Interspeech 2018
2-6 September 2018, Hyderabad

Comparison of an end-to-end trainable dialogue system with a modular statistical dialogue system

Norbert Braunschweiler[1], *Alexandros Papangelis*[1]

[1]Toshiba Research Europe Limited, Cambridge Research Laboratory, Cambridge, UK
{norbert.braunschweiler,alex.papangelis}@crl.toshiba.co.uk

Abstract

This paper presents a comparison of two dialogue systems: one is end-to-end trainable and the other uses a more traditional, modular architecture. End-to-end trainable dialogue systems recently attracted a lot of attention because they offer several advantages over traditional systems. One of them is the avoidance to train each system module independently, by creating a single network architecture which maps an input to the corresponding output without the need for intermediate representations. While the end-to-end system investigated here had been tested in a text-in/out scenario it remained an open question how the system would perform in a speech-in/out scenario, with noisy input from a speech recognizer and output speech generated by a speech synthesizer.

To evaluate this, both dialogue systems were trained on the same corpus, including human-human dialogues in the Cambridge restaurant domain, and then compared in both scenarios by human evaluation. The results show, that in both interfaces the end-to-end system receives significantly higher ratings on all metrics than the traditional modular system, an indication that it enables users to reach their goals faster and experience both a more natural system response and a better comprehension by the dialogue system.

Index Terms: spoken dialogue systems, end-to-end trainable, statistical dialogue systems, human-computer interaction, human evaluation

1. Introduction

To enable humans to converse more naturally with artificial devices remains one of the big challenges in dialogue systems research. While much progress has been achieved in this field, there is still a gap which surfaces in the perception of mishaps in speech recognition, delays in system response, lack of comprehension by the dialogue system, and inappropriate ways of system responses in both linguistic structure and synthesized expression. The complexity of the whole pipeline from speech recognition to speech synthesis poses significant challenges. Recently, a lot of research has been dedicated to end-to-end systems [1, 2, 3, 4, 5] which provide significant advantages over more traditional modular systems, e.g. the avoidance to train each system module independently, removing the need for tediously hand-crafted rules or expensive hand-labelled training data.

The success of end-to-end systems has been dominantly in non-task oriented domains [1, 2, 6] while it is more limited for task-oriented applications. Non-task oriented systems typically predict the next response in a chat system for e.g. social media [7], while in a task-oriented setting, the dialogue system needs to be capable to comprehend a user request, interact with a knowledge base and compile useful information into their response aiming to achieve this within the smallest possible number of dialogue turns.

The end-to-end system (E2E) evaluated here, belongs to the task-oriented family. It is based on the system presented in [8] which is an end-to-end trainable, neural network driven dialogue system. While this model is end-to-end trainable, i.e. each system module can be trained from data except for a database operator, it is still modularly connected. Therefore representing an intermediate system between a traditional modular system and a pure end-to-end system. The E2E-system achieved a high success rate and performed very competitively despite being trained on a relatively small corpus of 676 dialogues [8].

However, this E2E-system had been evaluated by using only text as input and output modalities, but it remained an open question how it would perform when noisy speech recognition input would be used and the output would be generated by speech synthesis.

Therefore, this paper addresses this question and adds speech recognition at the input side and speech synthesis at the output side to the E2E-system. This E2E-spoken dialogue system is then compared to a baseline system, which is a more traditional modular spoken dialogue system (BASE-SDS), in an evaluation where humans interact with both systems to find restaurants in Cambridge, UK that fulfil certain criteria.

Both dialogue systems were first evaluated in the text-in/out condition to enable a comparison with the subsequent evaluation in the speech-in/out scenario. A detailed description of the evaluations is provided, in which subjects judged the quality of the systems in terms of dialogue success, naturalness of response, and comprehension ability. Results indicate that the E2E-system is very competitive. A discussion addresses some of the issues experienced during the evaluation and raises some aspects for future investigations.

2. Method

2.1. Enabling the speech interface

To enable a speech-based evaluation, the E2E-system first had to be connected with automatic speech recognition and speech synthesis engines, capable of running in real-time or faster for fluent system interaction.

For speech recognition, a python-speech recognition module [9] was used in combination with the PyAudio package [10] and Google's speech recognition API [11]. The default recognition language, American English (en-US) was used and the setting 'adjust_for_ambient_noise' was enabled with the microphone always on. The text output of the recognizer, i.e. the top-listed transcription was fed into the dialogue systems.

