Syntactic Manipulation for Generating More Diverse and Interesting Texts

Jan Deriu
Zurich University of Applied Sciences
jan.deriu@zhaw.ch

Mark Cieliebak
Zurich University of Applied Sciences
mark.cieliebak@zhaw.ch

Abstract

Natural Language Generation plays an important role in the domain of dialogue systems as it determines how users perceive the system. Recently, deep-learning based systems have been proposed to tackle this task, as they generalize better and require less amounts of manual effort to implement them for new domains. However, deep learning systems usually adapt a very homogeneous sounding writing style which expresses little variation.

In this work, we present our system for Natural Language Generation where we control various aspects of the surface realization in order to increase the lexical variability of the utterances, such that they sound more diverse and interesting. For this, we use a Semantically Controlled Long Short-term Memory Network (SC-LSTM), and apply its specialized cell to control various syntactic features of the generated texts. We present an in-depth human evaluation where we show the effects of these surface manipulation on the perception of potential users.

1 Introduction

In this paper, we describe our end-to-end trainable neural network for producing natural language descriptions of restaurants from meaning representations (MR). Recently, data-driven natural language generation (NLG) systems have shown great promise, especially as they can be easily adapted to new data or domains. End-to-end systems based on deep learning can jointly learn sentence planning and sentence realization from unaligned data. However, a recurrent problem, which we found with the existing solutions for NLG, is that the generated utterances express a very homogeneous writing style. More precisely, most utterances start by using the restaurant name, the follow-up sentences usually begin with the pronoun "It", and each attribute-value pair is expressed using the same formulation across different utterances (see Table 1).

Green Man is a family friendly japanese restaurant in riverside near Express by Holiday Inn.
Clowns is a pub near Crowne Plaza Hotel with a customer rating of 5 out of 5.
Wildwood is an italian pub located near Raja Indian Cuisine in the city centre. It is not family-friendly.
The Cricketers provides chinese food in the 20-25 price range. It is located in the riverside. It is near All Bar One. Its customer rating is high.

Table 1: Examples to highlight the homogeneity of the utterances generated by state-of-the-art systems.

The publicly available E2E dataset by (Novikova et al., 2017) provides pairs of Meaning Representations (MR's) and several human generated reference utterances for the restaurant-domain. It is the first dataset to provide large amounts of training data with an open vocabulary, complex syntactic structures, and more variabilty in expressing the attributes. In this work, we exploit these characteristics of the dataset to generate utterances which express a higher diversity in their writing style. For this, we extend the Semantically Conditioned Long Short-term Memory Network (SC-LSTM) proposed by (Wen et al., 2015b) with surface features to control the manipulation of the surface realization.

Since the data contains a large variety of formulations for an attribute-value pair, a simple delexicalization of the utterance is not possible. This fact also increases the difficulty of evaluating the utterances for their correctness. Thus, we introduce a semantic reranking procedure based on classification algorithms trained to rate whether

Proceedings of The 11th International Natural Language Generation Conference, pages 22–34,
Tilburg, The Netherlands, November 5-8, 2018. ©2018 Association for Computational Linguistics

the attributes are rendered correctly.

We evaluate our model on the E2E dataset and report the BLEU, NIST, METEOR, ROUGE-L and CIDEr scores. We measure the diversity of the generated utterances by counting the number of different uni- and bi-grams. Further, to evaluate the correctness of the generated utterances, we employ a soft metric based on the aforementioned classifiers. Finally, we present an in-depth human evaluation where we measured the effects of these more diverse utterances on the perceptions of potential users. More precisely, humans evaluated the quality and naturalness of an utterance, which of the attributes comprehensible, concise, elegant, and professional fits to the text, and which of the different systems generated the most preferred outputs. We release the code and all the scripts.[1]

2 Related Work

The task of NLG is usually divided into separate subtasks such as content selection, sentence planning, and surface realization (Stent et al., 2004). Traditionally, the task has been solved by relying on rule-based methods, but these methods do not scale and are hardly adaptable to new domains. Recently, deep learning techniques have become more prominent for NLG. With these techniques, there now exists a large variety of different network architectures, each tackling a different aspect of NLG: (Wen et al., 2015b) propose an extension to the vanilla LSTM (Hochreiter and Schmidhuber, 1997) to control the semantic properties of an utterance, whereas (Hu et al., 2017) use variational autoencoder (VAE) and generative adversarial networks to control the generation of texts by manipulating the latent space; (Mei et al., 2016) employ an encoder-decoder architecture extended by a coarse-to-fine aligner to solve the problem of content selection; (Wen et al., 2016) apply data counter-fitting to generate out-of-domain training data for pretraining a model where there is little in-domain data available; (Semeniuta et al., 2017; Bowman et al., 2015) use a VAE trained in an unsupervised fashion on large amounts of data to sample texts from the latent space; and (Dušek and Jurcicek, 2016) use a sequence-to-sequence model with attention to generate natural language strings as well as deep syntax dependency trees from dialogue acts. All these approaches solve different aspects of the NLG task.

In our work, we tackle the aspect of generating texts that display more complex and diverse syntactic structures. The dialogue system community has proposed most work on this topic, as the end-to-end trainable algorithms tend to produce the same universal answer to each input. In (Li et al., 2016a) the authors develop a new loss function based on mutual information, (Li et al., 2016b) propose a new decoding algorithm based on a modified beam search, which favors hypotheses from different parent nodes. In (Li et al., 2017) the authors aim to increase the diversity by removing training examples, which are similar to the most commonly used utterances. In (Shao et al., 2017) the authors propose a sequence-to-sequence model with an augmented attention mechanism, which takes into account parts of the target sentence. Finally, the authors adapt the beam-search ranking to work at a segment level and, thus, injecting diversity earlier during the decoding.

