Document Grounded Dialogue Agent

Anuja Tayal
657111812
UIC,Chicago
atayal4@uic.edu

ABSTRACT

The approach of training models started with replicating human minds and so started feeding them with more and more data. As a result, models like jeopardy were born which were good in answering trivia questions. But with that approach, researchers quickly realised that models don't need to memorise everything. In fact its okay to not learn everything but know how to fetch the required answers. Gradually, a paradigm shift took place and instead of training models with millions of data source- example wikipedia pages and others, these documents were indexed, stored and then fetched when required. This grounding documents technique has the potential make a generic dialogue agent model which can be used for different domains.

In this work, we construct a dialog agent which is grounded on government document (dmv) composed to guide users/customers through various information regarding motor vehicles department. We use the dmv subset of MultiDoc2dial dataset which is a dataset constructed from the government website in form of question answer pair. Using the dataset, we will generate the response to user inquiry. To construct the dialogue agent and generate the response of each query, it is absolutely necessary to understand the query or user utterance or user question, then find the document or context where response is present and finally generate the correct response. BLEU score is used to evaluate the model. With experiments, we show that by appending the dialog history to user questions, accuracy of the model increases as the model generates the context and query based on dialog history.

INTRODUCTION

A dialog system or conversational system converses with a human either to chit-chat (Alexa, Siri, Google Assistant) or to accomplish a goal. The dialog system which tries to accomplish a specific goal are known as goal oriented dialog systems. Till now, most goal oriented dialogue systems are restricted to restaurant booking, movie ticket, hotel or flight booking [4] but much less work has been done on task oriented dialogue systems based on grounded documents such as government websites (dmv,student aid, ssa).

Traditional Task Oriented Dialog System When a user requests something, text is recognised using automatic speech recognition, which is passed to Spoken Language Understanding (SLU) module where the system tries to understand the text by identifying the domain and the intent of the user. The output is fed as input to Dialog State Tracking (DST) module which maintains the state of the dialog according to the slots filled. Based on the gaps left unfilled, dialog manager and policy module by taking the help of the knowledge providers, generates the next response or dialog act which is passed as input to Natural Language Generation (NLG) module.

The knowledge providers used by the Dialog Management of goal oriented dialogue system is generally restricted to a database of movies or hotel booking. But it can also be extended to a single or multiple documents. We aim to construct a dialog agent grounded on a government document (dmv) which provides information regarding driving licence and other related answer queries related to motor vehicles to the customers.

Moreover, to find the correct and accurate response to user query, it is very important to find the correct context. If incorrect context is retrieved, the answer generated from it will definitely be wrong. On the other hand, if the correct context is retrieved, the chances of generating correct output is more. So the accuracy of the response indirectly depends on retrieving correct context.

To summarise, we propose to build a dialog agent grounded on dmv document in which the system needs to understand user request, retrieve the correct context, and provide an answer to the user query. In summary, following is our contribution:

- To understand the user query, we propose a dialog act prediction model.
- After understanding the user query by predicting the dialogue act, we propose to build a retrieval model which will retrieve the most relevant context from all the contexts given the query.
- We model the task of generating the response of the agent as a question-answering task where question represent the user's dialogue or query based on the document and the output is the span of the context containing the answer.
- When the most relevant context is retrieved, then we will build a question answer module from the query and context to extract the relevant portion of text which will constitute an answer.

ISBN 978-1-4503-2138-9.
DOI: 10.1145/1235

RELATED WORK

In this section, we will discuss some of the related works associated with the different modules presented for grounding document in a dialogue agent.

Question Answering Module

Question Answering is a very crucial task in Natural Language Processing which allows a user to ask question in natural language and get an immediate or brief response.

Machine Reading Comprehension

Machine Reading Comprehension is the task in Natural Language Processing where system understand an unstructured text and answer questions based on it. But normally in this architecture an assumption is made that answer will be continuous span of the passage due to which the problem is simplified to output 2 integers representing the start and end position of the passage.

Doc2Dial Shared Task

In [7], the authors enriched the RoBERTa-large model by appending different embeddings for predicting the knowledge span. While to generate the responses, the authors leverage curriculum learning in BART model.

Contextual Question Answering Systems

In [1], the authors constructed a student-teacher model question answering dataset based on wikipedia articles but the questions are more realistic, not every question has an answer, and the answers are only meaningful if the context of the article is known. But in the dataset, the student did not asked any follow-up or clarifying questions which was resolved in [10].