For the speech synthesis program, Toshiba's in-house text-to-speech (TTS) engine ToSpeak (see [12] for a front-end description) was chosen, which is a statistical parametric speech

10.21437/Interspeech.2018-1679

synthesizer using Hidden Markov models (HMM-TTS) to model spectrum, excitation and duration of speech. A British English male voice was selected for this experiment which previously had been identified as the most preferred voice in the context of dialogue system outputs.

The output text generated by the dialogue systems was passed to the Toshiba HMM-TTS to synthesize spoken system responses. However, the text generated by both systems was slightly modified to avoid any un-desired output stemming from irregular orthography. In particular, the marking of possessive case is generated by the E2E-system as "restaurant s" which was converted to "restaurant's" and the conversion of derived adverbs such as "cheap -ly" to "cheaply".

2.2. Training corpus

To enable a comparison on level terms, both dialogue systems were trained on the same corpus. That means, for the BASE-SDS the trainable part, i.e. the policy model was trained on the same corpus as used for training the end-to-end system. The training corpus included 1200 dialogues of human-human data in the Cambridge restaurants domain and is available at https://mi.eng.cam.ac.uk/˜nm480/woz_2.0.zip.

It was collected by [8] and [13] in a Wizard-of-Oz style using Amazon Mechanical Turk users in which both the roles of the user and the system are played by humans. Users had to follow a given task which included the specifications of particular restaurants in Cambridge, UK which they had to find. Users typed their inputs instead of using speech. The role of the dialogue system was also played by Amazon Mechanical Turk users, who had to record the information provided in the last user turn, which was then entered into a search in the Cambridge restaurant database, and given the output of that search, they would then type their response.

An example of a dialogue in the corpus is given in Table 1 for the following task:

Task: You are looking for a moderately priced restaurant and it should be in the north part of town. Make sure you get the address and type of food.

Table 1: *Example of a dialogue in the training corpus.*

User:	*I would like a moderately priced restaurant in the north part of town.*
System:	Golden Wok is a moderately priced restaurant in the north side of town.
User:	*What type of food does Golden Wok serve?*
System:	The Golden Wok serves Chinese food. Would you like more information?
User:	*Yes what is the address for the business?*
System:	They are located at 191 Histon Road Chesterton
User:	*Great. I have everything I need, Thank you*
System:	You're welcome

To enable the usage of the corpus for training the baseline system and the E2E-SDS it was annotated with success-flags, by using the following guidelines: a) check that the restaurant provided meets all the constraints in the task, and b) check that the system informs the user about all the requested information. Also, dialogue acts were annotated and the corpus was transformed so that it could be used with the training environment provided in the E2E-system. The corpus was divided by a ratio of 3:1:1 into training, validation and testing sets.

2.3. Baseline dialogue system (BASE-SDS)

As baseline system a more traditional, modular statistical dialogue system (BASE-SDS) was chosen which was based on the PyDial toolkit [14].

The BASE-SDS uses a hand-crafted semantic parser, a deep Q-network-based policy model with two hidden layers of size 60 and 40 [15], and domain independent parametrisation (DIP) features [16, 17] for input; the rule-based focus tracker [18] for belief tracking and a template based language generator.

The deep Q-network based policy model of the BASE-SDS was retrained with the 1200 dialogues corpus, i.e. trained from data not using simulators.

The corpus based evaluation using user simulators resulted in an objective success rate of 84.85%. An agenda-based user simulator was used, provided by [14]. A dialogue is classified as successful if a) the item offered by the system matches the constraints of the user, and b) all user requests have been satisfied. Please note, that the user simulator operates at the dialogue act level, therefore bypassing automatic speech recognition and spoken language understanding (SLU).

From the output of the user simulator, a semantic error simulator generates an n-best list of SLU hypotheses by changing either the dialogue act type, slot, or slot value according to the semantic error rate. To calculate the objective success rate, the semantic error rate was set to 15%. This approach enables one to quickly evaluate dialogue managers, but the learned dialogue policies tend to take advantage of characteristics of the simulator. As such, it might not reflect characteristics of real users who might not always comply to system requests or provide more fuzzy inputs.

For speech input/output, the BASE-SDS used the same ASR and TTS engines as in the end-to-end system.

