BART for Knowledge Grounded Conversations

Maxime De Bruyn
Computational Linguistics & Psycholinguistics Research Center
University of Antwerp
maxime.debruyn@uantwerpen.be

Ehsan Lotfi
Computational Linguistics & Psycholinguistics Research Center
University of Antwerp
ehsan.lotfi@uantwerpen.be

Jeska Buhmann
Computational Linguistics & Psycholinguistics Research Center
University of Antwerp
jeska.buhmann@uantwerpen.be

Walter Daelemans
Computational Linguistics & Psycholinguistics Research Center
University of Antwerp
walter.daelemans@uantwerpen.be

ABSTRACT

Transformers have shown incredible capabilities for conversation modeling, however, they store factual knowledge in their learned parameters, which is costly to update with new knowledge since it requires retraining. Models trained before the Coronavirus pandemic do not know about COVID-19.

In this paper, we investigate how a BART model can be adapted to a knowledge grounded conversational setup. We introduce the notion of *key* and *query* tokens to retrieve knowledge stored in an external database, that can easily be updated with new knowledge. As factual knowledge can hardly be reduced to a single sentence or vector, we allow the model to retrieve multiple sentences from the memory.

Our analysis shows perplexity decreases with the number of passages retrieved from memory. Second, our analysis shows a shared encoder for knowledge retrieval, and conversation understanding reduces the model size and perform as well as a specialized module.

CCS CONCEPTS

• **Computing methodologies → Discourse, dialogue and pragmatics**.

KEYWORDS

neural networks, transformers, conversational agents, knowledge retrieval, shared encoder

ACM Reference Format:
Maxime De Bruyn, Ehsan Lotfi, Jeska Buhmann, and Walter Daelemans. 2020. BART for Knowledge Grounded Conversations. In *Proceedings of KDD Workshop on Conversational Systems Towards Mainstream Adoption (KDD Converse'20).* ACM, New York, NY, USA, 6 pages.

1 INTRODUCTION

Large transformer-based language models have shown excellent capabilities in generating human-like conversations [1, 11]. While powerful, these models have a major drawback: they cannot expand their factual knowledge of the world without being trained on new data [7]. As an example, all models trained before the COVID-19 outbreak have no knowledge about the coronavirus epidemic.

It should be possible to allow open-domain conversational models to use additional external knowledge sources for factual knowledge. Their knowledge source should be easily extendable with recent information.

Current knowledge grounded open-domain agents limit the external world knowledge to one sentence [3, 11], or to a single vector [4]. We believe limiting models this way is insufficient for open-domain conversational agents, and show that increasing the number of passages retrieved from memory leads to more human-like replies from the agent.

2 RELATED WORK

Knowledge grounded dialog systems can be described as sequence-to-sequence models [12] conditioned on an external knowledge source. Grounding a conversation in external knowledge requires two different abilities: retrieving the right knowledge amongst multiple candidates and effectively using this knowledge to generate the next utterance.

One way of providing context to the model is to concatenate the chat history with the knowledge source. Budzianowski and Vulic [2] concatenated the context, belief state, and database as input to a task-oriented GPT2 model [10]. Wolf et al. [17] concatenated the agent's persona with the previous chat history. Liu et al. [8] find that this approach struggles with handling longer contexts. Wang et al. [14] separate source and context encoding and interleave source and context attention when decoding.

In some cases, the length of the context may be too large to be concatenated with the chat history (e.g. multiple Wikipedia articles). Dinan et al. [3] introduce the Transformer Memory Network models, capable of retrieving and attending to knowledge and outputting a response, either in retrieval or generative mode. Fan et al. [4] present a KNN-based information fetching module that learns to identify relevant information from external knowledge sources in the context of a dialogue dataset. The Wizard Generative Model

[11] uses a Poly-encoder [5] to retrieve a single sentence from an external knowledge source.

Retrieval dialog systems [3, 15] can also be framed as knowledge grounded agents where the knowledge source is a fixed set of utterances and the task is to select the most appropriate utterance given the context.

In this paper, we expand on the work of Dinan et al. [3]. We fine-tune a BART model [6] to retrieve multiple sentences (instead of a single one) from an external knowledge source, and use it effectively to generate the next utterance in a conversation.

