Natural Language Generation at Scale: A Case Study for Open Domain Question Answering

Alessandra Cervone[1]*, **Chandra Khatri**[2]†, **Rahul Goel**[3]‡, **Behnam Hedayatnia**[4], **Anu Venkatesh**[4], **Dilek Hakkani-Tür**[4], **Raefer Gabriel**[4]
[1]Signals and Interactive Systems Lab, University of Trento, Italy
[2]Uber AI, [3]Google, [4]Amazon Alexa AI
alessandra.cervone@unitn.it, chandrak@uber.com,
{behnam, anuvenk, hakkanit, raeferg}@amazon.com,
goelrahul@google.com

Abstract

Current approaches to Natural Language Generation (NLG) for dialog mainly focus on domain-specific, task-oriented applications (e.g. restaurant booking) using limited ontologies (up to 20 slot types), usually without considering the previous conversation context. Furthermore, these approaches require large amounts of data for each domain, and do not benefit from examples that may be available for other domains. This work explores the feasibility of applying statistical NLG to scenarios requiring larger ontologies, such as multi-domain dialog applications or open-domain question answering (QA) based on knowledge graphs. We model NLG through an Encoder-Decoder framework using a large dataset of interactions between real-world users and a conversational agent for open-domain QA. First, we investigate the impact of increasing the number of slot types on the generation quality and experiment with different partitions of the QA data with progressively larger ontologies (up to 369 slot types). Second, we perform multi-task learning experiments between open-domain QA and task-oriented dialog, and benchmark our model on a popular NLG dataset. Moreover, we experiment with using the conversational context as an additional input to improve response generation quality. Our experiments show the feasibility of learning statistical NLG models for open-domain QA with larger ontologies.

1 Introduction

In dialog literature Natural Language Generation (NLG) is framed as the task of generating natural language responses that faithfully convey the semantic information given by a Meaning Representation (MR). A MR is typically a structure consisting of a Dialog Act (DA) and a list of associated slots. While the DA (Stolcke et al., 2000; Mezza et al., 2018) expresses the intent of the utterance to be generated (e.g. "inform" in Table 1), the slots, organized as slot type-slot value pairs (e.g. *food*:'french' in Table 1), represent the information which has to be conveyed in the generated text.

	Input		Output
	context	MR	Text
Task Oriented	-	**inform** *name*: 'fringale' *food*: 'french'	'fringale is a french restaurant' 'fringale serves french food'
QA	'when was kentucky founded'	**inform** *timepoint*: '1792' *objStr*: 'kentucky' *claStr*: 'state' *relStr*: 'founded'	'1792' 'kentucky formed in 1792' 'kentucky founded in 1792'

Table 1: Examples of input-output pairs from a task-oriented (Task) NLG (SFX (Wen et al., 2015)) and a Question-Answering (QA) dataset. In NLG the input is typically a Meaning Representation (MR) and the output is its textual realization (Text). Each MR is composed of a Dialog Act (bold) and a list of slot type (italic)-value pairs. Compared to most NLG datasets, our QA corpus also has the previous question (context) as input. While in the task-oriented setting we observe a one-to-one relation between slots in the input and the ones realized in the text, the same is not true for QA.

So far statistical NLG for dialog has mainly been investigated in research for task-oriented applications (e.g. restaurant reservation, bus information) in narrow, controlled environments with limited ontologies, i.e. considering a small set of DAs and slot types (respectively 12 and 8 in the popular San Francisco restaurant dataset (SFX) (Wen et al., 2015), 8 and 1 in the recent E2E NLG challenge (Novikova et al., 2017)). Furthermore, most datasets consider MRs in isolation (Novikova et al., 2017) i.e., they lack con-

* Work done during internship at Amazon Alexa AI.
† Work done while working at Amazon Alexa AI.
‡ Work done while working at Amazon Alexa AI.

Proceedings of The 12th International Conference on Natural Language Generation, pages 453–462,
Tokyo, Japan, 28 Oct - 1 Nov, 2019.

versational context, even though the previous utterances in the dialog have been shown to improve the performance of task-oriented NLG (Dušek and Jurcicek, 2016). These characteristics of current approaches to NLG can be linked to the fact that a vast majority of dialog NLG research is tested on a single domain where the dialog agent performs simple tasks such as giving information about a restaurant, with few exceptions (Wen et al., 2016).

However, with the rise of conversational agents such as Amazon Alexa and Google Assistant, there is an increasing interest in complex multi-domain tasks. These systems typically rely on hand-crafted NLG, but this approach cannot scale to the complex ontologies which may be required in real-world applications (e.g. booking a trip).

In this work we explore the applicability of current NLG models for task-oriented dialog, based on a MR-to-text framework using Encoder-Decoder architectures, to open-domain QA. This allows us to investigate the performance of current NLG research in an environment with (1) much larger numbers of slot types, and (2) a different application compared to task-oriented dialog. We generate the QA datasets for our experiments using as source a large corpus of open-domain QA pairs from interactions between real-world users and a conversational agent. For evaluation, we utilize both objective metrics and human judgment. We observe that NLG for open-domain QA poses its own challenges compared to task-oriented dialog, since correct answers to the same question do not necessarily convey all slot types in the MR (see Table 1).