2.4. End-to-End dialogue system (E2E-SDS)

The end-to-end dialogue system can be roughly divided into an encoder and decoder part as follows: On the encoder side, the system uses an LSTM (Long Short-term Memory [19]) network for encoding the user utterance, a set of RNN and CNN-based slot trackers that are designed to keep track of each slot-value pair across turns, and a database operator which enables the system to access a discrete database.

On the decoder side, the E2E-system uses a policy network for decision-making and producing the vector for decoding, finally an LSTM decoder network to generate the system response [8].

The E2E-system was re-trained with the 1200 dialogues corpus described in section 2.2, i.e. the same corpus used for training the BASE-SDS. The training included first, the training of belief trackers, which is then followed by the training of the remaining parts of the model.

Belief trackers can be seen as simple semantic parsers, which enable a dialogue system to map freely expressed natural language sentences into a fixed set of slot-value pairs (e.g. slot=*food*, value=*Italian*). The latter can then be used to conduct a search in a database. Belief trackers provide also a certain robustness towards variations on the input which helps to handle more noisy speech recognition input [20, 8].

The full model includes three trackers for *informable* slots (*food, pricerange, area*) and seven trackers for *requestable* slots (*address, phone, postcode, name*, and the three informable slots). *Informable* slots can be used to constrain a search, for instance by *food type* or *area*, while users can ask for a value

for *requestable* slots, such as *address*. Using the slot-based belief trackers introduces a set of intermediate labels which is a crucial difference to a pure end-to-end system [8].

After re-training the E2E-system with the new corpus the performance of the resulting model can be assessed by a corpus-based evaluation, i.e. the model is used to predict each system response in the test set left out of the training.

While this evaluation provides an indication of the models performance it also needs to be mentioned that many factors can influence the results, starting from the type of the additional training data in the new training corpus up to the version number of helper applications used in the training software package as mentioned in the documentation included in [21]. Therefore, results are not directly comparable with the benchmark results provided in [21] achieved on the 676 dialogues corpus available here: https://www.repository.cam.ac.uk/handle/1810/260970.

The joint performance of the trackers for informable and requestable slots after re-training with the 1200 dialogues corpus is shown in Table 2.

While performance of the informable slots is lower than the benchmark tracker performance in [21] the performance for requestable slots is very close to the benchmark result.

Table 2: *Tracker performance in the end-to-end system.*

Tracker type	Prec.	Recall	$\mathbf{F_1}$
Informable	97.39%	92.64%	94.95%
Requestable	98.07%	93.73%	95.85%

For the E2E-SDS the model with attention-mechanism [22, 23, 8] and weighted decoding [24] was used, similar to the system set-up used in [8] for the human evaluation.

For finding the best model a search over the initial learning rates 0.002 – 0.012 was conducted and the model with an initial rate of 0.004 resulted in the best performance in the corpus-based evaluation of 78.10% success rate and a BLEU-score of 0.2582. All other settings were left as provided by the exemplar configuration file provided in the model distribution available at [21]. It should be pointed out, that this task success rate is not directly comparable with the task success rate for the BASE-SDS, because the latter is calculated by a user simulator while the E2E-system compares with real human generated data.

3. Evaluation by human interaction

To evaluate the performance of the end-to-end system it was compared with the traditional modular system by humans interacting with the dialogue systems.

The comparison was conducted for both dialogue scenarios, i.e. text-in/out and speech-in/out, enabling one to identify any differences in the two use cases. For both scenarios, subjects were asked to interact with the systems given the general goal of finding a restaurant in Cambridge, UK, but following some specific tasks presented to them as follows:

> *Task: You want to find a restaurant in the centre which serves mediterranean food. You don't care about the price range. You want to know the name and address of the restaurant.*

Subjects were free to choose their own words to fulfil their tasks. Each subject had to do five tasks in one of the dialogue systems, then switch to the other system to complete another five tasks. Subjects were told they are dealing with two systems, but no information was given to what extend they differed and on the surface the systems looked identical. The starting order of systems was changed from subject to subject, i.e. subject1 would start with the BASE-SDS, while subject2 would start with the E2E-SDS, and so on. Subjects were asked to provide their feedback after each task was finished, either successfully or un-successfully, by answering the following three questions:

1. How successful would you rate the dialogue? (Success)
2. How natural do you rate the system response? (Naturalness)
3. How would you rate the comprehension of the system? (Comprehension)

Each rating category used a 5-point scale as follows: 1-Bad 2-Poor 3-Average 4-Nice 5-Excellent. For the first question, the rating of dialogue success, subjects were instructed to rate a "5" when the dialogue system provided all the required information, i.e. *name* and *address* of the restaurant in the example above, to choose any score between five and one when only part of the information was provided and to select "1" when they could not get the required information. The goal could range from getting a single item ("name of restaurant") up to the full range of items available: *name, address, postcode, phone number, post code*. After subjects finished their tasks they were asked which system they preferred.