3 Task Definition

Natural language generation for dialogue systems describes the task of converting a meaning representation (MR) into an utterance in a natural language. The E2E training data consist of 50k instances in the restaurant domain, where one instance is a pair of a MR and an example utterance or reference. The data is split into training, development and test in a 76.5%-8.5%-15%-ratio. Each MR consists of 3-8 attributes and their values, see Table 2 for the domain ontology. The split ensures that the MRs in the different dataset-splits are distinct. The dataset contains an open vocabulary and more complex syntactic structures than other similar datasets, as shown in the dataset definition (Novikova et al., 2017). Especially, it contains various ways of expressing a single value of an attribute: for instance, the value *1 of 5* is expressed in the data as “one star rated”, “rated with 1 of 5 stars”, or “rated one out of five”. In this work, we exploit this variety of formulation to produce utterances that express a more varied writing style.

4 Model

The goal of our model is to generate a text while providing the ability of controlling various semantic and syntactic properties of this text. Our model has two components: i) the generator and ii) semantic classifiers that rate the correctness of an ut-

[1] https://github.com/jderiu/e2e_nlg

Attribute	Type	Example Values
name	verbatim string	Alimentum, ..
eatType	dictionary	restaurant, pub, coffee shop
familyFriendly	boolean	yes, no
food	dictionary	Italian, French, English, ...
near	verbatim string	Burger King
area	dictionary	riverside, city center
customerRating	dictionary	1 of 5, 3 of 5, 5 of 5, low, average, high
priceRange	dictionary	<£20, £20-25, >£30 cheap, moderate, high

Table 2: Domain ontology of the E2E dataset.

terance.

We use the Semantically Conditioned Long Short-term Memory Network (SC-LSTM) proposed by (Wen et al., 2015b) as our generator, which has a specialized cell to process the one-hot encoded MR-vector. The semantic classifiers (SC) are trained for each attribute separately: they classify which value the generator rendered. With this, the correctness of an utterance can be determined, which is relevant when dealing with contradictory constraints during the generation of more diverse texts.

4.1 Semantically Conditioned LSTM

The SC-LSTM (Wen et al., 2015b) extends the original LSTM (Hochreiter and Schmidhuber, 1997) cell with a specialized cell, which processes the MR. The MR is represented as a one-hot encoded MR-vector d_0, which represents the value for each attribute. This cell assumes the task of the sentence planner, as it treats the MR-vector as a checklist to ensure that the information is fully represented in the utterance. The cell acts as a forget gate, keeping track of which information has already been consumed.

We briefly introduce the SC-LSTM as defined in (Wen et al., 2015b), which we will later on modify to meet our needs. Let $w_t \in \mathbb{R}^M$ be the input vector at time t, $d_t \in \mathbb{R}^D$ the MR-vector at time t, and N be the number of units of an SC-LSTM cell, then the formulation of the forward pass is defined as:

$$\begin{pmatrix} i_t \\ f_t \\ o_t \\ r_t \\ g_t \end{pmatrix} = \begin{pmatrix} \sigma \\ \sigma \\ \sigma \\ \sigma \\ \tanh \end{pmatrix} \mathbf{W}_{5n,2n} \begin{pmatrix} w_t \\ h_{t-1} \end{pmatrix}$$

$$\begin{aligned} d_t &= r_t * d_{t-1} \\ c_t &= i_t * g_t + f_t * c_{t-1} + \tanh(W_d d_t) \\ h_t &= o_t * \tanh(c_t) \end{aligned}$$

where σ is the sigmoid function, and $i_t, f_t, o_t, r_t \in [0,1]^N$ are the input, forget, output, and MR-reading gates, and $h_t, c_t \in [0,1]^N$ are the hidden state and the cell state. The weights $\mathbf{W}_{5n,2n}$ and $W_d \in \mathbb{R}^{D \times M}$ are the model parameters to be learned.

The prediction of the next token is performed by sampling from the probability distribution:

$$w_t \sim P(w_t | w_{0:t-1}, d_t) = \text{softmax}(W_s h_t)$$

where $W_s \in \mathbb{R}^{N \times M}$ is a weight matrix to be learned during training. During the training procedure the inputs to the SC-LSTM are the original tokens w_t from the training set. On the other hand, when generating new utterances we use the previously generated token as input to generate the next token.

Loss To ensure that the SC-LSTM consumes the MR correctly, two conditions are defined: i) the MR-vector at the last time step d_T has to be zero, which ensures that all the required information has been rendered, and ii) the gate should not consume too much of the dialogue act in one time step, i.e. the difference $\|d_t - d_{t-1}\|$ should be minimised. From these criteria, the reconstruction loss is adapted to:

$$F(\theta) = \sum_t p_t^T log(y_t) + \|d_T\| + \sum_{t=0}^{T-1} \eta \xi^{\|d_t - d_{t-1}\|}$$

where the first term is the reconstruction error, which sums the cross-entropy loss for each time step and the following two terms ensure the two criteria defined above.