In particular, in our first set of experiments, we investigate the effect of using increasingly larger ontologies with regards to slot types on the performance of our NLG models for QA. We find that, notwithstanding the larger ontologies and the noisiness of our dataset, models' performance does not degrade significantly in terms of naturalness of generated text and efficiency in encoding the MR information (i.e. Slot Error Rate). Interestingly, we find it improves for some of the human evaluation metrics. We also observe that using conversational context improves the quality of generated responses. In our second set of experiments, we investigate whether jointly training NLG models for task-oriented dialog and QA improves performances. To this end, we experiment with learning NLG models in a multi-task setting between our QA data and SFX. Our experiments show that learning models in a multi-task setting lead to better performances in terms of naturalness of the generated output for both tasks.

This work has several contributions:

1. We apply the MR-to-text framework (typical of NLG for task-oriented dialog) to a open-domain QA application.
2. We explore the importance of adding the previous conversational context to improve the quality of the generated output.
3. We investigate the possibility of learning NLG models using a MR-to-text approach with increasingly larger ontologies in terms of slot types.
4. We experiment with multi-task learning for NLG between open-domain QA and task-oriented dialog.
5. Finally we also propose new evaluation metrics (see Section 5) to capture the variability of output in open-domain QA compared to NLG for task-oriented dialog.

2 Related work

While classical approaches to NLG involve a pipeline of modules such as content selection, planning, and surface realization (Gatt and Krahmer, 2018), recently a large part of the literature investigated end-to-end neural approaches to NLG. The tasks tackled include dialog, text, and QA. While these tasks share some similarities, each comes with its own set of challenges and requires specific solutions.

NLG for dialog State of the art NLG models for dialog (Dušek and Jurcıcek, 2016; Juraska et al., 2018) mostly use end-to-end neural Encoder-Decoder approaches with attention (Bahdanau et al., 2014) and re-ranking (Dušek et al., 2018). Ensembling is another technique employed to boost model performance (Juraska et al., 2018). Using delexicalization (Henderson et al., 2014), i.e., the process of substituting slot values with slot types in the generated text, has also shown improvements in many settings. However, recent work also depicted the disadvantages of delexicalization (Nayak et al., 2017). In our work, we compare and combine both delixecalized and lexicalized inputs for the NLG system.

NLG for dialog has been mostly tested in controlled environments using task-oriented, single domain datasets with limited ontologies (Wen et al., 2015; Novikova et al., 2017; Balakrishnan et al., 2019). Although Wen et al. (2016) perform multi-domain task-oriented NLG experiments, the ontologies used are still limited for such settings. Finally, while research has shown how encoding the previous utterance leads to better performances (Dušek and Jurcicek, 2016), most settings consider the turns in isolation (Wen et al., 2015; Novikova et al., 2017).

In our work, we perform open-domain NLG with significantly larger ontologies and also evaluate the impact of adding the context to the input.

NLG for text and QA Recent work around NLG for text involves generating text using structured data using the encoder-decoder networks (Mei et al., 2016). Similarly to dialog, NLG for text has also been addressed in controlled environments such as weather forecast (Liang et al., 2009) with few exceptions (Lebret et al., 2016).

In the literature for QA, most approaches retrieve answers directly or generate answers jointly with the retrieval, and answers are usually entities or lists of entities (Dodge et al., 2015). On the contrary, in NLG we assume the answer has already been retrieved, and the goal is to generate text matching it. The field of QA which most strictly relates to our work is answer generation, where current approaches are also based on encoder-decoder networks encoding information directly from a knowledge base (Yin et al., 2016; He et al., 2017; Wei and Zhang, 2019). An additional challenge to answer generation is that there are no publicly available datasets for this task (Fu and Feng, 2018).

Our approach differs from answer generation in that we structure the task as in NLG dialog literature with a MR-to-text approach.

3 Datasets

3.1 Question Answering

Source data Our source for generating the MR-text pairs are thousands of open-domain factual question-answer pairs from commercial data. The domains covered in this data are manifold, including geography (e.g. 'is canada bigger than united states' in Table 7), history (e.g. 'when was kentucky founded' in Table 1), present-day knowledge (e.g. 'will ferrell's wife' in Table 7), grammar ('is there a plural form of pegasus') and even mathematics ('what is one modulo seven'). Pairs are grouped according to the type of question asked. Each group consists of a list of specific questions (e.g. "who is the wife of barack obama", "tell me the wives of henry the viii") of the same type (e.g. "who is the wife of") asked by real users to a conversational agent. Each specific question additionally has: (1) the answer to the question (e.g. "michelle obama is obama's wife") generated by the NLG of the conversational system, either using information retrieval or a knowledge base search coupled with templates; (2) relevant noun and verb phrases (e.g. "michelle obama", "barack obama", "wife") used by the system to generate the answer, including the ones from the question. Noun phrases are tagged according to their semantic type (examples of semantic types are *timepoint* and *human being*), while verb phrases are tagged as "relation" types (see "founded" tagged as *relStr* in Table 1).

