekaterina Novikova, Ondřej Dušek, Amanda Cercas Curry, and Verena Rieser. 2017. Why we need new evaluation metrics for nlg. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2241–2252.

Jekaterina Novikova, Oliver Lemon, and Verena Rieser. 2016. Crowd-sourcing nlg data: Pictures elicit better data. In *9th International Natural Language Generation Conference*, pages 265–273. Association for Computational Linguistics.

Sinno Jialin Pan and Qiang Yang. 2009. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359.

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.

Alec Radford, Karthik Narasimhan, Tim Salimans, and Ilya Sutskever. 2018. Improving language understanding by generative pre-training. *URL https://s3-us-west-2. amazonaws. com/openai-assets/researchcovers/languageunsupervised/language understanding paper. pdf.*

Ehud Reiter. 2018. A structured review of the validity of bleu. *Computational Linguistics*, 44(3):393–401.

Thibault Sellam, Dipanjan Das, and Ankur Parikh. 2020. Bleurt: Learning robust metrics for text generation. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, pages 7881–7892.

Hiroki Shimanaka, Tomoyuki Kajiwara, and Mamoru Komachi. 2019. Machine translation evaluation with bert regressor. *arXiv e-prints*, pages arXiv–1907.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *international conference on intelligent text processing and computational linguistics*, pages 341–351. Springer.

Johan AK Suykens and Joos Vandewalle. 1999. Least squares support vector machine classifiers. *Neural processing letters*, 9(3):293–300.

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.

TH Wen, M Gašić, N Mrkšić, PH Su, D Vandyke, and S Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Conference Proceedings-EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Tianyi Zhang, Varsha Kishore, Felix Wu, Kilian Q Weinberger, and Yoav Artzi. 2019. Bertscore: Evaluating text generation with bert. In *International Conference on Learning Representations*.

Yin Zhang, Rong Jin, and Zhi-Hua Zhou. 2010. Understanding bag-of-words model: a statistical framework. *International Journal of Machine Learning and Cybernetics*, 1(1-4):43–52.