Semantic Classifiers For each attribute a we train a CNN-based classifier D_a. Each classifier is trained to detect which of the possible values for the attribute a is rendered in the utterance or if the attribute is present in the utterance at all. We train the classifiers on the training set, where the input is the utterance and the output is the value for the attribute a, which is defined in the MR. These classifiers measure the semantic correctness of the produced utterances by comparing the output of the classifier to the MR. If the classifier output corresponds to the value defined in the MR then we regard the attribute as being rendered correctly.

5 Syntactic Control

The utterances produced by the basic model described in Section 4 lack syntactic variety, they all follow the trivial structure. To control the syntactic expressions of an utterance we expand the MR-vector with syntax specific features. More specifically, in this work we control three different surface features: i) the first word of the utterance, ii) the first word of each follow-up sentence in the utterance, and iii) for each attribute-value pair the formulation used to express it. For each of these control mechanisms, we produce one-hot encoded vectors and append these vectors to the MR-vector d_0. Through this mechanism, we provide the SC-LSTM with more prior information on the structure of the utterance. Thus, it learns to correlate how to render the surface based on the surface information provided. In the following, we describe the three control mechanisms in detail.

First Word Control Most utterances generated by the vanilla SC-LSTM begin by using the restaurant name. The main reason for this behaviour is that 59% of all utterances in the dataset have this characteristic. All the other starting words are used much less frequently: e.g. only 7% of all utterances start with the word "There", which is the second most used word. The model optimizes to generate the utterance, which yields the lowest average loss. Without additional information, this equates to the most common structure of utterances found in the training set. The first word used in an utterance greatly impacts how the rest of the utterance is rendered. Thus, using different first words increases the diversity of the rendered utterances. To generate more uncommon utterances, we provide the model with the information about the first word in the utterance during training. For this, we select all the words that appear more than $t = 60$ times as first word in the training data, which results in a set of $n = 20$ different words[2]. We then extend the MR-vector by adding a one-hot encoded vector $u_0 \in \mathbb{R}^{n+1}$, where the vector is set to '1' at the index of the first word in the utterance of the training sample. During the training, we use a dummy-index at $n + 1$ in case the first word of the utterance is not present in the list of first words. During test-time the first word is sampled from the set of n first words. To improve the semantic correctness we use the sampling procedure to over-generate, i.e. m different words are sampled to generate m different utterances. Using the semantic classifiers, the produced utterances are ranked by their correctness score.

Follow-up First Word Control We observe that the follow-up sentences in an utterance, which are produced by the vanilla SC-LSTM also follow the same pattern. More precisely, in cases where the utterance uses multiple sentences, the follow-up sentences usually begin with the pronoun 'It' which refers to the restaurant name mentioned in the first sentence. Similarly, to the First-Word-Control, we control the first word of follow-up sentences by using one-hot encoded vectors. The encoding states which word is used as first word of each follow-up sentence. As most utterances are composed between one and four sentences, we use three vectors to encode the first word of the first three follow-up sentences.

There are $n = 22$ different first words used in follow-up sentences, thus, each vector f_i is of length $n + 1$, where $i \in \{2, 3, 4\}$ denotes the sentence enumeration. We add an extra dimension to denote the case where the number of sentences is less than i. This representation provides the ability to control the first word used in each follow-up sentence as well as the number of sentences rendered.

Attribute-Value Formulation Control We observe that the vanilla SC-LSTM learns to use the most common formulation for an attribute-value pair. On average over all the attribute-value pairs, the most common formulation is used in 76% of the cases in the training set. It turns out that the most used formulation for most attribute-value pairs is equivalent to the surface form of the value itself. For example, the value "5 out of 5" is mostly expressed using the formulation: "... with a customer rating of 5 out of 5", instead of "It has an excellent customer rating" or other formulations.

To extract the different formulations of an attribute-value pair, we use a simple TF-IDF approach based on unigrams. For the complete list of formulations refer to Table 11 in Appendix A. For each attribute, we treat the utterances for each value as one document, thus, the corpus is made of as many documents as there are values for this attribute. The score is computed as $1 + \log(\text{tf}^a_{iv}) * \log(1 + \frac{N}{\text{df}^a_i})$ where tf^a_{iv} is the term frequency of

[2] $Name, Located, For, In, A, $Near, An, Near, There, On, $Food,The ,With ,Serving , If, At, Riverside, By, You, Family

term i for value v and df_i^a is the document frequency of term i in the documents of attribute a. We keep only those terms whose score is higher than 3. We apply manual filtering to clean the list from terms, which do not describe the attribute-value pair. With this method, we get on average 4.2 terms per attribute-value pair. We extend the MR-vector with one one-hot encoded vector for each attribute-value pair.

6 Experimental Setting

The goal for our application is to generate descriptions for restaurants. The dataset from (Novikova et al., 2017) contains 50k utterances for 5,751 different MRs. On average, each MR is composed of 5.43 attributes and there are 8.1 different references for each MR on average. For the evaluation, we report various corpus-based metrics: BLEU-4 (Papineni et al., 2002), NIST (Doddington, 2002) METEOR (Lavie and Agarwal, 2007), ROUGE-L (Lin, 2004), and CIDEr (Vedantam et al., 2015). Furthermore, we report various measures for lexical diversity: number of different tokens (#tokens), the type-token ratio (TTR) (Chotlos, 1944), the moving average type-token ratio (MSTTR) (Covington and McFall, 2010), and the measure of lexical diversity(MLTD) (McCarthy, 2005). Finally, we perform a human evaluation to measure the effect of the proposed manipulations on the user's perception.