Tasks were created by a task generator doing random generation from the available values in the database. The verbatim description of the task and its goals was done by hand.

All evaluations were conducted at the Toshiba Cambridge Research Laboratory. Subjects were mainly recruited from PhD students not working in the field of speech technology. All subjects were rewarded for their participation.

3.1. Evaluation of text-in/out scenario

First, the text-interface was evaluated to set the baseline performance of the systems without speech interface. 10 subjects participated in the evaluation. Subjects were asked to interact with the systems by typing their inputs and feedback by the systems was also provided via text. Since the training material was also produced by a text-in/out setup, this evaluation should reveal how the systems compare in terms of their ability to understand user inputs and their capacity to generate natural responses.

Results are shown in Table 3 below for the BASE-DS and the E2E-DS as average scores across subjects and dialogues. The numbers in round brackets represent standard deviations.

Table 3: *Mean opinion scores and standard deviations in brackets for comparison of BASE-DS vs. E2E-DS by textual interaction, i.e. typing. All ratings are on a 1–5 scale.*

Metric	BASE-DS	E2E-DS
Success	3.34 (1.71)	4.72 (0.75)
Naturalness	3.02 (1.23)	4.16 (0.93)
Comprehension	2.86 (1.40)	4.38 (0.98)
Average # of turns	6.51	4.79
# dialogues	50	50

Results show, that the E2E-DS is consistently rated higher on all metrics, especially in terms of perceived comprehension

capabilities. The BASE-DS also has a wider range of ratings reflected in higher standard deviation values, indicating that subjects made more mixed experiences, i.e. some dialogues which were successful and smooth but also a number of dialogues with problems and/or ending unsuccessfully. nine out of ten subjects expressed their preference for the E2E-system, one subject did not have a preference. The E2E-system allowed subjects to reach their goals faster (lower number of average turns), while getting a more natural response and feeling better understood.

3.2. Evaluation of speech-in/out scenario

For the speech-in/out evaluation, subjects listened and talked to the system via a closed-ear headset (Beyerdynamic MMX 300) and the output of the speech recognition engine was shown on the computer screen as feedback. The evaluation was conducted in a regular office room with the experimenter and the subject present. 15 subjects participated in the evaluation.

In order to provide some feedback to the user with respect to the speech recognition system the subjects were shown the following prompts when the speech recognition was listening, when processing happened and also the output of the recognition after the "You said" prompt.

InfoDesk: Say something!
InfoDesk: Got it, recognising...
InfoDesk: You said: ...

Given that information, subjects could judge whether the speech recognition worked or whether the dialogue system was not responding to their input as expected. Otherwise the set-up of the experiment was identical to the set-up used in the text-in/out evaluation. Subjects could also see the text of the dialogue system response on the screen in addition to the spoken version they heard over the headphones.

The results are shown in Table 4 and generally resemble the results found in the text-only evaluation, i.e. the E2E-SDS is consistently rated higher than the baseline system on all metrics.

Comparing the results from the text-in/out scenario with the ones from the speech-in/out scenario shows, that ratings are generally slightly lower in the speech-in/out case (except for comprehension in the BASE-SDS which actually was rated slightly better), showing the impact of the more noisy speech interface, i.e. some recognition problems on the input side and also a few synthesis issues on the output side.

Table 4: *Mean opinion scores and standard deviations in brackets for comparison of BASE-SDS vs. E2E-SDS by spoken interaction. All ratings are on a 1–5 scale.*

Metric	BASE-SDS	E2E-SDS
Success	3.17 (1.73)	4.40 (0.98)
Naturalness	3.18 (1.30)	4.04 (0.97)
Comprehension	2.98 (1.71)	4.10 (1.15)
Average # of turns	6.96	5.25
# dialogues	75	75

Thirteen subjects preferred the E2E-system, two subjects did not declare a preference, i.e. the majority of subjects expressing their preference for the E2E-SDS. Subjects mentioned that they felt better understood by the E2E-SDS, achieved their goals faster, again reflected by a lower number of average turns, and felt less restricted or forced to follow a certain conversation scheme as in the BASE-SDS.