Search

Symmetric vs Asymmetric Semantic Search

If the query and context are of same length, symmetric semantic search is used but when context is very large as compared to the question(query), asymmetric semantic search is used.

Cosine Similarity vs Dot Product

If we use cosine similarity, then short documents or context will be retrieved but if the context is large as in our case, then dot product should be preferred.

Store Context Embeddings

As we do not want to compute context embeddings again and again, we could calculate context embeddings once and store them.

Elastic Search-

If a lot of metadata is present about the context, Elastic search is performed to use cross cluster.

FAISS (Facebook AI Similarity Search)-

FAISS [5] is a C++ library developed by Facebook which allows developers to quickly search for embeddings of millions of documents. It is an improvement over traditional methods like SQL by providing more scalable options to search the similarity. Meta data information is maintained in a database which maps to the FAISS id. To conduct similarity search, it compares two vectors with L2 (Euclidean distance) or dot product. If the two vectors are very similar, they will have lowest L2 distance or the highest dot product.

Retrieval Models

There are two kinds of retrieval models namely Sparse and Dense Passage Retrievals.

Sparse Retrievers-

TF-IDF-TF-IDF [11] finds the similarity between two pieces of text and finds the most relevant passage based on number of words found in the context and query.

Term Frequency (TF)- refers to the number of words from the query that is matched to the context.

Inverse Document Frequency (IDF)- is the inverse of the fraction of the documents containing the word. As it gives more preference to unique words.

BM25- BM25 is an improved version of TF-IDF in which the term frequency metric is weighted and also considers the document length as it gives preference to small document as opposed to long documents.

Dense Retrievers-

SBERT- SBERT[9] based on BERT encoder to generate encode the question as well as the passages and then generate relevance scores by performing the dot product or cosine similarity between the question vector and the vector of all passages and then choosing the passage with smallest angle. As smaller the angle between the vectors, higher will be the dot product between the two implying higher relevance score as shown in Figure 1.

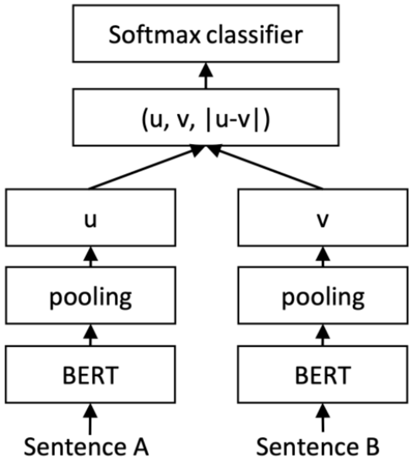


Figure 1. Sentence Bert

Dense Passage Retrieval (DPR) - DPR [6] is similar to SBERT in the sense that it uses BERT and dot product to calculate the relevance between the question vector and context vectors. However, as the length of question is generally smaller than the context, DPR uses 2 different BERT encoder-one to encode question and other to encode the passages.

DATASET

For the purpose of the project, we will be using dmv domain of MultiDoc2Dial[2] dataset, an extension of Doc2Dial[3] dataset, a new dataset consisting of goal oriented dialogues that are grounded in 4 government documents mainly dmv, ssa, veteran affairs and student aid which guides users to access information related to them. Multidoc2dial dataset consists of 4796 conversations from four domains where each dialogue turn is annotated with role, dialogue act, human generated utterance and the grounding span with document information. In MultiDoc2Dial A statistics of data is presented in Table 1 whereas a sample dialogue is shown in Figure 2. It is to note that each utterance is appended with the dialogue history as shown in Figure 3.

Multidoc2dial dataset contains complex and diverse dialogues whose dialogues are dynamic in nature. The dialogues are also personalised so they seem more human. One of the thing to note is that sometimes context have hyperlinks and references which points to different part of the document. Also, as it is a government document it also contains HTML mark ups, lists and document is sectioned with titles.

The authors have uploaded the dataset in the HuggingFace library [12] also so to make it easy for the researchers to load the dataset. There are 3 subsets available mainly multidoc2dial, dialogue domain and document domain. Multidoc2dial subset was used to retrieve the correct context and query while dialogue domain was used to predict the dialogue act.

Documents	149
Dialogues	1328
2 Segments	781
>2 Segments	257
Single	290

Table 1. DMV Dataset Statistics

METHODOLOGY

To create a conversational or dialogue agent which is grounded on the dmv domain, we proposed three tasks. First, to understand the query of the user, there is an NLU module which predicts the dialogue act of the user. Secondly, there is a Context retrieval module to find the correct context of the query. Because if the context retrieved is wrong, the answer retrieved would be wrong. Finally after retrieving the context of the query, it is given as the input along with the query to the Question Answer Model to retrieve the correct answer. The whole methodology can be viewed more clearly in Figure 4.