Preprocessing Each utterance is treated as a string of characters, where each character is represented as a one-hot encoded vector. We replace the *name* and *near* values with the tokens 'X-name" and "X-near" respectively. The high diversity of the various formulations found for the attribute-value pairs, impedes us from replacing other attributes with placeholders. To generate the lexical features, we apply the Spacy-API[3] for word and sentence tokenization.

System Setup We train the SC-LSTM and the classifiers using AdaDelta (Zeiler, 2012) to optimize the loss function. We apply a softmax with decreasing temperature as proposed in (Hu et al., 2017) to approximate the discrete representation, which is used as input to the LSTM during the decoding stage. For the LSTM cell we use a hidden state of size 1024 and apply dropout as suggested in (Yarin and Ghahramani, 2016). For the classifiers we use a 2-layer CNN with 256 kernels of length 3.

We use our character-based version of the SC-LSTM (*vanilla*) as well as the sequence-to-sequence model by (Dušek and Jurcicek, 2016) (*tgen*) as baseline. We evaluate different versions of our model: the model where we control only the first word of the utterance (*utt-fw*), the model where we only control the first words of the follow-up sentences (*follow-fw*), the model where we only control the formulations of the attribute-value pairs (*form*), and the model where we control all three factors (*full*).

System	BLEU	NIST	METEOR	ROUGE_L	CIDEr
vanilla	0.634	8.270	0.428	0.653	1.9281
tgen	**0.661**	**8.550**	**0.446**	**0.687**	**2.201**
utt-fw	0.581	7.983	0.427	0.591	1.810
follow-fw	0.572	7.665	0.436	0.643	1.819
form	0.623	8.161	0.432	0.657	1.992
full	0.505	7.455	0.422	0.558	1.616

Table 3: Scores achieved for the corpus-based metrics by the different systems. The value of the best system for each score is highlighted in bold.

Output Generation The input to the system is a meaning representation (MR) which is converted into the MR-vector d_0. For each MR, the system samples the syntactic control values at random, i.e. it samples the first word of the utterance, the first words of each of the follow-up sentences and the formulation for each attribute-value pair randomly from the list of their respective possibilities. Then, these syntactic features are encoded into the one-hot format as described above. The input to the SC-LSTM is composed of both the MR-vector and the syntactic control vector. To ensure that the sampling of the syntactic features did not introduce semantic error, the system samples 10 different values for each of the three control types and produces one utterance for each combination, e.g. the *full* system produces 1000 sentences for each MR. We then use the classifiers (previously trained to evaluate if the utterance rendered the MR correctly) to rank the 1000 utterances w.r.t. their correctness. Finally, the system samples the final utterance from the set of utterances with the highest score (as there can be multiple utterances with the same score).

[3] https://spacy.io/

name	eatType	price	rating	near	food	area	fam.
1.0	0.97	0.90	0.84	0.99	0.95	0.94	0.91

Table 4: Validation Accuracy scores for each classifier.

System	vanilla	tgen	utt-fw	follow-fw	form	full
ERR$_{sc}$	0.158	0.192	0.093	0.100	0.100	**0.056**
ERR$_{rule}$	0.086	0.059	0.028	0.054	0.040	**0.015**

Table 5: Error Rate for each system, best system is highlighted in bold. The *sc* subscript denotes the scores computed by the classifiers.

7 Results

7.1 Evaluation Metrics

We report the scores for the automatic evaluation. This includes the metrics BLEU, ROUGE-L, METEOR, NIST, and CIDEr score, which rely on the comparison between the predicted utterance and multiple reference utterances. Table 3 shows that the surface manipulation leads to a decrease in all of these scores. The best scores for each metric is achieved by the *tgen* system. Its BLEU score is 3 points above the score achieved by *vanilla*. The *full* system achieved the lowest scores in each metric. Generally speaking, the deeper the impact of the syntactic manipulation the lower the word-overlap based score. This behaviour is explained by the fact that the baseline systems generate utterances which are syntactically similar to the most used structure in the gold-standard. The other systems generate sentences whose style and structure is much rarer in the gold-standard. For example, 59% of the reference utterances start with the standard pattern, whereas only 3% of the sentences generated by the *full* system follow this pattern. Although there are multiple reference utterances, it is not likely that one of these follows the syntactic choices of the syntactically controlled systems. Table 6 displays the various lexical diversity scores for each system as well as for the human-written text for reference. As expected, the human-written texts display the highest diversity across all scores. The *full* system achieves the highest scores out of all systems. Furthermore, both the *vanilla* and the *tgen* system obtain the lowest scores, thus, showing that the syntactic control mechanisms generate more diverse texts.

System	#tokens	TTR	MATTR	MTLD
vanilla	106	0.0070	0.5410	31.4811
tgen	120	0.0081	0.5175	30.5444
utt-fw	131	0.0082	0.5980	34.2865
follow-fw	141	0.0084	0.5745	33.5055
form	155	0.0098	0.5748	33.4892
full	**224**	**0.0134**	**0.6310**	**35.7831**
human	425	0.0280	0.6373	36.4466

Table 6: Diversity scores for each system and the human texts. The highest score of a system is marked in bold.

7.2 Classifier Performance

Since we use semantic classifiers to evaluate the correctness of the generated sentences, it is important to assess the quality of these classifiers. Table 4 shows the accuracy score for each of the classifiers on the testset. We note that all classifiers have a score greater than 0.9 except for the *customer rating*. The errors of the *customer rating* and the *price* classifiers stem from the semantic equivalence between the numerical and the verbal values which were used interchangeably in the references, e.g. when "price range is over £30" is expressed as "high-priced".