	Size	Slots	DAs	Words	Domain	Context
E2E	51k	8	1	2453	restaurant	no
SFX	5k	12	8	438	restaurant	no
QA.1	6k	147	1	702	open	yes
QA.2	16k	210	1	1528	open	yes
QA.3	67k	369	1	2963	open	yes

Table 2: Our QA NLG datasets compared to popular (task-oriented) NLG datasets: San Francisco restaurant (SFX) and the NLG E2E challenge (E2E). We report the full size of datasets in terms of MR-text pairs, the number of slot types, DAs, words (computed after delexicalization), domain and whether the dataset comprises the previous utterance or not.

The answers in the source data are varied, and range from a simple entity to a fully formed answer, as in Table 1 example where valid answers to the question "when was kentucky founded" can be "1792" or "kentucky formed in 1792". This shows an interesting difference between our QA data and task-oriented NLG datasets. While for task-oriented NLG all valid responses for a single MR have the same slot types (i.e., the ones in the input MR), in our dataset this is not always true.

QA NLG datasets We generate the NLG input-output pairs for QA from our source data. In order to perform cross-application experiments, we maintain the same MR-text format as task-oriented dialog NLG. The target output is the text

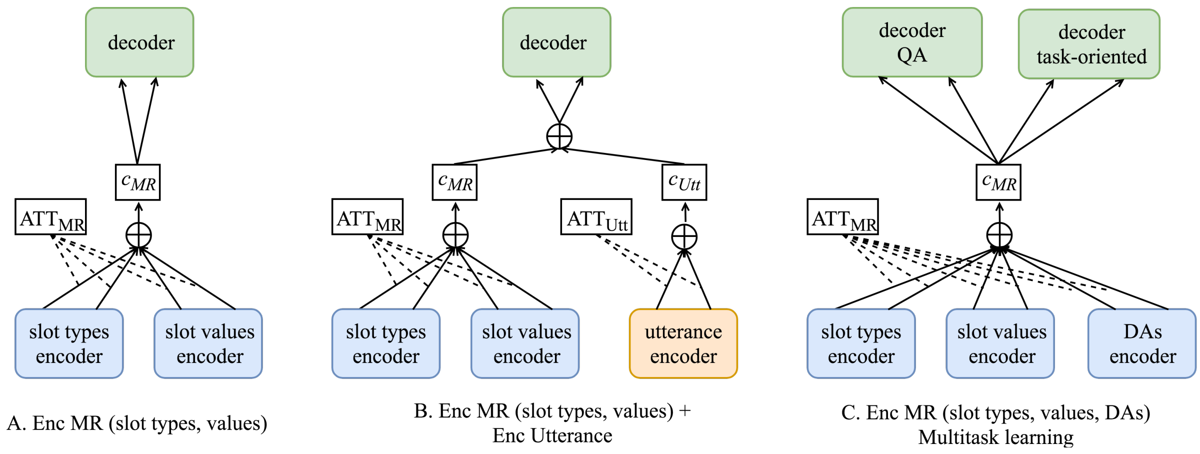

Figure 1: Our baseline model (A) and the models with the previous utterance (B) and for multi-task learning (C). While our baseline model Enc MR (slot types, values) is composed by two encoders for the MR, one for slot types and one for slot values; our model in subfigure B extends this baseline by adding an encoder for the previous utterance. In the multitask learning setting, on the hand, where we do not have the previous context but might have different Dialog Acts (DAs), we add a corresponding encoder (see subfigure C).

of the answers in the source data. To generate the input MRs we assumed only one DA across all answers, i.e. "inform"; for the slots, we used the semantic types and relations for noun phrases and verb phrases in the source data as slot types, while the actual entity or verb was used as the corresponding slot value. [1] On top of the generated MR we use, as additional input, the previously asked question as context.

Answers are delexicalized (Henderson et al., 2014) to improve generalization. Since we do not have alignment between entities in the input and the generated text, we use a heuristic-based aligner which we also use to filter out data that could not be appropriately aligned. All noun phrases are delexicalized while verb phrases are not. Furthermore, similar to (Juraska et al., 2018), we use delexicalization for data augmentation. We generate additional references for each MR, besides the original one, by considering all delexicalized answers in the question group as candidate template answers for each specific question in the group and then substituting (where possible) slots values which are already available in the input. The text of the previous question is also delexicalized.

Finally, to investigate performances across different ontology sizes, we generate 3 different partitions of the data (QA.1, 2 and 3 in Table 2) with a progressively larger number of slot types. Each QA partition was split in train, test and development set (using a 80-10-10 split) according to the type of question asked. We ensured there was no overlap between the different sets to test if models generalize to previously unseen questions.

