

Adapting Document-Grounded Dialog Systems to Spoken Conversations using Data Augmentation and a Noisy Channel Model

David Thulke,^{*1,2} Nico Daheim,^{*1} Christian Dugast,² Hermann Ney^{1,2}

¹ Human Language Technology and Pattern Recognition Group, RWTH Aachen University, Germany

² AppTek GmbH, Aachen, Germany

{thulke, daheim, ney}@i6.informatik.rwth-aachen.de
cdugast@apptek.com

Abstract

This paper summarizes our submission to Task 2 of the second track of the 10th Dialog System Technology Challenge (DSTC10) “Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations”. Similar to the previous year’s iteration, the task consists of three subtasks: detecting whether a turn is knowledge seeking, selecting the relevant knowledge document and finally generating a grounded response. This year, the focus lies on adapting the system to noisy ASR transcripts. We explore different approaches to make the models more robust to this type of input and to adapt the generated responses to the style of spoken conversations. For the latter, we get the best results with a noisy channel model that additionally reduces the number of short and generic responses. Our best system achieved the 1st rank in the automatic and the 3rd rank in the human evaluation of the challenge.

1 Motivation

While research on document-grounded dialog systems has been focusing on written dialogs, many of their applications, such as voice assistants, are within the domain of spoken conversations. However, the two domains are fundamentally different. First of all, spoken conversations are more spontaneous and may include interruptions, repetitions, corrections, and other disfluencies. If the speech was transcribed automatically, errors propagated from automatic speech recognition (ASR) systems introduce further challenges. Hence, it is not clear a priori whether a dialog system trained on written data would also fit spoken data sufficiently well. Indeed, Gopalakrishnan et al. (2020) and Kim et al. (2021b) show that the performance of existing models trained on written conversations strongly degrades when evaluated on spoken data.

The second track of the 10th Dialog System Technology Challenge (DSTC10) hosts a shared task on “Knowledge-grounded Task-oriented Dialogue Modeling on Spoken Conversations”. In continuation of the first track of the DSTC9 challenge (Kim et al. 2021a), three subtasks are proposed. First, the dialog system has to identify user turns that seek knowledge not defined by an API and only available in the form of unstructured documents. This task is called Knowledge-seeking Turn *Detection*. Next, the system has

to retrieve a document and generate a response based on this document and the dialog context. These tasks are called *Knowledge Selection* and *Response Generation*, respectively. To address the aforementioned issues, this time evaluation is only done on automatic transcripts of spoken conversations including ASR errors to test the robustness of dialog models. Only a small validation set of spoken conversations is provided to develop models.

We explore different methods to adapt document-grounded dialog systems to spoken conversations. For the detection and selection subtasks, we explore simple text pre-processing techniques applied to the training data in order to make the models more robust to ASR transcripts. In addition, we make use of the provided n-best lists to capture the uncertainties of the transcriptions. For the generation subtask, we additionally focus on adapting the style of generated responses so that they are more naturally connected to the dialog. Therefore, we experiment with a few-shot transfer using the DSTC10 validation data and explore approaches to integrate an external (ungrounded) response generation model trained on spoken conversations. One of these approaches is a Noisy Channel reformulation of the generation task. In addition to the integration of the external model, this formulation allows penalizing short and generic responses that are often generated by the baseline model. We hypothesize that this increases the faithfulness of generated responses.

2 Task Description

The DSTC10 challenge does not provide a dedicated training set. Rather, teams were encouraged to use the training set of the DSTC9 challenge consisting of 72,518 dialogs in written style. This set is an augmented version of MultiWOZ 2.1 (Eric et al. 2020) where turns were added that are grounded in external knowledge documents as opposed to the existing API-based turns. Documents were collected from FAQ pages and consist of a question-answer pair. Thus, they are relatively short. They span the four different domains: hotel, restaurant, train, and taxi. The first two are divided further into entities. In the following we refer to the set of documents as knowledge base. While the validation set makes use of the same documents and locality as the training set, the test set introduced a new locality - San Francisco - and unseen documents, some of which stem from a

^{*}These authors contributed equally.

new domain called attraction. Around half of the 4,181 dialogs from the test set were obtained again by augmenting the MultiWOZ dataset. The other half was collected from human-to-human conversations covering touristic information in San Francisco. Around a tenth of these conversations are from the spoken domain and thus have a similar style as the DSTC10 data. However, the DSTC9 test data does not include ASR transcripts but rather human transcripts without errors. The validation data of DSTC10 are the same 263 dialogs as the spoken part of the DSTC9 test data but transcribed with an ASR system. The test data consists of 1,988 additional dialogs collected with the same locality. The ASR system that was used to transcribe the validation and test data is described in Kim et al. (2021b). The system achieved a word error rate of 24.09% resulting in strongly perturbed outputs. The knowledge base is the same as in the DSTC9 test set. For more details on the datasets, we refer to Kim et al. (2021b).

Accordingly, the tasks of this challenge are also the same as the ones of its predecessor. The three tasks are defined as follows:

Detection In Knowledge-seeking Turn Detection, the system has to decide whether the current user turn is covered by the given API or whether it requires unstructured knowledge access, i.e. one of the FAQ documents contains the required information.

Selection In Knowledge Selection, the system has to find the FAQ document K' that answers the last knowledge-seeking user turn u_T .

Generation Finally, response generation is the task of generating an appropriate agent response u_{T+1} based on the selected knowledge document K' and dialog context u_1^T .