7.3 Correctness

We evaluate the correctness using a rule based system. We report the average error rate achieved by a system, as proposed by (Wen et al., 2015a), in Table 5, line **ERR**$_{rule}$. The best error-rate is achieved by the *full* system, followed by *utt-fw* and *form*. This shows that our approach to rerank the utterances with the semantic classifiers works very well. For comparison, we also report the error-rates when using the semantic classifiers themselves to determine the correctness of an utterance **ERR**$_{sc}$. It turns out that there is a mismatch between the scores achieved by the two metrics, especially for the *tgen* and *vanilla* system. This is due to the fact that the classifiers are used to filter the incorrect utterances, which leads the scores to be biased. Thus, it shows that the classifiers themselves are not suitable to compute a correctness score.

7.4 Qualitative Evaluation

In Table 8 two representative (cherry picked) examples are shown. For one MR we compare the outputs of all systems. In both examples the *tgen* and *vanilla* system produce utterances which follow the trivial pattern. The *uff-fw* and *full* systems produce a different style of utterance by starting the sentence with a preposition. The *follow-fw* system adds more variability to the utterance by starting the follow-up sentences with verbs (e.g.

"Located") or nouns ("Children") instead of pronouns referring to the restaurant name. The *form* system adds more variability by using different ways of phrasing an attribute-value pair (e.g. replacing "high price range" with "expensive"). We added a list of randomly sampled (non-cherry-picked) examples in Appendix B.

System	Quality	Naturalness
vanilla	3.979	**2.732***
tgen	4.013	2.591
utt-fw	4.007	2.605
follow-fw	3.992	2.576
form	**4.035**	2.577
full	4.033	2.540

Table 7: Quality and naturalness results from the user study. Here, * implies a statistical significant difference between a system and the *tgen* system, measured with two-tailed Student's t-test with $p < 0.05$

7.5 Human Evaluation

To measure the effectiveness of our approach, we performed an extensive human evaluation. For this, we recruited judges from the Figure-Eight[4] platform. For each experiment the sentence is rated by three different judges.

Quality and Naturalness To show that the syntactic manipulations do not deteriorate the utterances, we evaluated the *quality* and *naturalness* of the utterances produced by the different systems. Here, *quality* is defined to measure the grammatical correctness, the fluency and the correctness of the content, whereas *naturalness* measures the likelihood that the utterance was written by a human. For this, we sampled 250 MR's and generated the respective utterances for each system. The judges rated all utterances on a Likert scale from 1 to 5 for *quality* and on a scale from 1 to 3 for *naturalness*[5]. Table 7 shows the results for both the *quality* and *naturalness* evaluation. Statistical significance is measured by means of a two-tailed Student's t-test between the *tgen* system and the other systems. For *quality* there is no statistically significant difference between the *tgen* system and any other system. For *naturalness* there is no statistically significant between *tgen* and the syntactically controlled systems. However, there is a significant difference between *tgen* and *vanilla*. In fact, the *vanilla* system is rated significantly higher in terms of *naturalness* than any other system. For both metrics, the scores of all systems are very high, thus, we conclude that the syntactical control mechanisms do not deteriorate the utterances.

Subjective Analysis The main goal of the human evaluation is to understand how humans *perceive* the new utterances. For this, we compare the utterances of *tgen* and the *full* system by first sampling a MR, generate the utterance for each system, and let the human judges decide which of the two utterances they prefer. Since preference is a very subjective measure that might not give complete insight, we asked the judges to also state which utterance they find more *comprehensible* (is the utterance easier to understand), more *concise* (does the utterance convey the information clearly with as little text as possible), more *elegant* (is the utterance more nicely written, more poetic, display higher variability) and more *professional* (could this text be written by an experienced and well trained writer). Table 9 shows the ratio at which the system was selected for each of the five aforementioned categories alongside the inter-annotator agreement computed with the Fleiss' κ. The results show that none of the two systems is significantly preferred by the judges, nor is any of the two systems rated as being more comprehensible. However, the judges perceive the *full* system to produce significantly more elegant and professional utterances, i.e. in 71.6% of the comparisons the utterances by the *full* system were rated as more elegant and in 66.6% as being more professional. On the other hand, the judges rate the utterances of the *tgen* system as being significantly more concise (75%).

There is a moderate correlation between the preference of an utterance and the elegance (Spearman's Rho $\rho = 0.557$) or professionalism ($\rho = 0.569$). Furthermore, there is a weak correlation between comprehensibility of a system and the professionalism ($\rho = 0.468$). However, we found that there is a strong correlation between the preference and the comprehensibility of an utterance ($\rho = 0.719$). The evaluation shows that although the utterances by the *full* system are rated as being more elegant and professional, they are not necessarily preferred. The strongest indicator for preference is, thus, the comprehensibility.