3.2 Task-oriented Dialog

As a task-oriented NLG corpus for our multi-task learning experiments we use the popular San Francisco restaurants (SFX; Wen et al. (2015)) dataset. Statistics about the dataset is shown in Table 2. Although SFX is not large (6k examples), compared to the E2E NLG corpus it presents more variation for DA (although less in style). For all our datasets, we use the TGEN library [2] (Dušek and Jurcıcek, 2016) to delexicalize all slot types except binary values.

4 Model and Architectures

In this section we present the variety of different architectures used in our experiments. Although all our models are based on the Encoder-Decoder framework, we investigate architectures with different number of Encoders (up to 3). Given this variety, for clarity, we follow a template Enc <Encoder type> for naming our different models. The type of Encoder, in particular, can be of Meaning Representation (MR) type, when we encode parts of the MR, such as slot types, values or Dialog Act; or it can be of Utterance type, when we encode the previous utterance context.

[1] Although we use the original tags of the source data, a similar representation could be produced by tagging noun phrases with their Named Entity type and verb phrases with a "relation" slot type.

[2] https://github.com/UFAL-DSG/tgen

Encoder-Decoder with Attention Following recent state-of-the-art approaches to NLG for dialog (Juraska et al., 2018; Balakrishnan et al., 2019), our models are based on the Encoder-Decoder with Attention framework. In particular, we use bidirectional Gated Recurrent Units (GRU) and Luong general attention (Luong et al., 2015) as our baseline. While we also experimented with other types of architectures, such as using Long-Short-Term Memory Units (Hochreiter and Schmidhuber, 1997) instead of GRUs and different types of attention (including Bahdanau attention (Bahdanau et al., 2014) and Luong dot attention (Luong et al., 2015), this combination gave us the best results for our setting. Depending on the encoder used, either slot type or slot value, we refer to this model as Enc MR (slot types) or Enc MR (slot values).

Multi-Encoder, Single Decoder: We expand the baseline (Enc MR) models using multiple inputs from the MR (slot types, values, DAs), each encoded by a different encoder. The attention is performed on their concatenated output to produce the MR context vector c_{MR}. Figure 1 A shows an example of such an architecture using two encoders, one for slot types and one for slot values. Furthermore, we experimented with adding the previous utterance as input with an additional encoder (Enc Utterance). In this case, the context vector for the previous utterance c_{Utt} is produced by an independent attention mechanism and the outputs of both attentions (c_{MR} and c_{Utt}) are concatenated (see Figure 1 B).

Multi-Encoder, Multi-Decoder: We also performed multi-task learning, jointly training the models for both QA and task-oriented NLG. As shown in Figure 1 C, we shared the encoders and corresponding input layers across multiple tasks while we maintained multiple decoders for individual tasks. We alternated between mini-batches from various data sources to perform multitasking.

5 Evaluation

As word overlap metrics may not have a good correlation with human judgment for NLG output evaluation (Stent et al., 2005), we use both objective metrics and human evaluation.

Objective metrics Besides the standard BLEU score (obtained using the official E2E NLG challenge evaluation script [3]), we report different types of Slot Error Rate (SER). In dialog NLG approaches SER shows the number of correct slots in the output compared to the input MR. We refer to this metric as SER_{mr} to differentiate it from its modified versions we introduce next. The formula (Wen et al., 2015) is:

$$SER_{mr} = \frac{p_{mr} + q_{mr}}{N_{mr}} \tag{1}$$

where N_{mr} is the total number of slots in the input MR and p_{mr} , q_{mr} are respectively the number of missing and redundant slots in the output. This formula works well for task-oriented NLG approaches, but it assumes a one-to-one relationship between the slots in the input MR and the output text. We found this assumption might not hold for our QA datasets where not all slots in the input MR need to be realized for the output to be correct. An example of this is shown in Table 1, where the first QA reference text ('1792') would be penalized with 3 missing slots, while still being correct.

In order to capture this different behaviour we designed additional NLG metrics tailored for QA. Slot Error Rate Target (SER_{trg}) is a modification of SER_{mr} where we simply substitute the MR with the main reference text:

$$SER_{trg} = \frac{p_{trg} + q_{trg}}{N_{trg}} \tag{2}$$

SER_{trg} is designed to penalize both missing and redundant slots compared to the target sentence. Hence, using SER_{trg} the first QA reference text in Table 1 would not be penalized.

Slot Error Rate MultiTarget (SER_{mtrg}), on the other hand, penalizes redundant slots that did not appear in any of the references:

$$SER_{mtrg} = \frac{p_{mtrg}}{N_{mtrg}} \tag{3}$$

where N_{mtrg} are all slots appearing in any reference and p_{mtrg} are the slots in the output that did not appear in any reference sentence. To compute SER_{mtrg} for the model output "kentucky formed in 1792" given the QA MR in Table 1 we assume to have two references "1792" and "kentucky formed in 1792". In this case, SER_{mtrg} would consider the output correct as all of its slots appear in at least one of the references.