Dialogue Act Prediction

To predict the dialogue acts of each utterance, the utterances were analysed and found that each user or agent utterance can be classified into different dialogue acts mainly *Query Condition*, *Respond Solution*, *Query Solution* , *Response Positive*, *Response Negative*, *Respond No Solution*. A snippet of the utterances along with their respective dialogue acts can be viewed in Figure 2.

Therefore, we treat dialogue act prediction model as a multiclass classification problem with 6 classes. Each classes were assigned a label between 0-6. As the classes are imbalanced, so we used stratified split the data with 15%. To classify different dialogue acts, Bert Sequence Classification Model was trained on the dataset.

Context Retrieval

To find the correct and accurate response to user query, it is very important to find the correct context. If incorrect context is retrieved, the answer generated from it will definitely be wrong. On the other hand, if the correct context is retrieved, the chances of generating correct output is more. So the accuracy of the response indirectly depends on retrieving correct context.

As illustrated in the related work, we should perform asymmetric semantic search with dot product therefore we used SBERT **Msmarco-distilbert-base-dot-prod-v3** model from Sentence Transformer Library to encode the query and the contexts. Whereas to find the correct context given the query, FAISS [5] library has been used.

For finding the correct context for given query, following methodology was used

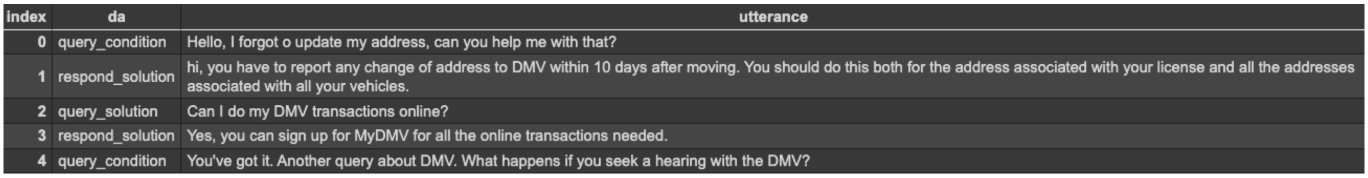
- Used SBERT to embed all the contexts.
- The context embeddings were indexed through FAISS.
- Whenever a user query is performed, embedding of the query is calculated
- The query embedding is searched through the indexed database of contexts to retreive top k results.

Question Answer Model

After retrieving the correct context, the next task is to extract the answer from the context given the query. As the answer or the solution to the query was a part of context, and was not the paraphrase of the context, BERT Model for Question Answering was used.

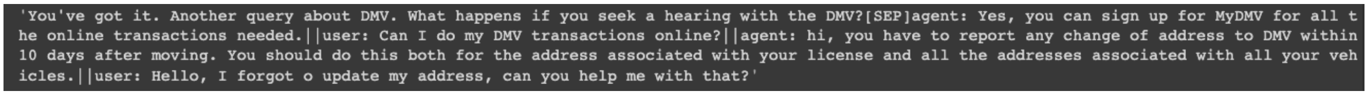
Bert or Bidirectional Encoder Representations from Transformers is designed to pre-train deep bidirectional representations taking into consideration both left and right contexts and can be fine-tuned very easily for a wide variety of NLP tasks. In Question Answering Module, BERT takes two parameters, the input question, and passage as a single packed sequence. The input embeddings are the sum of the token embeddings and the segment embeddings. Bert uses segment embedding to differentiate question from the passage. Or in other words, Segment "A" represent the question while Segment "B" represent the passage.

As these input tokens are passed through layers of transformers, sequences are tokenized and start vector is introduced. The probability of each word being the start token is calculated by taking the dot product between the word embedding and the start vector and taking the highest probability using the softmax function. Similar approach is used for predicting the word being the end token. The text between the start and the end token is considered the final answer as shown in Figure 6.



index	da	utterance
0	query_condition	Hello, I forgot o update my address, can you help me with that?
1	respond_solution	hi, you have to report any change of address to DMV within 10 days after moving. You should do this both for the address associated with your license and all the addresses associated with all your vehicles.
2	query_solution	Can I do my DMV transactions online?
3	respond_solution	Yes, you can sign up for MyDMV for all the online transactions needed.
4	query_condition	You've got it. Another query about DMV. What happens if you seek a hearing with the DMV?