3 DATASET

3.1 Wizard of Wikipedia

We use the Wizard of Wikipedia dataset [3] where two participants engage in chit-chat conversations. One of the participants is the wizard, and the other the apprentice. The wizard plays the role of a knowledgeable expert while the other is a curious learner. The goal of the wizard is to inform its conversation partner about a topic that one of them will choose. The wizard has access to an information retrieval system that shows paragraphs from Wikipedia possibly relevant to the conversation. Before each conversation turn, the wizard can read these paragraphs and then potentially base its next reply on that observed knowledge.

The authors collected 22,311 conversations with a total of 201,999 turns on 1365 open-domain dialog topics (e.g. commuting, gouda cheese, bowling).

The dataset is divided in a train, validation and test set. The validation and test sets are sub-divided into seen and unseen. The seen sets share conversation topics with the training set while the unseen sets do not.

4 MODEL

Our goal with this dataset is to train an agent capable of conversing about any domain. We use a model to replace the wizard in the conversations. To generate the next utterance x_{t+1}, the model has access to the previous conversation turns $x_1, ..., x_t$ and to a hierarchical knowledge source: M. Each element of the knowledge source, m_i, is composed of a topic and a sentence belonging to that topic.

Similar to the End-to-End (E2E) Transformer MemNet of Dinan et al. [3], we use an encoder-decoder Transformer [13] as our base sequence-to-sequence model. Instead of pre-training our model on a Reddit corpus, we use a pre-trained BART model [6]. An illustration of our model is shown in Figure 1.

The knowledge source M is filtered before each turn using the same procedures as in Dinan et al. [3].

4.1 Encoder

While some approaches choose to have a separate encoder for knowledge retrieval and for conversation modeling [3, 4], we use a shared encoder to encode the conversation context x, and the filtered knowledge source M. Every m_i is encoded independently.

We choose to share the encoder because the purpose of an encoder is to understand text, it does not make sense to have two encoders do the same thing but for different sources (chat history and knowledge memory). This architectural choices also reduces the model size.

To let the model recognize the difference between a knowledge piece and a conversation history, we use segment embeddings [17]. We introduce three segment embeddings, one for the wizard's turn, one for the apprentice's turn, and one for the knowledge passages. We prepend the conversation context x with a special token q, the query token. We prepend each knowledge source candidate with another special token k, the key token. After decoding, the query and key vectors are projected to a lower dimension using a linear layer.

4.2 Knowledge Selection

After the encoding step, our key and query tokens become the key and query vectors. We concatenate the key vectors k_i from the knowledge source encoding into the query matrix K.

We then train the model to recognize which single knowledge passage (the gold knowledge) k_i was selected by the wizard. The query vector q from the conversation history is compared against K to produce a dot product attention over the knowledge source. We train the model to select the gold knowledge passage with a cross-entropy loss over the knowledge attention.

4.3 Decoder

We concatenate the full sequence of the n first knowledge candidates m_i with the chat history. This context matrix is then given as memory to the decoder. To let the decoder know that it is generating the next utterance of the wizard, we use the same segment embedding for the wizard as in the encoding step. We train the model to minimize the negative log-likelihood of the response utterance x_{t+1}.

To summarize, our model uses a pre-trained BART model to retrieve relevant knowledge and to generate the next utterance. We improve on the current methods in two ways. First we introduce a *key* and *query* token to perform the retrieval step with a shared encoder, second we allow the model to retrieve multiple full (i.e. not a vector representation) passages from the memory.

5 EXPERIMENTS

We conduct several experiments to analyze the ability of our model to select and use knowledge in dialogue.

The model uses the large version of BART [6], which has 12 layers in the encoder and 12 layers in the decoder and an inner dimension of 1024. The model has approximately 400M parameters. We use the BART implementation from HuggingFace [16].

On top of the original implementation, we add a segment embedding layer and two additional tokens (query and key token) to the vocabulary. We also add two linear layers to project the key and query vector to a dimension of 512.

Before each turn, the wizard is presented with a varying number of memory passages retrieved by an IR system: the visible passages (see Dinan et al. [3] for a detailed description). We feed a subset of the visible passages to the model (40 sentences per utterance). The visible passages can be divided into positive (gold passage) and negative examples (non-gold passages). We pool together the negative examples of a single batch to increase the difficulty of the task at a reduced computational cost (the model has to choose from a larger pool of already computed *key* vectors).