[3] We do not report other word overlap metrics (e.g., METEOR) computed by the E2E evaluation scripts due to space limitations and correlation with the BLEU score.

	QA.1				QA.2				QA.3			
	BLEU	**SER_{mr}**	**SER_{trg}**	**SER_{mtrg}**	**BLEU**	**SER_{mr}**	**SER_{trg}**	**SER_{mtrg}**	**BLEU**	**SER_{mr}**	**SER_{trg}**	**SER_{mtrg}**
Enc MR (slot types, values)	0.85	**0.42**	0.21	0.023	0.77	0.44	0.3	0.057	0.66	**0.43**	0.37	**0.05**
+ Enc Utterance delex	0.89	0.44	0.19	0.014	0.83	**0.43**	0.23	**0.025**	0.7	0.45	0.32	0.054
+ Enc Utterance lex	**0.95**	0.46	**0.15**	**0.012**	**0.89**	0.47	**0.19**	0.03	**0.72**	0.46	**0.28**	0.067

Table 3: Objective metrics results on three QA NLG datasets with increasingly larger ontologies. The models under comparison are a baseline with two encoders, for slot types and slot values, and its extensions with a delexicalised or lexicalised previous utterance. While for BLEU score the higher the better, for all types of Slot Error Rate (SER) the lower the better.

	QA.1				QA.2				QA.3			
	Nat.	***Inf.***	***Conv.***	***Ans.***	***Nat.***	***Inf.***	***Conv.***	***Ans.***	***Nat.***	***Inf.***	***Conv.***	***Ans.***
Enc MR (slot types, values)	3.73	3.8	3.96	0.36	3.88	3.78	4.57	0.38	4.29	4.37	4.40	0.73
+ Enc Utterance delex	4.61	4.63	**4.59**	0.67	4.64	4.32	**4.88**	0.5	4.52	4.47	**4.51**	0.79
+ Enc Utterance lex	**4.7**	**4.69**	4.48	**0.78**	**5.15**	**4.88**	4.85	**0.67**	**4.57**	**4.57**	4.45	**0.80**

Table 4: Human evaluation results on three QA NLG datasets with increasingly larger ontologies. The models reported are a baseline with two encoders, for slot types and slot values, and its extensions with a delexicalised or lexicalised previous utterance. We report averages of Naturalness (*Nat.*), Informativeness (*Inf.*), and how conversational the response was judged (*Conv.*) on a scale of 1 to 6. Additionally, we report the average of whether responses could be considered an answer to the given question (*Ans.*), given to annotators as a binary choice.

	Dataset	**BLEU**	**SER_{mr}**	**SER_{trg}**	**SER_{mtrg}**
baseline	SFX	0.727	**0.40**	-	-
+ QA.3		**0.74**	0.413	-	-
baseline	QA.3	0.659	**0.429**	0.37	**0.05**
+ SFX		**0.673**	0.44	**0.368**	0.07

Table 5: Objective metrics of multitask learning experiments combining QA (QA.3) and task-oriented dialog (SFX) NLG. For all Slot Error Rate (SER) metrics the lower the better.

	Dataset	***Nat.***	***Inf.***	***Conv.***	***Ans.***
baseline	SFX	4.69	**5.50**	-	-
+ QA.3		**5.11**	5.40	-	-
baseline	QA.3	4.29	4.33	4.40	**0.73**
+SFX		**4.43**	**4.38**	**4.5**	0.72

Table 6: Results of multitask learning experiments combining NLG for QA (on QA.3) and task-oriented dialog (on SFX) according to human evaluation.

Human evaluation In all experiments, for each dataset, we selected a sample of 100 MR-text pairs from the test set. Pairs were randomly selected among those where all models under comparison in the experiment had generated different output text. Data for all reported experiments were annotated by 2 human annotators, and final ratings were averaged between the two. In all experiments annotators, presented with MR and all outputs of the systems under comparison, were asked to rate the *naturalness* and *informativeness* of the generated output using a 1-6 Likert score, as in previous NLG dialog evaluations (Gatt and Krahmer, 2018). Additionally, for the QA datasets annotators had also the previous question as context. Moreover, for the QA datasets annotators were asked to rate how *conversational* the output was, on the same Likert scale, and whether or not the output could ultimately be considered an answer to the question (*answer*), as a binary choice.

6 Experimental setup

The hyperparameters chosen for our models were empirically determined through various experiments. Both encoder and decoder in all our models had only one layer, as we noticed additional layers did not give improvements. All embeddings were trained from scratch with a fixed dimension of 50. Models were trained using a cross-entropy loss function and the Adam (Kingma and Ba, 2014) optimizer with a learning rate of 0.001, for 1000 epochs, with early stopping on the validation set. We used mini-batches of size 32.