3 Methods

3.1 Text Preprocessing

In contrast to the written training data, the ASR transcripts in the DSTC10 validation and test data are lower-cased and do not contain punctuation. This creates a mismatch between the training and evaluation data. One approach to solve this issue is to augment the ASR transcripts with punctuation and casing information. The downside of this approach is that external modules for these tasks are required that may introduce further errors in the pipeline.

An alternative approach is to remove this information from the written text so that it becomes more similar to the ASR transcripts. In addition to that, we write out numbers (e.g. $42 \mapsto \text{forty two}$) and spell out abbreviations (e.g. $mm \mapsto \text{millimeters}$).

3.2 Detection

As in the baseline model proposed by Kim et al. (2020), we model knowledge detection as a binary classification task. Similar to previous work (Mi et al. 2021; Jin, Kim, and Hakkani-Tur 2021), we fine-tune RoBERTa-large (Liu et al. 2019). Therefore, we add a simple classifier consisting of two linear layers on top of the first hidden state (corresponding to the begin of sentence token). To limit the length of the

input, we only pass the last three utterances to the model and additionally truncate the input sequence if it exceeds 384 tokens.

To adapt the model to the knowledge documents from the new domains and localities, we generate additional knowledge-seeking dialog samples based on the documents in our knowledge base. Therefore, for each document, we randomly select one dialog from the original MultiWOZ corpus in the same domain, replace the entity in the document with an entity from the dialog, and add the questions of the (faq) document as a new knowledge-seeking turn. This way, we add 16,675 new samples to the training data.

To make use of the n-best list provided by the ASR system, we pass each ASR hypothesis to our model and experiment with two different strategies. The *best* strategy just selects the highest score of all hypotheses and the *weighted* strategy calculates the weighted sum of all scores based on the (renormalized) probabilities of the ASR hypotheses. Even though the *weighted* strategy is the mathematically more sound option, as it treats the ASR hypotheses as a latent variable, we observe the highest F1 scores on the validation data with the *best* strategy.

For our final model, we use an ensemble of different model variants and fine-tune the decision threshold on the DSTC10 validation data.

3.3 Selection

The goal of the selection subtask is to find the most relevant document from the knowledge base given a dialog. The baseline system models this as a relevance classification task, i.e. for each pair the model predicts whether it is relevant or not and the document with the highest relevance score is chosen. As this requires a full pass through the model for each pair of context and document, the method becomes increasingly inefficient for larger knowledge bases. To avoid this, Thulke et al. (2021) propose a Hierarchical Selection model that is identified as a good tradeoff between efficiency and performance. To reduce the search space, they first use an entity model p_E to identify the most relevant domain and entity and then another document model p_D to identify the most relevant document of the selected entity. In the following, we discuss the extensions we made to this approach.

First, since the entity selection task is similar to other task-oriented dialog tasks, it allows us to make use of additional training data. Therefore, we collect those dialogs whose states contain entities of the relevant domains from Taskmaster-2¹ (Byrne et al. 2019) and the DSTC10 Track 2 Task 1 validation data (Kim et al. 2021b). In total 4,499 new training instances were added.

To train both models, we provide the reference entity/document as a positive sample and sample three negative samples in each epoch. For the entity selection model, we randomly sample one entity with a different domain and two entities with the same domain as the negative samples. For

¹<https://github.com/google-research-datasets/Taskmaster/tree/master/TM-2-2020>

the document selection model, we sample three documents of the same entity as negative samples.

In the original variant proposed by Thulke et al., a greedy search method was used. Specifically, during the document selection, only documents of the most relevant entity are considered. Instead, we propose to consider all entities whose entity relevance score $p_E(r|e, u_1^T)$ is within a threshold $t \leq 1$ of the most relevant entity \hat{e} :

$$\hat{k} = \arg \max_{\substack{k=(e,d) \\ p(e|u_1^T) > t \cdot p(\hat{e}|u_1^T)}} p_E(e | u_1^T)^\gamma \cdot p_D(d | e, u_1^T) \quad (1)$$

Additionally, we add a scaling factor γ to control the influence of both models on the final selection.

To make use of the ASR n-best lists, we experimented with the same strategies as for the detection task. Finally, for the DSTC10 validation and test data, we can further reduce the search space by only considering entities from the San Francisco locality.

3.4 Generation

To maintain a fluent conversation, generated responses should be naturally connected to the context of the dialog and thus match the style of preceding utterances. Hence, we study different methods to encourage the model to generate answers in spoken style with either no or only a few in-domain samples. In the following let $w := u_{T+1}$ denote the generated response and w_n its n-th token. An intuitive method is to train the standard model

$$p(w_n | w_1^{n-1}, u_1^T, K'),$$

to which we will refer as *direct model (dm)* in the following, on additional spoken dialogs. While the model could infer the style of the dialog already from the context u_1^T , we further introduce a style token as a form of explicit conditioning. Then, the model becomes

$$p(w_n | w_1^{n-1}, u_1^T, K', s),$$

where s is a special $\langle \text{written} \rangle$ or $\langle \text{spoken} \rangle$ token that is added to the vocabulary. Since only a few samples from the DSTC10 validation set are available to train this model, we seek methods that facilitate domain adaptation without requiring grounded dialogs, as such data is more readily available. First, we try out a *shallow fusion* (Gulcehre et al. 2015), i.e. a log-linear model combination, of the direct model and an *ungrounded* response generation model (*lm*) trained on spoken conversations. We combine the models locally such that the response is generated according to

$$p(w_n | w_1^{n-1}, u_1^T, K') \propto p_{\text{dm}}(w_n | w_1^{n-1}, u_1^T, K') \cdot p_{\text{lm}}(w_n | w_1^{n-1}, u_1^T)^\lambda,$$