Figure 2. Sample dmv dialogue



```
'You've got it. Another query about DMV. What happens if you seek a hearing with the DMV?[SEP]agent: Yes, you can sign up for MyDMV for all t
he online transactions needed.||user: Can I do my DMV transactions online?||agent: hi, you have to report any change of address to DMV within
10 days after moving. You should do this both for the address associated with your license and all the addresses associated with all your veh
icles.||user: Hello, I forgot o update my address, can you help me with that?'
```

Figure 3. Dialogue with History Appended

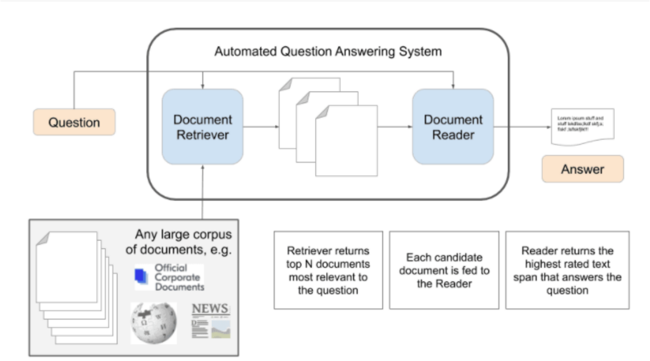

Automated Question Answering System
Question
Document Retriever
Document Reader
Answer
Any large corpus of documents, e.g.
Official Corporate Documents
NEWS
Retriever returns top N documents most relevant to the question
Each candidate document is fed to the Reader
Reader returns the highest rated text span that answers the question

Figure 4. Methodology of Document Grounded Dialog System

EXPERIMENTS

In this section, we will briefly explain the experiments performed on the 3 modules mainly Dialogue Act Prediction, Context Retrieval and Question Answering Module on the dmv domain of the Multidoc2dial [2] dataset. We have used transformers provided by the Hugging Face [12] library to fine-tune the models.

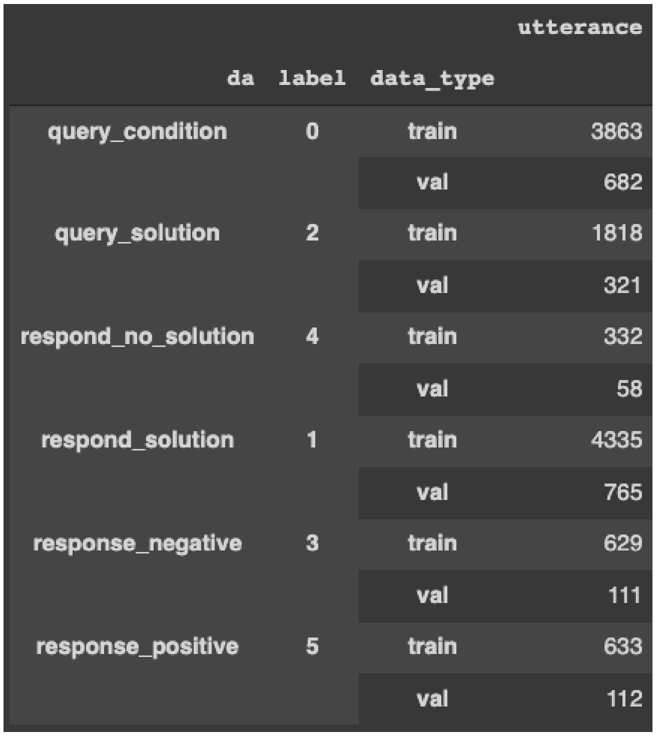
Dialogue Act Prediction

To classify the different dialogue acts into 6 classes, Bert Sequence Classification model was trained for multi class classification using Pytorch transformers for 4 epochs with batch size=4 and keeping the learning rate 1e-5. As the data was imbalanced, stratified split was used.

Context Retrieval

To retrieve the correct context, SBERT [9] was used to embed all the contexts. There were a total of 127 unique contexts with 6135 query and context values. SBERT **Msmarco-distilbert-base-dot-prod-v3** model from Sentence Transformer Library was used to encode the query and the contexts. After the contexts were embedded, FAISS library [5] was used to index the contexts. While searching the context of the query, the query was also embedded with SBERT and dot product calculated to get the correct context. Here top 3 contexts were retrieved and if the actual context is in one of the retrieved contexts, it was believed that correct context is extracted.