For the NLG models for QA, experiments on QA.1 (not reported due to space limitations) with different encoders combinations showed that the best performances were achieved using all input types (slot type, value, and previous context) with lexicalized (+ Enc Utterance lex) or delexicalized (+ Enc Utterance delex) previous context in terms of all metrics, except SER_{trg}. On this metric, the architecture with slot types and values, but without the previous context (Enc MR (slot types, values)) achieved the best performance (cf. Table 3). For this reason, we chose to report the performances of these architectures in our QA experiments.

7 Results

Open domain QA In our first batch of experiments we test various Encoder-Decoder architectures on our 3 different partitions of QA NLG data.

As we can see from Table 3, in general, the best performances across all QA datasets for both BLEU and $\mathrm{SER}_{\mathrm{trg}}$ are achieved by the model using as additional input the lexicalized previous question, followed by the model with the delexicalized one. However, $\mathrm{SER}_{\mathrm{mr}}$ results show the opposite picture, where the baseline with only slot types and values performs better (except for QA.2 where the score is close to the model with the delexicalized input) and the model with the lexicalized previous utterance is the worst. $\mathrm{SER}_{\mathrm{mtrg}}$ shows, on the other hand, that the context might slightly degrade performances with bigger ontologies in terms of all text references.

Human evaluation, on the other hand, seems in line with the picture depicted by BLEU and $\mathrm{SER}_{\mathrm{trg}}$. Table 4 shows the model with the lexicalized context is regarded as the best, closely followed by the model with the delexicalized one in every metric except for *conversational*, where delexicalized is better. This confirms our hypothesis that $\mathrm{SER}_{\mathrm{mr}}$ might be a less reliable metric to evaluate NLG QA output. Moreover, although we notice a consistent but not drastic degradation in terms of BLEU and $\mathrm{SER}_{\mathrm{trg}}$ in correlation with bigger ontologies, human evaluation shows an even more gentle degradation between QA.1 and 3 for many metrics. Interestingly, it seems the ability of all models to give a proper answer to the question (*answer*) increases from QA.1 to 3.

Multitask learning In our multitask learning experiments we combine the biggest QA dataset, QA.3, with a task-oriented corpus, SFX. We aim to investigate the possibility of transferring knowledge across different NLG systems, notwithstanding the diversity of the data in terms of domain, ontology size, DAs, application (QA vs. task-oriented). Since context is not available in SFX, the model we use has 3 MR Encoders (slot types, values, DAs) and 2 Decoders (one for each task).

Our experiments show that the NLG QA task improves the fluency on SFX both in terms of objective metrics (in Table 5) and human evaluation (in Table 6). However, training with QA seems to slightly degrade the model efficiency in generating the correct slots. This is to be expected given the difference in the relation between slots in MR and output (one-to-one in SFX, variable in QA.3). As for QA.3 results, it seems the task-oriented NLG task improves QA NLG performances in terms of fluency (BLEU and *Naturalness*) and slot errors ($\mathrm{SER}_{\mathrm{trg}}$ and *Informativeness*). $\mathrm{SER}_{\mathrm{mr}}$ and $\mathrm{SER}_{\mathrm{mtrg}}$, however, show a slight degradation. We observe task-oriented NLG also makes QA NLG more conversational, however slightly reducing its probability of being an answer to the posed question as well.

Finally, comparing all experiments on QA.3, we notice that although multi-task learning helps, the previous context (either lexicalized or delexicalized) plays a critical role in improving the overall performance.

8 Qualitative analysis

In this section we report the qualitative analysis we performed on the human annotated testset. Table 7 reports some output examples from different models given the same input MR. In particular, we are interested on the impact of adding various features and multi-tasking.

QA According to our qualitative analysis on the QA datasets, the baseline model is the one with most grammatical errors (e.g. "will ferrell 's 's wife is viveca paulin", "no , canada is not the bigger than united states ."), while in general adding "delex" and "lex" features generates more grammatical responses. This observation was confirmed from both the objective (in terms of BLEU score) and subjective (*naturalness*) evaluations performed.

We also notice how lexicalizing the previous question helps in producing generally correct (e.g. 'unilever.') however shorter answers, which can be regarded as less conversational. Delexicalizing the input, on the other hand, produces more conversational (e.g. 'popsicle's manufacturer is unilever') but also more factually incorrect answers. These observations seem also in line with the subjective evaluation results, which on average reported the best scores for the model with lexicalized previous context (+ Enc Utterance lex) on whether the generated text could be considered an answer to the given question (*answer*), except for the metric rating how *conversational* the output was, for which the model with delexicalized previous context (+ Enc Utterance delex) was regarded