where λ is a scaling factor that is tuned on the validation data. Further, to reduce the influence of the written style from the training data, we experiment with subtracting the output of a response generation model trained only on the written in-domain data without document grounding (*ilm*). Here, we draw inspiration from the Density Ratio method

to external language model integration in automatic speech recognition (McDermott, Sak, and Variani 2019) and hence refer to the model by this name. This results in the following model combination:

$$p(w_n | w_1^{n-1}, u_1^T, K') \propto p_{\text{dm}}(w_n | w_1^{n-1}, u_1^T, K') \cdot p_{\text{lm}}(w_n | w_1^{n-1}, u_1^T)^{\lambda_1} \cdot p_{\text{ilm}}(w_n | w_1^{n-1}, u_1^T)^{-\lambda_2}$$

The advantage of these two approaches is that the ungrounded response generation model can be trained on a large amount of dialog data without document grounding which allows a zero-shot domain transfer. At the same time, however, this model might skew the distribution towards degenerate words that introduce inconsistencies w.r.t. the grounding if the scaling factor is too large and not influence the style sufficiently if it is too low.

Noisy Channel formulation Since it is not clear how shallow fusion influences the faithfulness of the generated responses, we seek to find a model that explicitly enforces faithfulness to document grounding. Hence, we use Bayes Theorem to derive a noisy channel formulation for document-grounded response generation as follows:

$$\begin{aligned} & \arg \max_w p(w | u_1^T, K') \\ &= \arg \max_w p(w, u_1^T, K') \\ &= \arg \max_w \underbrace{p(K' | w, u_1^T)}_{\text{channel model}} \cdot \underbrace{p(w | u_1^T)}_{\text{response generation model}} \end{aligned}$$

First of all, we can see that the advantage of having an ungrounded response generation model which can be trained on large amounts of textual data in the new domain without requiring document grounding is retained. Furthermore, the channel model now encourages that the response explains the document grounding sufficiently well which could prevent the model from leaving out important details and mitigate the explaining-away effect (Liu et al. 2021b).

However, decoding the noisy channel model directly is computationally intractable. Hence, we use two different approximate decoding methods. First of all, we experiment with *reranking* generations obtained from a proposal model, for which we use the direct model. That is, we first decode k sequences from the direct model and then obtain the final response as the highest scoring sequence under the log-linear model combination

$$\hat{w} = \arg \max_w \log p(w | u_1^T, K') + \lambda_2 \cdot \log p(K' | w, u_1^T) + \lambda_1 \cdot \log p(w | u_1^T) \quad (2)$$

We interpolate with the direct model to encourage sequences with high direct model likelihood which has proven beneficial in other tasks (Yu et al. 2017; Liu et al. 2021b). While comparatively efficient, the method is limited by the proposal model, since the noisy channel formulation can only re-rank an n-best list of complete sequences. For example, if

Table 1: Effect of different text preprocessing techniques in the detection task on the DSTC10 validation data.

method	F1
baseline (RoBERTa-large)	75.3
+ lowercasing	78.4
+ no punct.	79.7
+ numbers written out	83.7
+ no abbrev.	84.1

all sequences contain false information, this error can not be corrected by the noisy channel model.

Therefore, we also derive a simple *online decoding* algorithm that allows an early interaction between language and channel model. At each timestep, the algorithm first decodes k beams obtained from the direct model for each hypothesis on the beam. Then, the beams are rescored using the log-linear model combination in Equation 2. Since online decoding is based on partial sequences, we also need to ensure that the channel model is trained on partial inputs to not create a mismatch between train and test time. Hence, as in Liu et al. (2021b), we truncate w according to a uniform distribution over all lengths from 1 up to sequence length during training.

4 Experiments

The experiments have been done using HuggingFace Transformers (Wolf et al. 2020), HuggingFace Datasets (Lhoest et al. 2021), and Sisyphus (Peter, Beck, and Ney 2018)². All models were trained on Nvidia GTX 1080 Ti or RTX 2080 Ti GPUs. For the evaluation, we use the same metrics as proposed by Kim et al. (2021a) and that are used for the final ranking. In the selection and generation subtasks which depend on the results of previous tasks, we evaluate the methods on the ground truth labels to facilitate comparability.

4.1 Text Preprocessing

We experimented with the different proposed text processing strategies on the detection task. Table 1 shows the results on the DSTC10 test data. We observed that each method gives a slight improvement in final performance. Therefore, we decided to apply these pre-processing methods in the detection and selection tasks.

4.2 Detection

Table 2 shows the results of our proposed methods for the detection task on the DSTC10 validation and test data. First, augmenting the training data with additional samples generated from the knowledge base gave us a strong improvement on the validation and a small improvement on the test data. The additional, in-domain pretraining of the RoBERTa model further improved the results by 1%. Next, we experimented with the two proposed ASR n-best strategies and observed better results with the max strategy. Finally, we included the DSTC9 test and DSTC10 validation data into the

Table 2: F1 scores of the detection subtask on the DSTC10 validation and test data.

model	val	test
baseline (+ text preprocessing)	84.8	84.1
+ data augmentation	91.9	85.1
+ in-domain pretraining	93.5	86.0
+ ASR n-best (weighted)	94.5	86.5
+ ASR n-best (max)	94.7	87.7
+ DSTC9 test + DSTC10 val	-	90.5
+ ensemble	-	91.1

Table 3: R@1 scores of the selection subtask on the DSTC10 validation and test data.

model	val	test
baseline (+ text preprocessing)	71.2	70.0
+ Beam Search	74.0	73.5
+ Taskmaster & DSTC10 data	78.8	76.3
+ in-domain pretraining	83.7	77.0
+ ASR n-best (max)	79.8	77.7
+ ASR n-best (weighted)	81.7	77.7
+ DSTC9 test + DSTC10 val	-	77.3
+ ensemble	-	77.6

training of the model and created an ensemble of different training runs.