As the queries in the dataset are appended with the history, initial hypothesis was that after removing the history, model will be able to predict context better but that was not the case. History of the queries helped to better predict the correct context vector with 70% accuracy as opposed to without history whose accuracy was just 50% as can be seen from the results in Table 2.



			utterance
da	label	data_type	
query_condition	0	train	3863
		val	682
query_solution	2	train	1818
		val	321
respond_no_solution	4	train	332
		val	58
respond_solution	1	train	4335
		val	765
response_negative	3	train	629
		val	111
response_positive	5	train	633
		val	112

Figure 5. Dialog Act data Statistics

	Without History	With History
True	2788	4200
False	3347	1935
Accuracy	50%	70%

Table 2. Retrieval Model Accuracy

Question Answer Model

After retrieving the correct context, the next task is retrieve the answer, given the context and the query. Contexts were preprocessed to remove all hyperlinks, special characters and brackets. As the answers in the dataset only contained the

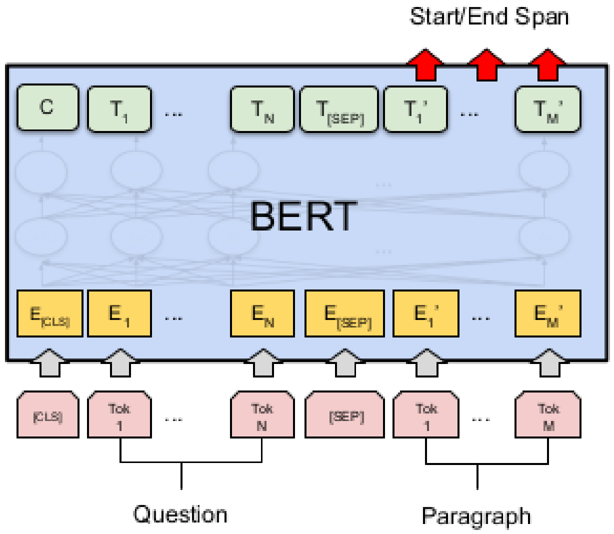


Figure 6. Question Answering System Using Bert

start index along with the correct answer while for training the BERT QA model, end index is needed so it was calculated by adding length of the answer with the start index. It should be noted that in some cases, start index was off by 1 or 2 characters.

As the length of the contexts are large with maximum length 5000, but Bert only allows maximum of 1024 words, rest of the words were truncated. As the answer of the query was part of the context, BertForQuestionAnswering Model from Transformers library was trained on the dmv dataset for 3 epochs with batch size 4 and learning rate 5e-5.

To evaluate the performance of the model BLEU (Bilingual Evaluation Understudy) [8] Score was used which is a metric used for evaluating a generated sentence to the reference sentence as it correlates with the human evaluation. It generates a score between 0 and 1 with 1 being a perfect match and 0 being imperfect or incorrect match. While comparing the answers generated with the expected answers, we got an average BLEU score of 0.6 which was good enough. Some of the results obtained can be seen in Table 3. As can be seen from the table, it generated the correct output for even large query question with long history.

FUTURE WORK

For now, we have considered conversations related to one field (dmv) but in future, we will also consider other domains as well. Even when considering other domains, in conversations context switch can take place. In the sense that conversation can involve multiple topics and user can query the system between domains as well (include contest switching) which will be included in future.

For embedding the contexts in the FAISS, we used SBERT to calculate the embeddings. But as the length of contexts and query is different, it is more appropriate to use different encoders and an appropriate model to do so is Dense Passage Retrieval [6] which we plan to use in future.

Question	Answer	Expected Answer
Hello I need information about paying the suspension termination fee[SEP]	suspension means your driver license or privilege to drive is taken away for a period of time before it is returned there are two types of suspensions definite	A suspension means your driver license or privilege to drive is taken away for a period of time before it is returned. There are two types of suspensions.
it is suggested that military personnel file a Notification of Military Service pdf [1] MV -75 to ensure that the license record is not purged once the expiration date has passed[SEP]agent: DMV has a procedure in place to exempt you from administrative actions upon your return. You must provide a notarized statement that the vehicle was not used during the time in question ...?	it is suggested that military personnel file a notification of military service pdf [1] mv - 75 to ensure that the license record is not purged once the expiration date has passed	it is suggested that military personnel file a Notification of Military Service pdf [1] MV -75 to ensure that the license record is not purged once the expiration date has passed.
Hello I need information about How to renew a registration[SEP]	need to renew your registration before it expires your new registration documents will arrive in the mail in about days renewing early will not change the new expiration date for your new registration members of the	You need to renew your registration before it expires. Your new registration documents will arrive in the mail in about 10 days. Renewing early will not change the new expiration date for your new registration.