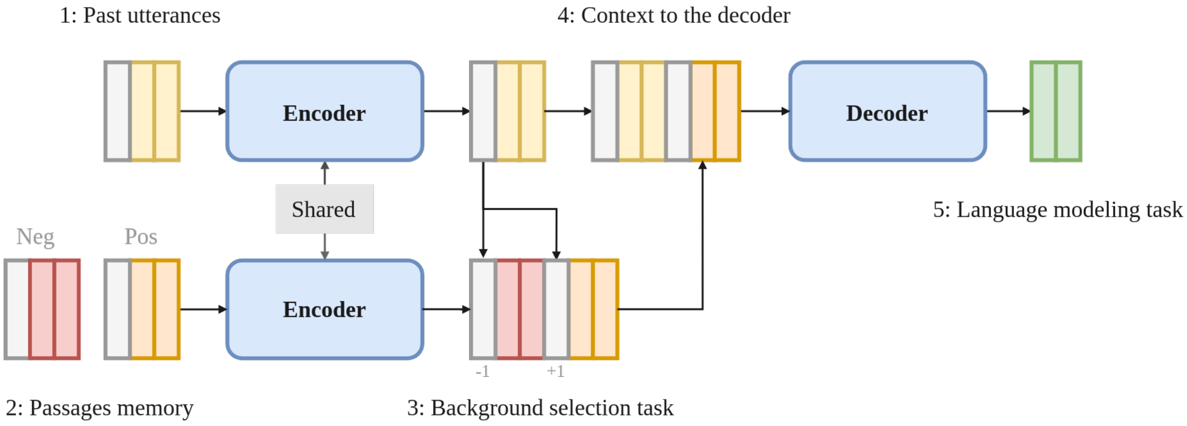
1: Past utterances
4: Context to the decoder
Encoder
Decoder
Shared
5: Language modeling task
Neg
Pos
Encoder
-1
+1
2: Passages memory
3: Background selection task

Figure 1: BART model adapted for knowledge grounded conversations.
1: The chat history is tokenized, a *query* token is prepended, the result is encoded by the encoder. 2: Each memory passage is tokenized, and prepended with a *key* token. The resulting matrix is encoded by the encoder. 3: The *query* vector is compared against the *key* vectors from the memory with a dot product attention. The first k passages with the highest score are selected. 4: The full sequence of the chat history and the full sequence from the selected memory passages are concatenated and given as context to the decoder. 5: Generation of the next utterance of the Wizard.

During training, we use a forcing teacher strategy and disregard the results from the knowledge retrieval step. Instead, we give as context (memory) to the decoder, the first five passages from the gold topic (the gold passage is always the first one). By feeding it multiple sentences, the model is trained to further select the relevant piece of information in the decoder. We believe this makes the decoder more robust to noise coming from the knowledge retrieval step.

We train the model to simultaneously optimize for the knowledge selection task and the language modeling task for three epochs, with a constant learning rate of 10^{-5}, linearly increased from zero over 1000 steps.

We did not test using a separate encoder (see *Shared* link in Figure 1) as this would increase the parameters count by 50%.

6 RESULTS

We analyze the performance of the model on two axes: knowledge retrieval and next utterance generation.

6.1 Retrieval Task

Similar to Dinan et al. [3] we use recall@1 and unigram F1 between the retrieved knowledge and the gold knowledge item as evaluation metrics. The results are displayed in Table 1.

Our model uses a shared encoder to encode the conversation context and to retrieve the relevant knowledge. It is therefore best compared against the Generative E2E Transformer MemNet of Dinan et al. [3], the other models have a separate knowledge retrieval module. We show that a shared encoder can achieve similar performance on this task as specialized modules.

As our model is capable of handling multiple knowledge pieces in the decoder, we also report recall@5 and recall@10 in Figure 2. The first five results contain the gold passage around 50% of the time.

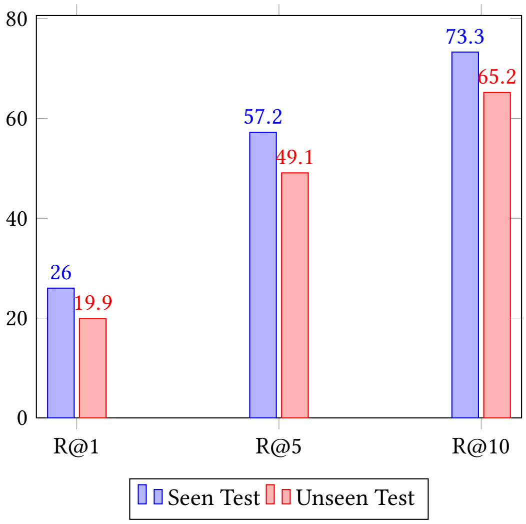

Figure 2: Recall metrics on seen and unseen test set.
The model retrieves the right memory passage 26% of the time on the seen test set. When retrieving the first 10 passages, the gold passage is included in the retrieved results 73% of the time. These results show the importance of retrieving multiple passages from the memory.