Dataset	Input		baseline	+delex	+lex	multitask
	context	MR				
QA.1	‘sing your song writer’	*human being*: ’vic ruggiero’, *a*:’sing your song’ *b*:’writer’	‘sing your song ’ coach is vic ruggiero .’	‘sing your song ’ writer is vic ruggiero .’	‘vic ruggiero .’	-
QA.1	‘did abraham lincoln have a dad’	*true*:’positive polarity’, *lp*:’dad’ *ro*:’abraham lincoln’	‘yes . abraham lincoln has at least one dad .’	‘yes , abraham lincoln had a mother.’	‘yes , abraham lincoln had a father.’	-
QA.1	‘what is the masters starting date’	*timepoint*:’1999’, *a*:’the masters’ *b*:’starting date’	‘1999 ’s starting date point is 1999 .’	‘the the masters ’s starting date point is 1999 .’	‘the masters was created on 1999 .’	-
QA.3	‘is canada bigger than united states’	*false*:’negative polarity’, *r*:’bigger than’, *y*:’united states’, *x*:’canada’	‘no , canada is not the bigger than united states .’	‘no , canada is not bigger than united states .’	‘no , canada is not bigger than united states .’	‘no , canada is not bigger than united states .’
QA.3	‘will ferrell’s wife’	*human being*: ’viveca paulin’, *a*:"will ferrell ’s", *b*:’wife’	‘will ferrell ’s wife is viveca paulin .’	‘will ferrell ’s ’s wife is viveca paulin .’	‘viveca paulin .’	‘will ferrell ’s wife is viveca paulin .’
QA.3	‘popsicle maker’	*business*:’unilever’, *a*:’popsicle’ *b*:’maker’	‘popsicle ’s maker is unilever .’	‘popsicle ’s manufacturer is unilever .’	‘unilever .’	popsicle ’s maker is unilever .’
SFX	-	**inform** (*name*:‘sanjalisco’, *kidsallowed*:‘yes’)	sanjalisco allows kid -s and is located	-	-	sanjalisco allows kid -s
SFX	-	**inform** (*name*:‘red door cafe’, *area*:‘cathedral hill’) *goodformeal*:‘breakfast’) *kidsallowed*:‘no’)	red door cafe is a nice restaurant in the cathedral hill does not allow kid -s and is good for breakfast	-	-	red door cafe is a nice restaurant in cathedral hill that is good for breakfast and does not allow kid -s
SFX	-	**inform** (*name*:‘darbar restaurant’, *food*:‘pakistani’) *goodformeal*:‘lunch’) *kidsallowed*:‘yes’)	darbar restaurant is a pakistani restaurant that allows kid -s and is good for lunch	-	-	darbar restaurant is a nice restaurant that serves pakistani food and allows kid -s

Table 7: Examples of different outputs from our models when given the same input Meaning Representation (and previous context when available) on two of our Question-Answering datasets (QA.1, QA.3) and on a task-based (SFX) dataset.

as the best one across all QA partitions.

Multitask Looking at the output of the models trained in a multi-task learning setting, we observe that the baseline tends to be more prone to grammatical errors compared to models jointly trained with another task (e.g. in Table 7 ‘sanjalisco allows kid-s and is located’). Due to multi-tasking the models generate more grammatically correct and natural responses for both SFX and QA.3.

9 Conclusions

In this work, we apply the traditional dialog MR-to-text approach to NLG to an open-domain QA setting, with sensibly larger ontologies compared to current task-oriented dialog approaches. Our goal was to test the reliability of current approaches to NLG for dialog in an environment where the number of slots could be substantial, a requirement that is critical to meet if we want to move towards an integrated NLG module across different domains.

The experiments presented show the feasibility of learning a NLG module for QA using a MR-to-text approach. NLG models performances on datasets with progressively bigger ontologies reported a continuous but not drastic decline for most metrics. Moreover, our multitask learning experiments showed that learning NLG models jointly for QA and task-oriented dialog improves single tasks performances in terms of fluency. Results across different experimental settings also point towards the vital role played by the previous utterance context (delexicalized and especially lexicalized) to improve NLG models for open-domain QA.

While we envision our approach as a first step towards an integrated statistical NLG module for a dialog system, still much remains to be done in order to achieve such a challenge. In this work, for example, we saw the importance of adapting approaches to NLG typical of task-oriented dialog when moving to an open-domain QA setting. This is important not only in terms of modelling (the essential role of the previous utterance), but also in terms of evaluation (designing metrics able to capture the relative importance of some slots in a given answer compared to others).

As our future work, we would like to expand our multi-task learning experiments to novel NLG datasets, for example recently proposed datasets of reviews (Oraby et al., 2019). Another possibility would be to explore transfer learning, rather than multi-task learning for NLG in the MR-to-text approach. Additionally, another interesting research direction would be the investigation of evaluation metrics for NLG in a QA setting, for example to better capture the centrality of some slots (or entities) compared to others when answering a given question.

References

Dzmitry Bahdanau, KyungHyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. *arXiv preprint arXiv:1409.0473*.

Anusha Balakrishnan, Jinfeng Rao, Kartikeya Upasani, Michael White, and Rajen Subba. 2019. Constrained decoding for neural nlg from compositional representations in task-oriented dialogue. *arXiv preprint arXiv:1906.07220*.