4.3 Selection

Table 3 shows the results of applying our proposed methods for the selection task on the DSTC10 validation and test data. Using our proposed Beam Search approach instead of always taking the entity with the highest score results in an improvement of around 3% absolute. Further, training the domain and entity selection model on additional data from Taskmaster and DSTC10 Task 1 gives an additional improvement. On the validation data, we observed slight degradations with our strategies to handle the ASR n-best list. On the test set, this resulted in improvements. We assume that the observed degradations can be attributed to the small size of the validation set. In contrast to the detection task, including the DSTC9 test and DSTC10 validation data resulted in small degradations. Finally, an ensemble of different training runs slightly improved the results again.

4.4 Generation

Table 4: Zero-shot transfer on the DSTC10 validation set.

Method	BLEU-1	METEOR	ROUGE-L
Direct model	21.1	24.4	19.6
+ Shallow Fusion	20.6	25.5	22.9
+ Density Ratio	21.2	26.0	22.9
+ Noisy Channel _{Onl.}	22.5	24.6	22.3

Similar to previous work (Thulke et al. 2021), we fine-tuned BART large for all sequence-to-sequence models discussed

²Code is available at github.com/dthulke/dstc10-track2

Table 5: Trained on DSTC9 train and 10-fold cross-validation on DSTC10 val.

Method	BLEU-1	METEOR	ROUGE-L
Direct Model	40.6	49.0	44.2
+ Noisy Channel	41.9	50.7	46.0
+ Style Token	42.3	50.2	45.7
+ Noisy Channel _{Onl.}	42.5	51.3	46.2

in this section. Furthermore, the ungrounded response generation model is always trained on the spoken two-person subset of Taskmaster-1 (Byrne et al. 2019), Taskmaster-2, CCPE (Radlinski et al. 2019), as well as the indicated DSTC splits. We study different settings for the generation task. In the first setting, we train the models only on the written dialog training set from the DSTC9 challenge. Then, we perform a zero-shot evaluation concerning the spoken domain dialogs of the DSTC10 validation set. Table 4 shows results obtained with different models in this setting. We can see that both the shallow fusion and density ratio approach give slight improvements. By using noisy channel online decoding with the direct model as proposal model we obtain similar results. However, these methods require substantial additional computational effort and the increased number of parameters from combining multiple models might also contribute to the improvements.

In general, we note that the generations in the zero-shot setting are substantially different from the references on the spoken test set and particularly contain follow-up questions such as "Do you need anything else?", which are not contained in the responses of the test data. We have also experimented with removing these phrases in the training data which however brought degradations in terms of automatic metrics, mainly due to the included length penalty. The test data contains different rather colloquial phrases such as the prefix "sure let me check that for you".

To also adapt the style of these phrases in the response, we experiment with training our model on the DSTC10 validation data. We split the data in 10 cross validation splits to still be able to evaluate the resulting models on the full validation set. The results are highlighted in Table 5. We can see that using a style token gives good improvements. Furthermore, the noisy channel model with online decoding gives further improvements for both proposal models.

Table 6: Trained on DSTC9 train and DSTC10 val, evaluated on DSTC10 test.

Method	BLEU-1	METEOR	ROUGE-L
Direct Model	43.1	49.8	48.0
+ Shallow Fusion	43.4	50.1	48.2
+ Style Token	43.8	50.3	48.7
+ Noisy Channel _{Re.}	44.3	51.2	48.8
+ Noisy Channel _{Onl.}	45.7	52.0	50.2

Finally, Table 6 shows our results on the DSTC10 test data. All models were trained on the DSTC9 training data

as well as both validation sets of DSTC9 and 10, i.e. a small amount of in-domain training data containing ASR errors and spontaneous speech. In this setting, there is only a minor improvement by interpolating with a spoken response generation model. Explicit conditioning on the style using a style token gives stronger improvements without necessitating the training of an additional model and is thus our preferred method for obtaining proposals for the noisy channel model which shows improvements with both decoding strategies. The online decoding method forms our best model and our submission to the shared task. This indicates that the early interaction of response generation and channel model can indeed improve performance and at the same time that the reranking approach is too limited by the proposal model. We discuss the improvements observed by the model in Section 4.6. We, however, need to note that the model has more parameters than the direct model such that some of the improvements might be attributed to an increase in modeling capacity. Nevertheless, as shown by Liu et al. (2021b), noisy channel formulations can still outperform the direct modeling approach with the same amount of parameters for task-oriented dialog. Hence, we do not study this particular aspect in our paper.

4.5 Challenge Results

Table 7 shows the official results of DSTC10 Track 2 Task 2 of the best five teams according to the human evaluation. The baseline system is the original baseline system proposed by Kim et al. (2020) for DSTC9. In contrast to previous experiments, models for the selection and generation subtasks are applied to the outputs of the previous tasks. In total, 16 teams participated in the challenge.

Automatic Evaluation Our best system achieved 4th place in the detection (F1) and selection (R@1) subtasks and 1st place in the generation subtask. In the official ranking according to the automatic metrics, our system achieved 1st place as the generation metrics have a higher influence on the final ranking than the other tasks.