The difference in retrieval performance between the seen and unseen set could indicate that the model overfitted the training set, or that the size of the dataset is too limited to generalize to unseen topics.

Method	Seen Test		Unseen Test	
	R@1	F1	R@1	F1
Random	2.7	13.5	2.3	13.1
IR baseline	5.8	21.8	7.6	23.5
BoW MemNet	23.0	36.3	8.9	22.9
Transformer	22.5	33.2	12.2	19.8
Transformer (+Reddit pretraining)	24.5	36.4	**23.7**	**35.8**
Transformer (+Reddit pretraining, +SQuAD training)	25.5	36.2	22.9	34.2
Retrieval Transformer MemNet (no auxiliary loss)	12.9	24.6	14.6	26.3
Generative E2E Transformer MemNet (no auxiliary loss)	13.4	28.3	11.8	25.9
Generative E2E Transformer MemNet (w/ auxiliary loss)	21.1	32.8	14.3	22.8
BART	**26.0**	**38.9**	19.9	33.9

Table 1: Knowledge retrieval performance on the seen and unseen test set.

The BART model outperforms all methods on the seen test set (unigram F1 and perplexity) and comes close to the best performing methods on the unseen test set, even though it does not have a separate module specialized in knowledge retrieval (as the Transformer models). The seen test set shares conversation topic with the training set, while the unseen test does not.

Method	Seen Test		Unseen Test	
	PPL	F1	PPL	F1
Repeat last utterance		13.8		13.7
E2E Transformer MemNet (no auxiliary loss)	66.5	15.9	103.6	14.3
E2E Transformer MemNet (w/ auxiliary loss)	63.5	16.9	97.3	14.4
Two-Stage Transformer MemNet	54.8	18.6	88.5	17.4
Two-Stage Transformer MemNet (w/ K.D.)	46.5	18.9	84.8	17.3
KIF-Augmented Transformer*		**25.9**		**22.3**
BART	12.2	20.1	14.9	19.3

Table 2: Next utterance generation performance on the seen and unseen test set.

The BART model outperforms the shared encoder methodologies (E2E) and non-shared encoder methodologies (Two-Stage) of Dinan et al. [3], but falls short of the KIF-Augmented Transformer [4] in terms of unigram F1. Perplexity number cannot be directly compared because of differences in vocabulary sizes. *Fan et al. [4] did not report perplexity numbers.

6.2 Generation Task

The second objective of our model is to use the past conversation and the retrieved knowledge to generate the next utterance.

Although BART was pre-trained on a denoising task, it quickly adapted to dialog generation.

Similar to Dinan et al. [3], we use the perplexity of the gold utterance and unigram F1 between the generated utterance and the gold utterance as evaluation metrics, see Table 2. The model achieves a better performance than [3] in terms of unigram F1 but falls short of Fan et al. [4]. The perplexity numbers cannot be directly compared between models because of differences in vocabulary size.

As our model is capable of handling more than one passage of knowledge, we also report the numbers for 0, 1, 5, 10, 15 and 20 knowledge passages retrieved, see Figures 3 and 4. In terms of perplexity, the more passages are retrieved, the better the performance. The higher the number of passages retrieved, the higher the probability of it containing the gold passage used by the Wizard to generate the next utterance (see Figure 2 for recall numbers). Hence, the model is less perplexed by this particular utterance. This phenomenon is true for the seen and unseen test set.

In terms of unigram F1, the performance reaches a maximum at 1 passage retrieved, while the unseen test reaches a maximum at 10. Unigram F1 and perplexity tell two different stories: perplexity says it is beneficial to include at least 10 passages, while unigram F1 says one is enough.

Adiwardana et al. [1] show perplexity is correlated with SSA (Sensibleness and Specificity Average) and state that optimizing for perplexity is a good proxy for optimizing the human likeliness of a model. Using their result as hypothesis, increasing the number of passages retrieved from memory results in more human-like models.

7 FUTURE WORK

An unsupervised pre-training of the BART model for simultaneous context retrieval and generation could help bridge the gap between seen and unseen performance.