[4] www.figure-eight.com

[5] For naturalness we asked if the utterance is likely to be written by a human, by a machine or if it is not clear

#	System	Utterance
1	**MR**	name=The Punter, eatType=pub, food=English, priceRange=high, area=city-centre, familyFriendly=no, near=Raja Indian Cuisine
	vanilla	The Punter is a pub that serves English food in the high price range and is located in the city centre near Raja Indian Cuisine.
	tgen	The Punter is an english pub in the city centre near Raja Indian Cuisine. It has a high price range and is not child friendly.
	utt-fw	**In** the city centre near Raja Indian Cuisine is a pub called The Punter. It serves English food and is not children friendly. The price range is high.
	follow-fw	The Punter is a pub that serves English food. **Located** in the city centre near Raja Indian Cuisine, it is not children friendly and has a high price range.
	form	The Punter is an **expensive** English Pub near Raja Indian Cuisine in the city centre. It is not child friendly.
	full	**If** you are looking for a pub serving English food, try The Punter. It is located in the city centre near Raja Indian Cuisine. **Prices** are on the **higher end** and it is not child friendly.
2	**MR**	name=Giraffe, eatType=restaurant, food=French, area=riverside, familyFriendly=yes, near=Raja Indian Cuisine
	vanilla	Giraffe is a family friendly restaurant that serves French food. It is located near Raja Indian Cuisine.
	tgen	Giraffe is a family friendly french restaurant near Raja Indian Cuisine in riverside.
	utt-fw	**A** French restaurant called Giraffe is located in the riverside area near Raja Indian Cuisine. It is child friendly.
	follow-fw	Giraffe is a restaurant that serves French food. **The** restaurant is located near Raja Indian Cuisine in the riverside area. **Children** are welcome.
	form	Giraffe is a French restaurant in the riverside area near Raja Indian Cuisine. It is family friendly.
	full	**In** the riverside area there is a French restaurant called Giraffe. **You** will find it near Raja Indian Cuisine. **Yes**, it is family friendly.

Table 8: Sample output of the vanilla SC-LSTM (V) and the First Word Control (F) for four different MRs where one attribute-value is changed.

Question	tgen	full	κ
Preference	0.476	0.523	0.587
Comprehensibility	0.476	0.523	0.555
Conciseness	0.750*	0.250	0.545
Elegance	0.283	0.716*	0.545
Professional	0.333	0.666*	0.529

Table 9: Results of the native speaking preference test. Significance is computed using a two-tailed binomial test. Where * denotes $p < 0.005$ and $N = 200$

Question	tgen	full	κ
Preference	0.593	0.406	0.456
Comprehensibility	0.682*	0.317	0.453
Conciseness	0.949**	0.050	0.312
Elegance	0.424	0.575	0.497
Professional	0.740**	0.259	0.342

Table 10: Results of the non-native speaking preference test. Significance is computed using a two-tailed binomial test, here * denotes $p < 0.05$ and ** denotes $p < 0.005$ and $N = 200$

Native vs. non-native speakers We observed that depending on whether the judges were native speaker or not the results were different. Thus, we repeated the same experiment by recruiting judges from non-native speaking countries[6]. Table 10 shows the results of the evaluation performed by the non-native speaking group. The differences of the ratings are significant. The non-native speakers rate the *tgen* system as significantly more comprehensible, more concise as well as more professional. There is still a high correlation between the preference and the comprehensibility of an utterance (Spearman's Rho $\rho = 0.709$). However, for the non-native group there is a significantly higher correlation between the comprehensibility and the professionalism of an utterance (Spearman's Rho $\rho = 0.628$) and a very high correlation between the preference and the professionalism (Spearman's Rho $\rho = 0.714$). This shows that the non-native speaking group finds it easier to understand the utterances produced by *tgen* and rates them as more preferable and more professional.
The evaluation shows that the two groups have different preferences and perceptions of the utterances. An in-depth analysis on the reasons behind these differences is left to future work. Our experiments indicate that the differences are due to the differences in language proficiency, as there is

[6] Judges were mostly recruited from eastern European countries and Asia.

a high correlation between the preference and the comprehensibility. However, to test this assumption, more characteristics about the judges need to be known.

8 Conclusion

In this work, we presented an end-to-end trainable deep-learning based system for the natural language generation task. With a simple control mechanism the utterances can be rendered more diverse and interesting. The human evaluation revealed that this control mechanism does not deteriorate the quality of the utterances in terms of semantic or grammatical errors. It further revealed that more diverse utterances are perceived as being more elegant and professional sounding to native speakers. Not surprisingly, the corpus-based metrics deteriorate when a more diverse vocabulary is used. One major challenge of this approach is the fact that during the generation the syntactic control features have to be sampled randomly to generate many utterances which have to be ranked and filtered. The solution to this inefficiency is part of future work.

References

Samuel R Bowman, Luke Vilnis, Oriol Vinyals, Andrew M Dai, Rafal Jozefowicz, and Samy Bengio. 2015. Generating sentences from a continuous space. *arXiv preprint arXiv:1511.06349*.

John W Chotlos. 1944. Iv. a statistical and comparative analysis of individual written language samples. *Psychological Monographs*, 56(2):75.

Michael A. Covington and Joe D. McFall. 2010. Cutting the gordian knot: The moving-average typetoken ratio (mattr). *Journal of Quantitative Linguistics*, 17(2):94–100.

George Doddington. 2002. Automatic evaluation of machine translation quality using n-gram cooccurrence statistics. In *Proceedings of the Second International Conference on Human Language Technology Research*, HLT '02, pages 138–145, San Francisco, CA, USA. Morgan Kaufmann Publishers Inc.

Ondřej Dušek and Filip Jurcicek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the 54th Annual Meeting of the Association for Computational Linguistics (Volume 2: Short Papers)*, pages 45–51. Association for Computational Linguistics.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neural computation*, pages 1735–1780.