Human Evaluation For the 8 best systems according to the automatic evaluation an additional human evaluation was conducted by the task organizers. Crowd workers were asked to rate the accuracy and appropriateness of a system response on a scale from 1 to 5. For the former, they had access to the reference knowledge document and should rate the accuracy of the reference response. To rate the appropriateness, they only had access to the dialog context and should rate how well the response fits into the dialog. As in the previous year’s evaluation (Kim et al. 2021a), the results in the selection task have the strongest correlation to human judgments. Additionally, we observe that there is no apparent correlation between the automatic scores in the generation task and the human judgements. For example, Team 10 ranks first in the human evaluation even though they have the lowest scores in the automatic generation metrics among the top five teams. One possible explanation is that crowd workers did not prefer responses in spoken style, even though they should fit more naturally in the spoken conversations.

Table 7: Final results of the top 5 out of 16 teams ranked according to the human evaluation.

	detection	selection	generation			human evaluation	
	F1	R@1	BLEU-1	METEOR	ROUGE-L	Accuracy	Appropriateness
baseline	79.5	45.8	11.5	12.2	11.4	2.74	2.79
Team 10	92.3	79.3	16.2	21.0	21.9	3.49	3.35
Team 4	91.8	74.8	33.8	40.7	38.7	3.34	3.30
Team 8 (our)	91.1	71.0	40.1	46.0	44.1	3.34	3.26
Team 14	92.4	62.0	27.1	31.7	31.8	3.29	3.28
Team 2	90.4	69.3	37.3	43.9	41.2	3.29	3.22
ground truth						3.58	3.48

4.6 Error Analysis

Error analysis of the generation models shows different patterns. First of all, we observe that all models sometimes do not answer the previous user question but rather one in the dialog context. We hypothesize that the model can not make sense of the previous turn, also in combination with the knowledge document, and then generates a false response. This would suggest modeling only the previous user turn, which however did not result in improvements in terms of automatic metrics in our initial experiments. Furthermore, the model often repeats ASR errors, especially when an entity is not recognized, and generates an answer containing the false transcription such as in the last row of Table 8 as an example, where both models repeat the word "scoffer". When comparing the direct model with the noisy channel formulation, we can see that the direct model sometimes leaves out important details which the noisy channel model retains. For example, in the first row of Table 8, we can see that the former only mentions that smoking is allowed while the latter model points out designated areas. Furthermore, the direct model sometimes fails to interpret the document correctly, for example in line three where the direct model incorrectly says that no parking is available which the noisy channel method is able to identify correctly. It is also worth noting that such effects depend on the scaling factors of the noisy channel model. In our experiments, too much influence of the response generation model led to hallucinations and too much influence of the channel model led to the model mostly repeating the document.

5 Related Work

It has been shown that models trained on written dialogs generalize poorly towards spoken dialogs with ASR errors (Gopalakrishnan et al. 2020; Kim et al. 2021b). Thus, different approaches of increasing the robustness of Spoken Language Understanding models to ASR errors have been proposed. Traditionally, many systems have used confusion networks (Tur et al. 2002; Henderson et al. 2012) to consider more than one ASR hypothesis. A different method is to already include automatic speech recognition errors in the training data. For example, Schatzmann, Thomson, and Young (2007) introduce ASR errors based on an n-gram confusion matrix obtained after Levenshtein alignment. Fazel-Zarandi et al. (2019) propose to use vocoders to create spoken sequences from textual training samples, adding noise

and then transcribing the samples with an ASR model to obtain perturbed training data. Simpler transformations such as removing punctuation (Gopalakrishnan et al. 2020) have also been proposed. The inverse transformation of correcting ASR errors has also been explored, for example in Weng et al. (2020).

Document-grounded dialog datasets have so far focused on the written domain. For example, Wizard-of-Wikipedia (Dinan et al. 2018) consists of open-domain dialogs grounded by Wikipedia articles, and the doc2dial dataset (Feng et al. 2020) consists of task-oriented dialogs grounded by long documents. In this paper, we are concerned with a continuation of the task "Beyond Domain APIs: Task-oriented Conversational Modeling with Unstructured Knowledge Access" hosted at DSTC9 (Kim et al. 2021a) which contains dialogs grounded in FAQ documents. Thus, we review different methods introduced for the tasks of Turn Detection, Knowledge Selection, and Response Generation. He et al. (2021) propose to model the first task by deciding whether a knowledge document or schema description obtained from MultiWOZ is more likely to be sought by the user. In the first case, the most likely knowledge document is selected and in the latter case, the turn is deemed not knowledge-seeking. Furthermore, similar to Jin, Kim, and Hakkani-Tur (2021) the authors propose different strategies to sample negative training examples, such as sampling documents from the same domain or entity. Tang et al. (2021) propose to select negatives by first training a model on the selection task with negatives sampled from the same entity or domain and then taking the documents likely to be confused under the model as negatives to train a stronger model. While this forms an explicit negative sampling, Thulke et al. (2021) explore to fine-tune the selection models end-to-end with the response generation task by using a retrieval augmented model (Gua et al. 2020; Lewis et al. 2020b), where the marginalization can be seen as an implicit batching of hard negatives. Furthermore, the authors propose to use a hierarchical selection approach and formulate knowledge selection as a metric learning problem using bi-encoders, similar to Karpukhin et al. (2020). Finally, Mi et al. (2021) and Kim et al. (2021b) propose different data augmentation methods to augment the training data by unseen knowledge documents.

In general, pretrained language models have found great success to perform the aforementioned tasks. Notably,

Table 8: Comparison of responses generated by the noisy channel and direct model.