The problem of knowledge selection is not one-to-one, but one-to-many. There are possibly many relevant passages for a single user query. The dataset could be updated to reflect that fact.

Unigram F1 has no semantic understanding of the generated text. Evaluating the model with USR [9], a reference-free metric that trains unsupervised models to measure several desirable qualities of dialog, could help in the comparison of models.

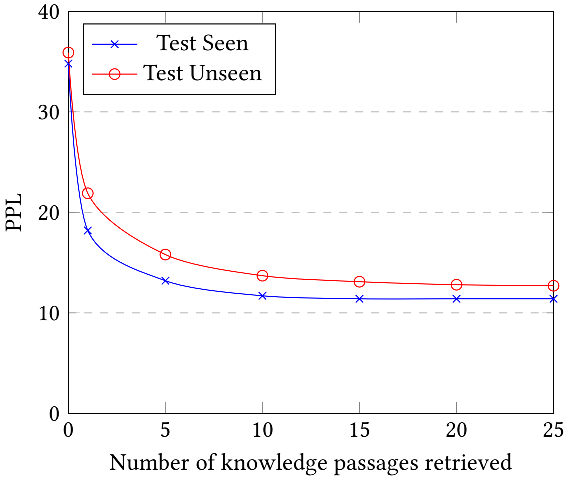

Figure 3: Perplexity results per number of passages retrieved.
The inclusion of knowledge has a significant impact on perplexity, on the seen and unseen test set. The performance of the model gets better as more knowledge passages are retrieved. There is no trade-off between the number of included passages and the model's performance in terms of perplexity.

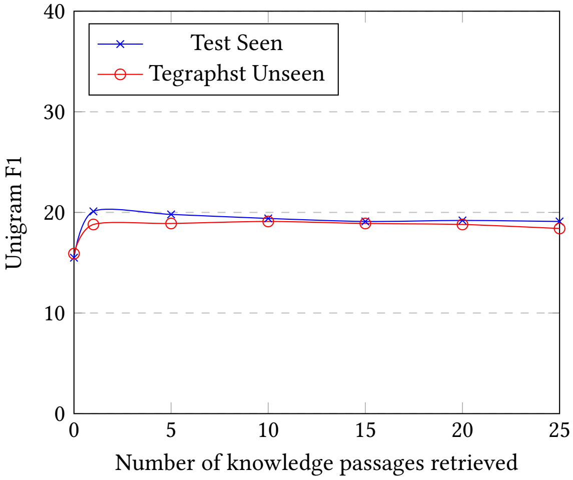

Figure 4: Unigram F1 results per number of passages retrieved.
The inclusion of knowledge has a significant impact on the model's performance in terms of unigram F1. Contrary to Figure 3, the model's performance peaks at one passage retrieved on the seen test set.

8 CONCLUSION

In this work, we showed how a BART [6] model can be extended to make use of an external memory. This model was successfully implemented in a knowledge grounded conversational setup using the Wizard of Wikipedia dataset [3].

Current models retrieve only one sentence or vector from the memory [3, 11]. Our analysis showed that it is limiting the potential of current models as retrieving multiple sentences from the memory diminishes the model's perplexity to the gold utterance.

We also showed it is not necessary to have a separate encoder for knowledge retrieval and context encoding. A shared encoder can achieve competitive results in the knowledge retrieval task, limiting the model size and complexity.

ACKNOWLEDGMENTS

This research received funding from the Flemish Government (AI Research Program).

REFERENCES

[1] Daniel Adiwardana, Minh-Thang Luong, David R. So, Jamie Hall, Noah Fiedel, Romal Thoppilan, Zi Yang, Apoorv Kulshreshtha, Gaurav Nemade, Yifeng Lu, and Quoc V. Le. 2020. Towards a Human-like Open-Domain Chatbot. arXiv:2001.09977 [cs.CL]

[2] Pawel Budzianowski and Ivan Vulic. 2019. Hello, It's GPT-2 - How Can I Help You? Towards the Use of Pretrained Language Models for Task-Oriented Dialogue Systems. *CoRR* abs/1907.05774 (2019). arXiv:1907.05774 http://arxiv.org/abs/1907.05774

[3] Emily Dinan, Stephen Roller, Kurt Shuster, Angela Fan, Michael Auli, and Jason Weston. 2019. Wizard of Wikipedia: Knowledge-powered Conversational Agents. In *Proceedings of the International Conference on Learning Representations (ICLR).*