Jesse Dodge, Andreea Gane, Xiang Zhang, Antoine Bordes, Sumit Chopra, Alexander Miller, Arthur Szlam, and Jason Weston. 2015. Evaluating prerequisite qualities for learning end-to-end dialog systems. *arXiv preprint arXiv:1511.06931*.

Ondřej Dušek and Filip Jurcicek. 2016. A context-aware natural language generator for dialogue systems. In *Proceedings of the 17th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, pages 185–190.

Ondrej Dušek and Filip Jurcıcek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *The 54th Annual Meeting of the Association for Computational Linguistics*, page 45.

Ondřej Dušek, Jekaterina Novikova, and Verena Rieser. 2018. Findings of the e2e nlg challenge. In *Proceedings of the 11th International Conference on Natural Language Generation*, pages 322–328.

Yao Fu and Yansong Feng. 2018. Natural answer generation with heterogeneous memory. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 185–195.

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

He He, Anusha Balakrishnan, Mihail Eric, and Percy Liang. 2017. Learning symmetric collaborative dialogue agents with dynamic knowledge graph embeddings. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, volume 1, pages 1766–1776.

Matthew Henderson, Blaise Thomson, and Steve Young. 2014. Robust dialog state tracking using delexicalised recurrent neural networks and unsupervised adaptation. In *Spoken Language Technology Workshop (SLT), 2014 IEEE*, pages 360–365. IEEE.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, 9(8):1735–1780.

Juraj Juraska, Panagiotis Karagiannis, Kevin Bowden, and Marilyn Walker. 2018. A deep ensemble model with slot alignment for sequence-to-sequence natural language generation. In *Proceedings of the 2018 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies, Volume 1 (Long Papers)*, volume 1, pages 152–162.

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

Rémi Lebret, David Grangier, and Michael Auli. 2016. Neural text generation from structured data with application to the biography domain. In *Proceedings of the 2016 Conference on Empirical Methods in Natural Language Processing*, pages 1203–1213.

Percy Liang, Michael I Jordan, and Dan Klein. 2009. Learning semantic correspondences with less supervision. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP: Volume 1-Volume 1*, pages 91–99. Association for Computational Linguistics.

Thang Luong, Hieu Pham, and Christopher D Manning. 2015. Effective approaches to attention-based neural machine translation. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*, pages 1412–1421.

Hongyuan Mei, TTI UChicago, Mohit Bansal, and Matthew R Walter. 2016. What to talk about and how? selective generation using lstms with coarse-to-fine alignment. In *Proceedings of NAACL-HLT*, pages 720–730.

Stefano Mezza, Alessandra Cervone, Evgeny Stepanov, Giuliano Tortoreto, and Giuseppe Riccardi. 2018. Iso-standard domain-independent dialogue act tagging for conversational agents. In *Proceedings of the 27th International Conference on Computational Linguistics*, pages 3539–3551.

Neha Nayak, Dilek Hakkani-Tur, Marilyn Walker, and Larry Heck. 2017. To plan or not to plan? discourse planning in slot-value informed sequence to sequence models for language generation. In *Proc. of Interspeech*.

Jekaterina Novikova, Ondřej Dušek, and Verena Rieser. 2017. The e2e dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206.

Shereen Oraby, Vrindavan Harrison, Abteen Ebrahimi, and Marilyn Walker. 2019. Curate and generate: A corpus and method for joint control of semantics and style in neural nlg. *arXiv preprint arXiv:1906.01334*.

Amanda Stent, Matthew Marge, and Mohit Singhai. 2005. Evaluating evaluation methods for generation in the presence of variation. In *International Conference on Intelligent Text Processing and Computational Linguistics*, pages 341–351. Springer.

Andreas Stolcke, Klaus Ries, Noah Coccaro, Elizabeth Shriberg, Rebecca Bates, Daniel Jurafsky, Paul Taylor, Rachel Martin, Carol Van Ess-Dykema, and Marie Meteer. 2000. Dialogue act modeling for automatic tagging and recognition of conversational speech. *Computational linguistics*, 26(3):339–373.

Mengxi Wei and Yang Zhang. 2019. Natural answer generation with attention over instances. *IEEE Access*, 7:61008–61017.

Tsung-Hsien Wen, Milica Gašic, Nikola Mrkšic, Lina M Rojas-Barahona, Pei-Hao Su, David Vandyke, and Steve Young. 2016. Multi-domain neural network language generation for spoken dialogue systems. In *Proceedings of NAACL-HLT*, pages 120–129.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Conference Proceedings-EMNLP 2015: Conference on Empirical Methods in Natural Language Processing*, pages 1711–1721.

Jun Yin, Xin Jiang, Zhengdong Lu, Lifeng Shang, Hang Li, and Xiaoming Li. 2016. Neural generative question answering. In *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence*, pages 2972–2978. AAAI Press.