Last turn u_T	selected document	Model: Generated response u_{T+1}
o great and do you know if the taj happens to uh allow smoking	Is there a smoking area at Taj Campton Place?	<i>Baseline</i> : yeah let me check that for you all right so it does say that smoking is allowed here
	Yes, the Taj Campton Place has a designated smoking area.	<i>Noisy Channel</i> : yeah let me check that for you all right so they do have a designated smoking area at this location
two stars ok do you know if they have a reservation or parking	Do I need to make a reservation for parking in advance?	<i>Baseline</i> : yeah let me go ahead and see here all right so it's saying that they do not have a reservation or parking at this location
	Parking spots are available on a first come, first served basis. No reservation is available now.	<i>Noisy Channel</i> : yeah let me go ahead and see here all right so it's saying that they do not have a reservation but parking spots are available on a first come, first served basis
can i bring a dog there	Are dogs allowed?	<i>Baseline</i> : they do not
	Aquarium Of The Bay does not allow dogs.	<i>Noisy Channel</i> : i'm sorry they do not allow dogs at this location
oh great and what's this place lt ike if p m b m	What is the atmosphere like?	<i>Baseline</i> : yeah let me go ahead and search that all right so it's not saying that you can't bring a dog there but it does say that it's a classy and casual place
	It has a classy and casual ambience with a great outdoor view of the water.	<i>Noisy Channel</i> : let me go ahead and search that all right so it's listed here as classy and casual with a great outdoor view of the water
l that sound great ill just wondering do they accept accept the scoffer card	What type of payment method do you accept?	<i>Baseline</i> : yeah let me go ahead and check if that's listed anywhere ok so they do accept all major credit cards including scoffer card
	The Nob Hill Motor Inn accepts, Visa, Mastercard, American Express and Discover.	<i>Noisy Channel</i> : yeah let me go ahead and check that for you. ok so they do accept all major credit cards including the scoffer card

RoBERTa (Liu et al. 2019) has often been used for the sequence classification tasks, while BART (Lewis et al. 2020a) has often been used for response generation, as is suggested by their corresponding pretraining objectives.

Parallel to our work on the Density Ratio model, Liu et al. (2021a) applied a similar method to controlled generation.

The Noisy Channel decomposition (Shannon 1948) has a long history in different language technology tasks, such as machine translation (Brown et al. 1993) or automatic speech recognition (Bahl, Jelinek, and Mercer 1983). With the advent of deep learning, modeling these tasks discriminatively has often been the preferred choice. Nevertheless, recently neural noisy channel modeling has been explored for different tasks, such as machine translation (Yu et al. 2017; Yee, Dauphin, and Auli 2019; Jean and Cho 2020; Subramanian et al. 2021), few-shot text classification (Min et al. 2021), and dialog (Liu et al. 2021b).

6 Conclusion

We have proposed different methods of adapting document-grounded dialog models to noisy transcriptions of spoken conversations. Notably, we achieve significant improvements in the tasks of turn detection and knowledge selection with rather simple text preprocessing steps. For both tasks,

we also propose to use multiple ASR hypotheses which can prevent some errors, for example when an important entity is present only in a few hypotheses. Furthermore, we highlight a data augmentation method to create samples for new knowledge documents by adapting dialogs from MultiWOZ. For the knowledge selection task, we propose a beam search for hierarchical selection models and pretrain the model on Taskmaster. We also explore different methods to improve the domain adaption of response generation models. In particular, we show the benefits of using a noisy channel factorization to leverage ungrounded dialog data and enforce the consistency of responses regarding their document grounding.

In the future, it would be interesting to explore a tighter integration with speech recognition models. This would allow for better error propagation and to make use of paralinguistic features like emphasis or emotion. Furthermore, we plan to study alternative online decoding algorithms for the noisy channel model to alleviate the strong dependency on the proposal model and especially create more diverse proposals. We also want to gain a better understanding of the model and its differences to the direct modeling approach. With this in mind, evaluation strategies that capture aspects such as diversity and especially faithfulness are of interest.

Acknowledgments

This work has received funding from the European Research Council (ERC) under the European Union’s Horizon 2020 research and innovation programme (grant agreement No 694537, project “SEQCLAS”). The work reflects only the authors’ views and the European Research Council Executive Agency (ERCEA) is not responsible for any use that may be made of the information it contains.