[4] Angela Fan, Claire Gardent, Chloe Braud, and Antoine Bordes. 2020. Augmenting Transformers with KNN-Based Composite Memory for Dialogue. arXiv:2004.12744 [cs.CL]

[5] Samuel Humeau, Kurt Shuster, Marie-Anne Lachaux, and Jason Weston. 2019. Real-time Inference in Multi-sentence Tasks with Deep Pretrained Transformers. *CoRR* abs/1905.01969 (2019). arXiv:1905.01969 http://arxiv.org/abs/1905.01969

[6] Mike Lewis, Yinhan Liu, Naman Goyal, Marjan Ghazvininejad, Abdelrahman Mohamed, Omer Levy, Ves Stoyanov, and Luke Zettlemoyer. 2019. BART: Denoising Sequence-to-Sequence Pre-training for Natural Language Generation, Translation, and Comprehension. arXiv:1910.13461 [cs.CL]

[7] Patrick Lewis, Ethan Perez, Aleksandara Piktus, Fabio Petroni, Vladimir Karpukhin, Naman Goyal, Heinrich Küttler, Mike Lewis, Wen tau Yih, Tim Rocktäschel, Sebastian Riedel, and Douwe Kiela. 2020. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. arXiv:2005.11401 [cs.CL]

[8] Peter J. Liu, Mohammad Saleh, Etienne Pot, Ben Goodrich, Ryan Sepassi, Lukasz Kaiser, and Noam Shazeer. 2018. Generating Wikipedia by Summarizing Long Sequences. In *6th International Conference on Learning Representations, ICLR 2018, Vancouver, BC, Canada, April 30 - May 3, 2018, Conference Track Proceedings*. OpenReview.net. https://openreview.net/forum?id=Hyg0vbWC-

[9] Shikib Mehri and Maxine Eskenazi. 2020. USR: An Unsupervised and Reference Free Evaluation Metric for Dialog Generation. arXiv:2005.00456 [cs.CL]

[10] Alec Radford, Jeff Wu, Rewon Child, David Luan, Dario Amodei, and Ilya Sutskever. 2019. Language Models are Unsupervised Multitask Learners. (2019).

[11] Stephen Roller, Emily Dinan, Naman Goyal, Da Ju, Mary Williamson, Yinhan Liu, Jing Xu, Myle Ott, Kurt Shuster, Eric M. Smith, Y-Lan Boureau, and Jason Weston. 2020. Recipes for building an open-domain chatbot. *arXiv preprint arXiv:2004.13637.*

[12] Ilya Sutskever, Oriol Vinyals, and Quoc V. Le. 2014. Sequence to Sequence Learning with Neural Networks. *CoRR* abs/1409.3215 (2014). arXiv:1409.3215 http://arxiv.org/abs/1409.3215

[13] Ashish Vaswani, Noam Shazeer, Niki Parmar, Jakob Uszkoreit, Llion Jones, Aidan N. Gomez, Lukasz Kaiser, and Illia Polosukhin. 2017. Attention Is All You Need. *CoRR* abs/1706.03762 (2017). arXiv:1706.03762 http://arxiv.org/abs/1706.03762

[14] Xinyi Wang, Jason Weston, Michael Auli, and Yacine Jernite. 2019. Improving Conditioning in Context-Aware Sequence to Sequence Models. arXiv:1911.09728 [cs.CL]

[15] Jason Weston, Emily Dinan, and Alexander Miller. 2018. Retrieve and Refine: Improved Sequence Generation Models For Dialogue. In *Proceedings of the 2018 EMNLP Workshop SCAI: The 2nd International Workshop on Search-Oriented Conversational AI.* Association for Computational Linguistics, Brussels, Belgium, 87–92. https://doi.org/10.18653/v1/W18-5713

[16] Thomas Wolf, Lysandre Debut, Victor Sanh, Julien Chaumond, Clement Delangue, Anthony Moi, Pierric Cistac, Tim Rault, R'emi Louf, Morgan Funtowicz, and Jamie Brew. 2019. HuggingFace's Transformers: State-of-the-art Natural Language Processing. *ArXiv* abs/1910.03771 (2019).

[17] Thomas Wolf, Victor Sanh, Julien Chaumond, and Clement Delangue. 2019. TransferTransfo: A Transfer Learning Approach for Neural Network Based Conversational Agents. *CoRR* abs/1901.08149 (2019). arXiv:1901.08149 http://arxiv.org/abs/1901.08149