Table 3. DMV Output

As the length of the contexts are large with maximum length 5000, but Bert only allows maximum of 1024 words, rest of the words were truncated. Due to this, some of the answers were truncated and were not generated correctly. So, in future we plan to use longformer to embed the contexts. An another alternative is to break the contexts into smaller contexts and index them or chained indexing can be used. For example, we can retrieve 1 big context and then subindex smaller contexts to more precisely get the answer context.

Till now the retrieval and answer generation module are separate but they need to be pipelined into a single model. Lastly, we are treating dialogue generation as question-answer model due to which answers generated are a bit bland and boring. So, final answer generation will be personalised in the future.

We have modeled the problem of modeling dialogues between user and agent as question answer system but the answers generated does not have any human touch in them. In future, we plan to personalise each user's response.

CONCLUSION

To conclude, we successfully tried to understand user utterances, query a large document by grounding/indexing the document- retrieve the required context and using the context retrieve the correct answer. Also, it was surprising to see that by appending the history to the user questions, accuracy of context retrieval and answer generation increased significantly.

REFERENCES

[1] Eunsol Choi, He He, Mohit Iyyer, Mark Yatskar, Wen-tau Yih, Yejin Choi, Percy Liang, and Luke Zettlemoyer. 2018. QuAC: Question Answering in Context. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2174–2184.

[2] Song Feng, Siva Sankalp Patel, Hui Wan, and Sachindra Joshi. 2021. MultiDoc2Dial: Modeling Dialogues Grounded in Multiple Documents. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Online and Punta Cana, Dominican Republic, 6162–6176.

[3] Song Feng, Hui Wan, Chulaka Gunasekara, Siva Patel, Sachindra Joshi, and Luis Lastras. 2020. doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 8118–8128.

[4] Donghoon Ham, Jeong-Gwan Lee, Youngsoo Jang, and Kee-Eung Kim. 2020. End-to-End Neural Pipeline for Goal-Oriented Dialogue Systems using GPT-2. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Online, 583–592.

[5] Jeff Johnson, Matthijs Douze, and Hervé Jégou. 2019. Billion-scale similarity search with GPUs. *IEEE Transactions on Big Data* 7, 3 (2019), 535–547.

[6] Vladimir Karpukhin, Barlas Oguz, Sewon Min, Patrick Lewis, Ledell Wu, Sergey Edunov, Danqi Chen, and Wen-tau Yih. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics, Online, 6769–6781.

[7] Boeun Kim, Dohaeng Lee, Sihyung Kim, Yejin Lee, Jin-Xia Huang, Oh-Woog Kwon, and Harksoo Kim. 2021. Document-Grounded Goal-Oriented Dialogue Systems on Pre-Trained Language Model with Diverse Input Representation. In *Proceedings of the 1st Workshop on Document-grounded Dialogue and Conversational Question Answering (DialDoc 2021)*. Association for Computational Linguistics, Online, 98–102.

[8] 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 of the Association for Computational Linguistics*. Association for Computational Linguistics, Philadelphia, Pennsylvania, USA, 311–318.

[9] Nils Reimers and Iryna Gurevych. 2019. Sentence-BERT: Sentence Embeddings using Siamese BERT-Networks. 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)*. Association for Computational Linguistics, Hong Kong, China, 3982–3992.

[10] Marzieh Saeidi, Max Bartolo, Patrick Lewis, Sameer Singh, Tim Rocktäschel, Mike Sheldon, Guillaume Bouchard, and Sebastian Riedel. 2018. Interpretation of Natural Language Rules in Conversational Machine Reading. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Brussels, Belgium, 2087–2097.

[11] Claude Sammut and Geoffrey I. Webb (Eds.). 2010. *TF–IDF*. Springer US, Boston, MA, 986–987.

[12] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, Rémi Louf, Morgan Funtowicz, Joe Davison, Sam Shleifer, Patrick von Platen, Clara Ma, Yacine Jernite, Julien Plu, Canwen Xu, Teven Le Scao, Sylvain Gugger, Mariama Drame, Quentin Lhoest, and Alexander M. Rush. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*. Association for Computational Linguistics, Online, 38–45.