References

- Bahl, L. R.; Jelinek, F.; and Mercer, R. L. 1983. A Maximum Likelihood Approach to Continuous Speech Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, PAMI-5(2): 179–190.
- Brown, P. F.; Della Pietra, S. A.; Della Pietra, V. J.; and Mercer, R. L. 1993. The Mathematics of Statistical Machine Translation: Parameter Estimation. *Computational Linguistics*, 19(2): 263–311.
- Byrne, B.; Krishnamoorthi, K.; Sankar, C.; Neelakantan, A.; Goodrich, B.; Duckworth, D.; Yavuz, S.; Dubey, A.; Kim, K.-Y.; and Cedilnik, A. 2019. Taskmaster-1: Toward a Realistic and Diverse Dialog Dataset. 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)*, 4515–4524. Hong Kong, China: Association for Computational Linguistics.
- Dinan, E.; Roller, S.; Shuster, K.; Fan, A.; Auli, M.; and Weston, J. 2018. Wizard of Wikipedia: Knowledge-Powered Conversational Agents. In *International Conference on Learning Representations*.
- Eric, M.; Goel, R.; Paul, S.; Kumar, A.; Sethi, A.; Ku, P.; Goyal, A. K.; Agarwal, S.; Gao, S.; and Hakkani-Tur, D. 2020. MultiWOZ 2.1: A Consolidated Multi-Domain Dialogue Dataset with State Corrections and State Tracking Baselines. In *Proceedings of The 12th Language Resources and Evaluation Conference*, 422–428.
- Fazel-Zarandi, M.; Wang, L.; Tiwari, A.; and Matsoukas, S. 2019. Investigation of Error Simulation Techniques for Learning Dialog Policies for Conversational Error Recovery. *ArXiv*, abs/1911.03378.
- Feng, S.; Wan, H.; Gunasekara, C.; Patel, S.; Joshi, S.; and Lastras, L. 2020. Doc2dial: A Goal-Oriented Document-Grounded Dialogue Dataset. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 8118–8128. Online: Association for Computational Linguistics.
- Gopalakrishnan, K.; Hedayatnia, B.; Wang, L.; Liu, Y.; and Hakkani-Tür, D. Z. 2020. Are Neural Open-Domain Dialog Systems Robust to Speech Recognition Errors in the Dialog History? An Empirical Study. In *INTERSPEECH*.
- Gulcehre, C.; Firat, O.; Xu, K.; Cho, K.; Barrault, L.; Lin, H.-C.; Bougares, F.; Schwenk, H.; and Bengio, Y. 2015. On Using Monolingual Corpora in Neural Machine Translation. *arXiv:1503.03535*.
- Guu, K.; Lee, K.; Tung, Z.; Pasupat, P.; and Chang, M.-W. 2020. REALM: Retrieval-Augmented Language Model Pre-Training. In *Proceedings of the 37th International Conference on Machine Learning*.
- He, H.; Lu, H.; Bao, S.; Wang, F.; Wu, H.; Niu, Z.; and Wang, H. 2021. Learning to Select External Knowledge with Multi-Scale Negative Sampling. *arXiv:2102.02096 [cs]*.
- Henderson, M.; Gašić, M.; Thomson, B.; Tsiakoulis, P.; Yu, K.; and Young, S. 2012. Discriminative spoken language understanding using word confusion networks. In *2012 IEEE Spoken Language Technology Workshop (SLT)*, 176–181.
- Jean, S.; and Cho, K. 2020. Log-Linear Reformulation of the Noisy Channel Model for Document-Level Neural Machine Translation. In *Proceedings of the Fourth Workshop on Structured Prediction for NLP*, 95–101. Online: Association for Computational Linguistics.
- Jin, D.; Kim, S.; and Hakkani-Tur, D. 2021. Can I Be of Further Assistance? Using Unstructured Knowledge Access to Improve Task-Oriented Conversational Modeling. In *Proceedings of the 1st Workshop on Document-Grounded Dialogue and Conversational Question Answering (Dial-Doc 2021)*, 119–127. Online: Association for Computational Linguistics.
- Karpukhin, V.; Oğuz, B.; Min, S.; Lewis, P.; Wu, L.; Edunov, S.; Chen, D.; and Yih, W.-t. 2020. Dense Passage Retrieval for Open-Domain Question Answering. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing (EMNLP)*, 6769–6781. Association for Computational Linguistics.
- Kim, S.; Eric, M.; Gopalakrishnan, K.; Hedayatnia, B.; Liu, Y.; and Hakkani-Tur, D. 2020. Beyond Domain APIs: Task-Oriented Conversational Modeling with Unstructured Knowledge Access. In *Proceedings of the 21th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, 278–289. 1st virtual meeting: Association for Computational Linguistics.
- Kim, S.; Eric, M.; Hedayatnia, B.; Gopalakrishnan, K.; Liu, Y.; Huang, C.-W.; and Hakkani-Tur, D. 2021a. Beyond Domain APIs: Task-Oriented Conversational Modeling with Unstructured Knowledge Access Track in DSTC9. In *AAAI 2021, Workshop on DSTC9*.
- Kim, S.; Liu, Y.; Jin, D.; Papangelis, A.; Gopalakrishnan, K.; Hedayatnia, B.; and Hakkani-Tur, D. 2021b. “How Robust r u?”: Evaluating Task-Oriented Dialogue Systems on Spoken Conversations. *arXiv:2109.13489 [cs]*.
- Lewis, M.; Liu, Y.; Goyal, N.; Ghazvininejad, M.; Mohamed, A.; Levy, O.; Stoyanov, V.; and Zettlemoyer, L. 2020a. BART: Denoising Sequence-to-Sequence Pre-Training for Natural Language Generation, Translation, and Comprehension. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics*, 7871–7880. Association for Computational Linguistics.
- Lewis, P.; Perez, E.; Piktus, A.; Petroni, F.; Karpukhin, V.; Goyal, N.; Küttler, H.; Lewis, M.; Yih, W.-t.; Rocktäschel, T.; Riedel, S.; and Kiela, D. 2020b. Retrieval-Augmented Generation for Knowledge-Intensive NLP Tasks. In *NeurIPS*.