4. Discussion

One of the aspects very much disliked by subjects, especially with the BASE-SDS, was the inability to grasp the meaning of an input even though it was written or recognized correctly. The BASE-SDS was often described as too schematic by following a strict scheme of answering single questions to tick off some boxes for internal entities ('slot-filling' approach), while the end-to-end system appeared to be less schematic and providing more variation in its responses.

Looking back at the results of the two evaluations shows, that the E2E-SDS is clearly preferred over the traditional baseline system by achieving a better subjective success rate, providing a more natural system response as well as a better comprehension.

The E2E-SDS achieves these advantages by generating less schematic system responses and by using a policy that does not need to go through a series of 'slot-filling' questions until a potentially user satisfying output can be generated. Table 5 provides an example of a typical dialogue with the E2E-SDS.

Table 5: *Example of a dialogue with the E2E-SDS.*

InfoDesk:	Welcome to the Cambridge restaurant information system. How may I help you?
Customer:	*I want an expensive british restaurant*
InfoDesk:	There are 3 restaurants serving british kind of food in the expensive price range, do you have a preference on side of town?
Customer:	*in the west part*
InfoDesk:	Graffiti is a british restaurant in the west area of town.
Customer:	*what is the phone number*
InfoDesk:	The phone is 01223 277977.
Customer:	*thanks bye*
InfoDesk:	Thank you, goodbye

Regarding the metrics used to assess the performance of dialogue models objectively it can be said that especially the BLEU-score is a poor predictor of performance when it comes to human interaction with the systems. An observation also addressed by Liu et al. [25]. The human evaluations conducted in this work, show that a rather low BLEU-score does not tell much about the performance of the model with real users, since there are multiple ways of a good system response which are just not reflected in the BLEU-score.

5. Conclusions

A comparison of a traditional, modular dialogue system with an end-to-end trainable dialogue system was conducted with human evaluators. Subjects interacted with the task-oriented dialogue systems and aimed to find restaurants in Cambridge, UK fulfilling certain requirements. Subjects provided their impressions of the dialogues in terms of subjective success, naturalness of response and comprehension of the systems. The comparison was conducted in two scenarios, i.e. text-in/out and speech-in/out. The results show that, in both scenarios, the end-to-end system is preferred on all metrics indicating that it enables users to reach their goals faster and experience both a more natural system response as well as a better comprehension by the dialogue system.

Future work will address the issue of scalability of the end-to-end architecture to more complex domains.

6. References

[1] O. Vinyals and Q. V. Le, "A neural conversational model," in *ICML Deep Learning Workshop*, Lille, France, 2015.

[2] L. Shang, Z. Lu, and H. Li, "Neural responding machine for short-text conversation," in *ACL*, Beijing, China, 2015, pp. 1577–1586.

[3] I. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Building end-to-end dialogue systems using generative hierarchical neural network models," in *AAAI*, 2016.

[4] R. Lowe, N. Pow, I. Serban, L. Charlin, C.-W. Liu, and J. Pineau, "Training end-to-end dialogue systems with the Ubuntu Dialogue Corpus," *Dialogue & Discourse*, vol. 8(1), pp. 31–65, 2017.

[5] A. Bordes, Y.-L. Boureau, and J. Weston, "Learning end-to-end goal-oriented dialog," in *ICLR*, Toulon, France, 2017.

[6] I. Serban, A. Sordoni, Y. Bengio, A. C. Courville, and J. Pineau, "Hierarchical neural network generative models for movie dialogues," 2015. [Online]. Available: https://arxiv.org/pdf/1507.04808v1.pdf

[7] A. Ritter, C. Cherry, and W. B. Dolan, "Data-driven response generation in social media," in *Proceedings of the Conference on Empirical Methods in Natural Language Processing*, ser. EMNLP '11, Stroudsburg, PA, USA, 2011, pp. 583–593. [Online]. Available: http://dl.acm.org/citation.cfm?id=2145432.2145500

[8] T.-H. Wen, D. Vandyke, N. Mrkšić, M. Gasic, L. M. Rojas Barahona, P.-H. Su, S. Ultes, and S. Young, "A network-based end-to-end trainable task-oriented dialogue system," in *EACL*, Valencia, Spain, April 2017, pp. 438–449. [Online]. Available: http://www.aclweb.org/anthology/E17-1042

[9] A. Zhang, "Python module for speech recognition (version 3.7.1) [software]," *Available from https://github.com/Uberi/speech_recognition*, 2017.