Zhiting Hu, Zichao Yang, Xiaodan Liang, Ruslan Salakhutdinov, and Eric P. Xing. 2017. Toward controlled generation of tex. *International Conference on Machine Learning*, pages 1587–1596.

Alon Lavie and Abhaya Agarwal. 2007. Meteor: An automatic metric for mt evaluation with high levels of correlation with human judgments. In *Proceedings of the Second Workshop on Statistical Machine Translation*, StatMT '07, pages 228–231, Stroudsburg, PA, USA. Association for Computational Linguistics.

Jiwei Li, Michel Galley, Chris Brockett, Jianfeng Gao, and Bill Dolan. 2016a. A diversity-promoting objective function for neural conversation models. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 110–119. Association for Computational Linguistics.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2016b. A simple, fast diverse decoding algorithm for neural generation. *CoRR*, abs/1611.08562.

Jiwei Li, Will Monroe, and Dan Jurafsky. 2017. Data distillation for controlling specificity in dialogue generation. *CoRR*, abs/1702.06703.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, pages 74–81, Barcelona, Spain. Association for Computational Linguistics.

Philip M McCarthy. 2005. An assessment of the range and usefulness of lexical diversity measures and the potential of the measure of textual, lexical diversity (mtld). *Dissertation Abstracts International*, 66(12).

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.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the 18th Annual Meeting of the Special Interest Group on Discourse and Dialogue*, Saarbrücken, Germany. ArXiv:1706.09254.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. Bleu: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*, ACL '02, pages 311–318, Stroudsburg, PA, USA. Association for Computational Linguistics.

Stanislau Semeniuta, Aliaksei Severyn, and Erhardt Barth. 2017. A hybrid convolutional variational autoencoder for text generation. *arXiv preprint arXiv:1702.02390*.

Yuanlong Shao, Stephan Gouws, Denny Britz, Anna Goldie, Brian Strope, and Ray Kurzweil. 2017. Generating high-quality and informative conversation responses with sequence-to-sequence models. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, pages 2210–2219. Association for Computational Linguistics.

Amanda Stent, Rashmi Prasad, and Marilyn Walker. 2004. Trainable sentence planning for complex information presentation in spoken dialog systems. In *Proceedings of the 42Nd Annual Meeting on Association for Computational Linguistics*, ACL '04, Stroudsburg, PA, USA. Association for Computational Linguistics.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *The IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Tsung-Hsien Wen, Milica Gasic, Dongho Kim, Nikola Mrksic, Pei-Hao Su, David Vandyke, and Steve Young. 2015a. Stochastic language generation in dialogue using recurrent neural networks with convolutional sentence reranking. pages 275–284. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, 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 the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*, pages 120–129. Association for Computational Linguistics.

Tsung-Hsien Wen, Milica Gašić, Nikola Mrkšić, Pei-Hao Su, David Vandyke, and Steve Young. 2015b. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing (EMNLP)*. Association for Computational Linguistics.

Gal Yarin and Zoubin Ghahramani. 2016. A theoretically grounded application of dropout in recurrent neural networks. *Advances in neural information processing systems*, pages 1019–1027.

Matthew D Zeiler. 2012. Adadelta: an adaptive learning rate method. *arXiv preprint arXiv:1212.5701*.

A Formulations of Attribute-Values

Attribute	Value	Formulations
customer rating	1 out of 5	1, one, poor
	3 out of 5	3, three
	low	low, one, poor, poorly
	5 out of 5	5, five, excellent
	average	average, an, three, averagely
	high	high, highly, between, ranging
familyFriendly	no	not, non, adult, adults, no, allowed, allow
food	English	English, British, breakfast, traditional
	Fast food	fast, fries, joint, American, burger
	French	French, wine, cheese, fine, drinks
	Italian	Italian, pasta
	Japanese	Japanese, sushi, bar
priceRange	cheap	cheap, inexpensive
	high	high, expensive, higher, end
	moderate	moderate, moderately, mid, medium, pricing
	less than£20	20, less, than, under, pounds, inexpensive, below, lower
	£20-25	20, from, between, mid, 20-25, ranging, around
	more than £30	30, more, than, expensive, over, higher, above, costs, euros, costing

Table 11: The most important formulations that appear in the training set for each attribute-value pair. Pairs with just a single formulation were omitted.