- Lhoest, Q.; Villanova del Moral, A.; Jernite, Y.; Thakur, A.; von Platen, P.; Patil, S.; Chaumond, J.; Drame, M.; Plu, J.; Tunstall, L.; Davison, J.; Šaško, M.; Chhablani, G.; Malik, B.; Brandeis, S.; Le Scao, T.; Sanh, V.; Xu, C.; Patry, N.; McMillan-Major, A.; Schmid, P.; Gugger, S.; Delangue, C.; Matussière, T.; Debut, L.; Bekman, S.; Cistac, P.; Goehringer, T.; Mustar, V.; Lagunas, F.; Rush, A.; and Wolf, T. 2021. Datasets: A Community Library for Natural Language Processing. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 175–184. Online and Punta Cana, Dominican Republic: Association for Computational Linguistics.
- Liu, A.; Sap, M.; Lu, X.; Swayamdipta, S.; Bhagavatula, C.; Smith, N. A.; and Choi, Y. 2021a. DExperts: Decoding-Time Controlled Text Generation with Experts and Anti-Experts. In *Proceedings of the 59th Annual Meeting of the Association for Computational Linguistics and the 11th International Joint Conference on Natural Language Processing (Volume 1: Long Papers)*, 6691–6706. Online: Association for Computational Linguistics.
- Liu, Q.; Yu, L.; Rimell, L.; and Blunsom, P. 2021b. Pretraining the Noisy Channel Model for Task-Oriented Dialogue. *Transactions of the Association for Computational Linguistics*, 9: 657–674.
- Liu, Y.; Ott, M.; Goyal, N.; Du, J.; Joshi, M.; Chen, D.; Levy, O.; Lewis, M.; Zettlemoyer, L.; and Stoyanov, V. 2019. RoBERTa: A Robustly Optimized BERT Pretraining Approach. *arXiv:1907.11692 [cs]*.
- McDermott, E.; Sak, H.; and Variani, E. 2019. A Density Ratio Approach to Language Model Fusion in End-to-End Automatic Speech Recognition. In *2019 IEEE Automatic Speech Recognition and Understanding Workshop (ASRU)*, 434–441.
- Mi, H.; Ren, Q.; Dai, Y.; He, Y.; Li, Y.; Sun, J.; Zheng, J.; and Xu, P. 2021. Towards Generalized Models for Beyond Domain API Task-Oriented Dialogue. In *AAAI 2021, Workshop on DSTC9*, 8.
- Min, S.; Lewis, M.; Hajishirzi, H.; and Zettlemoyer, L. 2021. Noisy Channel Language Model Prompting for Few-Shot Text Classification. *arXiv:2108.04106 [cs]*.
- Peter, J.-T.; Beck, E.; and Ney, H. 2018. Sisyphus, a Workflow Manager Designed for Machine Translation and Automatic Speech Recognition. In *Proceedings of the 2018 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 84–89. Brussels, Belgium: Association for Computational Linguistics.
- Radlinski, F.; Balog, K.; Byrne, B.; and Krishnamoorthi, K. 2019. Coached Conversational Preference Elicitation: A Case Study in Understanding Movie Preferences. In *Proceedings of the Annual Meeting of the Special Interest Group on Discourse and Dialogue (SIGDIAL)*.
- Schatzmann, J.; Thomson, B.; and Young, S. 2007. Error simulation for training statistical dialogue systems. In *2007 IEEE Workshop on Automatic Speech Recognition Understanding (ASRU)*, 526–531.
- Shannon, C. E. 1948. A mathematical theory of communication. *The Bell System Technical Journal*, 27(3): 379–423.
- Subramanian, S.; Hrinchuk, O.; Adams, V.; and Kuchaiev, O. 2021. NVIDIA NeMo Neural Machine Translation Systems for English-German and English-Russian News and Biomedical Tasks at WMT21. *arXiv:2111.08634*.
- Tang, L.; Shang, Q.; Lv, K.; Fu, Z.; Zhang, S.; Huang, C.; and Zhang, Z. 2021. RADGE: Relevance Learning and Generation Evaluating Method for Task-Oriented Conversational Systems. In *AAAI 2021, Workshop on DSTC9*, 7.
- Thulke, D.; Daheim, N.; Dugast, C.; and Ney, H. 2021. Efficient Retrieval Augmented Generation from Unstructured Knowledge for Task-Oriented Dialog. In *AAAI 2021, Workshop on DSTC9*.
- Tur, G.; Wright, J.; Gorin, A.; Riccardi, G.; and Hakkani-Tür, D. 2002. Improving spoken language understanding using word confusion networks. In *Proc. 7th International Conference on Spoken Language Processing (ICSLP 2002)*, 1137–1140.
- Weng, Y.; Miryala, S. S.; Khatri, C.; Wang, R.; Zheng, H.; Molino, P.; Namazifar, M.; Papangelis, A.; Williams, H.; Bell, F.; and Tur, G. 2020. Joint Contextual Modeling for ASR Correction and Language Understanding. *ICASSP 2020 - 2020 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, 6349–6353.
- Wolf, T.; Debut, L.; Sanh, V.; Chaumond, J.; Delangue, C.; Moi, A.; Cistac, P.; Rault, T.; Louf, R.; Funtowicz, M.; Davison, J.; Shleifer, S.; von Platen, P.; Ma, C.; Jernite, Y.; Plu, J.; Xu, C.; Le Scao, T.; Gugger, S.; Drame, M.; Lhoest, Q.; and Rush, A. 2020. Transformers: State-of-the-Art Natural Language Processing. In *Proceedings of the 2020 Conference on Empirical Methods in Natural Language Processing: System Demonstrations*, 38–45. Online: Association for Computational Linguistics.
- Yee, K.; Dauphin, Y.; and Auli, M. 2019. Simple and Effective Noisy Channel Modeling for Neural Machine Translation. 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)*, 5695–5700. Hong Kong, China: Association for Computational Linguistics.
- Yu, L.; Blunsom, P.; Dyer, C.; Grefenstette, E.; and Kocisky, T. 2017. The Neural Noisy Channel. In *ICLR*.