[10] H. Pham, "Python module for audio-in/out (version 0.2.11 [software]," *Available from https://people.csail.mit.edu/hubert/pyaudio/*, 2017.

[11] Google, "Google speech recognition API (version 2) [software]," *Accessible at http://www.google.com/speech-api/v2/recognize?*{}, 2017.

[12] S. Buchholz, N. Braunschweiler, M. Morita, and G. Webster, "The Toshiba entry for the Blizzard Challenge 2007," in *Proc. of 6th ISCA Speech Synthesis Workshop*, Bonn, Germany, 2007, pp. 264–269.

[13] N. Mrkšić, D. Ó Séaghdha, T.-H. Wen, B. Thomson, and S. Young, "The neural belief tracker: Data-driven dialogue state tracking," in *Proceedings of ACL*, Vancouver, Canada, 2017.

[14] S. Ultes, L. M. Rojas Barahona, P.-H. Su, D. Vandyke, D. Kim, I. Casanueva, P. Budzianowski, N. Mrkšić, T.-H. Wen, M. Gasic, and S. Young, "PyDial: A Multi-domain Statistical Dialogue System Toolkit," in *Proceedings of ACL 2017, System Demonstrations*, Vancouver, Canada, July 2017, pp. 73–78. [Online]. Available: http://aclweb.org/anthology/P17-4013

[15] A. Papangelis and Y. Stylianou, "Single-model multi-domain dialogue management with deep learning," in *Proceedings of International Workshop on Spoken Dialogue Systems*, Farmington, USA, 2017.

[16] Z. Wang, T.-H. Wen, P.-H. Su, and Y. Stylianou, "Learning domain-independent dialogue policies via ontology parameterisation," in *16th Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGdial)*, 2015, pp. 412–416.

[17] A. Papangelis and Y. Stylianou, "Multi-domain spoken dialogue systems using domain-independent parameterisation," in *Proceedings of Domain Adaptation for Dialogue Agents*, Riva del Garda, Italy, 2016.

[18] M. Henderson, B. Thomson, and J. Williams, "The second dialog state tracking challenge," in *Proceedings of Special Interest Group on Discourse and Dialogue (SIGdial)*, Philadelphia, USA, 2014.

[19] S. Hochreiter and J. Schmidhuber, "Long short-term memory," *Neural Comput.*, vol. 9, no. 8, pp. 1735–1780, Nov. 1997. [Online]. Available: http://dx.doi.org/10.1162/neco.1997.9.8.1735

[20] M. Henderson, "Machine learning for dialog state tracking: A review," in *Proceedings of The First International Workshop on Machine Learning in Spoken Language Processing (MLSLP)*, Aizuwakumatsu, Japan, 2015.

[21] T.-H. Wen, "NNDial - open source toolkit for building end-to-end trainable task-oriented dialogue models [software]," *Available from https://github.com/shawnwun/NNDIAL*, 2017.

[22] D. Bahdanau, K. Cho, and Y. Bengio, "Neural machine translation by jointly learning to align and translate," *CoRR*, vol. abs/1409.0473, 2014. [Online]. Available: http://arxiv.org/abs/1409.0473

[23] K. M. Hermann, T. Kociský, E. Grefenstette, L. Espeholt, W. Kay, M. Suleyman, and P. Blunsom, "Teaching machines to read and comprehend," *CoRR*, vol. abs/1506.03340, 2015. [Online]. Available: http://arxiv.org/abs/1506.03340

[24] T.-H. Wen, M. Gasic, N. Mrkšić, L. M. Rojas Barahona, P.-H. Su, S. Ultes, D. Vandyke, and S. Young, "Conditional generation and snapshot learning in neural dialogue systems," in *EMNLP*. Austin, Texas: ACL, November 2016, pp. 2153–2162.

[25] C.-W. Liu, R. Lowe, I. Serban, M. Noseworthy, L. Charlin, and J. Pineau, "How NOT to evaluate your dialogue system: An empirical study of unsupervised evaluation metrics for dialogue response generation," in *Proc. of Annual Meeting of the Association for Computational Linguistics (ACL)*, 03 2016, p. 13.