B Sampled Utterances

#	System	Utterance
1	**MR**	name=The Wrestlers, eatType=restaurant, food=Italian, priceRange=moderate, area=riverside, familyFriendly=no, near=Raja Indian Cuisine,
	vanilla	The Wrestlers is a moderately priced Italian restaurant near Raja Indian Cuisine that is not kid friendly.
	tgen	The Wrestlers is a moderately priced italian restaurant in riverside near Raja Indian Cuisine. It is not kid friendly.
	utt	There is a moderately priced Italian restaurant called The Wrestlers in the riverside area near Raja Indian Cuisine. It is not kid friendly.
	follow	The Wrestlers is a restaurant providing Italian food in the moderate price range. The are located in the riverside area near Raja Indian Cuisine. The restaurant is not kid friendly.
	form	The Wrestlers is a restaurant that serves Italian food at a moderate price range. It is located in riverside near Raja Indian Cuisine and is not kid friendly.
	full	With a moderate price range, The Wrestlers is a restaurant that serves Italian food. It is located near Raja Indian Cuisine in the riverside area. No children allowed.
	human	The Wrestlers is a restaurant that serves Italian food. Has a moderate price range, and is kids friendly. Located in the riverside area near Raja Indian Cuisine.
2	**MR**	name=Zizzi, eatType=coffee shop, customer rating=average, near=Burger King,
	vanilla	Zizzi is a coffee shop near Burger King with an average customer rating.
	tgen	Zizzi is a coffee shop near Burger King. It has an average customer rating.
	utt	If you are looking for a coffee shop near Burger King, try Zizzi. It has an average customer rating.
	follow	Zizzi is a coffee shop near Burger King. Zizzi has an average customer rating.
	form	Zizzi is a coffee shop near Burger King. It has an average customer rating.
	full	If you are looking for a coffee shop near Burger King, then you could try Zizzi. It has an average customer rating.
	human	Customers rate Zizzi coffee shop, near Burger King, average.
3	**MR**	name=The Punter, eatType=restaurant, food=Italian, priceRange=cheap, customer rating=average, area=city centre, familyFriendly=yes, near=Rainbow Vegetarian Café,
	vanilla	The Punter is a family-friendly restaurant located in the city centre near Rainbow Vegetarian Café. It is cheap and has an average customer rating.
	tgen	The Punter is an italian restaurant near Rainbow Vegetarian Café in the city centre. It is family-friendly and has a cheap price range and an average customer rating.
	utt	Rainbow Vegetarian Café is a family-friendly restaurant called The Punter that serves Italian food and has an average customer rating. It is located in the city centre.
	follow	The Punter is a cheap Italian restaurant in the city centre near Rainbow Vegetarian Café. The Punter is family friendly and has an average customer rating.
	form	The Punter is an inexpensive Italian restaurant in the city centre near Rainbow Vegetarian Café. It is family friendly and has an average customer rating.
	full	In the city centre is a family-friendly restaurant called The Punter. This is a cheap Italian restaurant near Rainbow Vegetarian Café. It has an average customer rating.
	human	There is a cheap, restaurant that serves Italian, named The Punter, in the city centre near Rainbow Vegetarian Café. It has an average customer rating and is family friendly

Table 12: Randomly sampled output. A meaning representation is sampled at random, the respective utterance from each system is displayed.

#	System	Utterance
1	**MR**	name=The Cricketers, eatType=coffee shop, customer rating=average, familyFriendly=yes, near=Café Sicilia,
	vanilla	The Cricketers is a coffee shop near Café Sicilia. It is family friendly and has an average customer rating.
	tgen	The Cricketers is a family friendly coffee shop near Café Sicilia with an average customer rating.
	utt	If you are looking for a child friendly coffee shop near Café Sicilia, try The Cricketers. It has an average customer rating.
	follow	The Cricketers is a coffee shop near Café Sicilia that is family friendly and has an average customer rating.
	form	The Cricketers is a coffee shop near Café Sicilia. It is family friendly and has an average customer rating.
	full	For a family friendly coffee shop near Café Sicilia, try The Cricketers. Customers have rated it as average.
	human	Near Café Sicilia there is a coffee shop named The Cricketers. It's family friendly with an average rating.
2	**MR**	name=The Waterman, eatType=restaurant, food=Italian, priceRange=less than 20, area=riverside, familyFriendly=no, near=Raja Indian Cuisine,
	vanilla	The Waterman is a restaurant providing Italian food in the less than 20 price range. It is located in the riverside. It is near Raja Indian Cuisine.
	tgen	The Waterman is an italian restaurant in the riverside area near Raja Indian Cuisine. It is not family-friendly and has a price range of less than 20.
	utt	Italian restaurant The Waterman is located in the riverside area near Raja Indian Cuisine. It is not family-friendly and has a price range of less than 20.
	follow	The Waterman is a restaurant located near Raja Indian Cuisine in the riverside area. The price range is less than 20. They serve Italian food and are not family-friendly.
	form	The Waterman is a restaurant providing Italian food in the low price range. It is located in the riverside area near Raja Indian Cuisine. It is not family friendly.
	full	Italian food is served at The Waterman restaurant in the riverside area near Raja Indian Cuisine. Prices are under 20. Not family-friendly.
	human	The Waterman is a mature Italian restaurant near the riverside with a competitive menu for under 20, near the Raja Indian Cuisine.
3	**MR**	name=The Wrestlers, eatType=pub, food=Italian, priceRange=less than 20, area=riverside, familyFriendly=yes, near=Raja Indian Cuisine,
	vanilla	The Wrestlers is a family friendly pub near Raja Indian Cuisine in the riverside area that serves Italian food for less than 20.
	tgen	The Wrestlers is a family-friendly pub near Raja Indian Cuisine in the riverside area. It serves italian food for less than 20.
	utt	Italian food is served at The Wrestlers pub located near Raja Indian Cuisine in the riverside area. It is family friendly and has a price range of less than 20.
	follow	The Wrestlers is a pub that serves Italian food. They are located in the riverside area near Raja Indian Cuisine. They are family friendly and the price range is less than 20.
	form	The Wrestlers is a family friendly pub serving Italian food in the low price range. It is located in the riverside area near Raja Indian Cuisine.
	full	On the riverside near Raja Indian Cuisine is a family friendly pub called The Wrestlers. The price range is less than 20 and they serve Italian food.
	human	The Wrestlers is a pub in the low price range that serves pasta. It is located near Raja Indian Cuisine and has a public restroom.

Table 13: Randomly sampled output. A meaning representation is sampled at random, the respective utterance from each system is displayed.