Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation

Raheel Qader, Khoder Jneid, François Portet, Cyril Labbé

Generation of Company descriptions using concept-to-text and text-to-text deep models: dataset collection and systems evaluation

Raheel Qader[1] **Khoder Jneid**[1] **François Portet**[2] **Cyril Labbé**[2]
Univ. Grenoble Alpes, LIG
38000 Grenoble, France
[1]firstname.lastname@univ-grenoble-alpes.fr
[2]firstname.lastname@imag.fr

Abstract

In this paper we study the performance of several state-of-the-art sequence-to-sequence models applied to generation of short company descriptions. The models are evaluated on a newly created and publicly available company dataset that has been collected from Wikipedia. The dataset consists of around 51K company descriptions that can be used for both concept-to-text and text-to-text generation tasks. Automatic metrics and human evaluation scores computed on the generated company descriptions show promising results despite the difficulty of the task as the dataset (like most available datasets) has not been originally designed for machine learning. In addition, we perform correlation analysis between automatic metrics and human evaluations and show that certain automatic metrics are more correlated to human judgments.

1 Introduction

Traditional approaches of Natural Language Generation (NLG) consist in creating specific algorithms in the consensual NLG pipeline (Reiter and Dale, 2000; Gatt and Krahmer, 2018). However, recently there has been a very quick and strong interest in End-to-End (E2E) NLG systems in particular in the Dialogue community (Mairesse and Young, 2014; Wen et al., 2015; Dusek and Jurcícek, 2016) which are data-driven NLG methods jointly learning sentence planning and surface realization. Probably the most well known current effort is the E2E NLG Challenge (Novikova et al., 2017) which has generated a high number of submissions and whose task was to perform sentence planing and realization from dialogue act-based Meaning Representation (MR) on *unaligned* data. This challenge was a great success as it gathered the community around this problem of data-driven NLG models and showed the diversity of techniques that has been proposed to deal with the proposed task. The challenge also revealed that sequence-to-sequence (seq2seq) attention models such as TGEN(Dusek and Jurcícek, 2016) are competitive, yet, other simpler template-based approaches can still be effective (Puzikov and Gurevych, 2018). It also showed that although automatic metrics are useful for learning, they cannot be blindly used to predict human performances in NLG (Reiter and Belz, 2009; Puzikov and Gurevych, 2018). Furthermore, the E2E data contained a lot of redundancy of structure and a limited amount of concepts plus a least 5 references for the same MR input. This is an ideal case for machine learning but is it the one that is encountered in all E2E NLG applications?

In this work, we are interested in applying E2E models in a real world application in which there is a low amount of resources and whose output quality must be at human-level. The task is to produce a short description of a company given either a semi-structured set of slots (MR) or a textual document. This work is performed in the context of a research project with the Skopai company whose aim is to use AI technique to support startup description for attracting investors. More precisely, the task will be to generate an abstract for the article that contains the main factual information about a company.

In this research, we focus on seq2seq models in order to generate a summary for an article for two approaches: **concept-to-text** and **text-to-text**. As emphasized by (Gatt and Krahmer, 2018), there seem to be a convergence of NLG and summarization techniques, that is why for both approaches were recently applied in the

text-to-text domain. Furthermore, text-to-text approaches have to explicitly deal with content selection and document planning, tasks that were not the focus of the E2E challenge. In the literature, there has been few work related to our objective. For instance, in (Lebret et al., 2016), authors used a neural model to generate short biographical summaries from structured data (concepts) using a dataset collected from Wikipedia. Similarly, in (Chisholm et al., 2017) a sequence-to-sequence autoencoder with attention was used to generate biographies from Wikipedia. However, both of these work concentrate on generating mainly one sentence summaries, while our aim is to be able to generate longer summaries. In addition, since their generated summaries are extremely short, the information in the summaries is almost always is guaranteed to be present in the concepts, while in our dataset this is not the case. Finally, most Wikipedia biographies have very similar and repetitive writing styles which makes it easier for models to learn them, but in the case of company descriptions, the task is more challenging since most summaries are written in a different style.

The contribution of the paper can be summarized as follow:

- a collection of a realistic dataset for concept-to-text and text-to-text task that is made available to the community;
- the implementation and evaluation of several E2E models for the two tasks;
- an evaluation by naive human subjects and a comparison with the automatic metrics.

The paper starts by describing the dataset collection in Section 2 and then the seq2seq methods in Section 3. Corpus-based experiments and human evaluation are described in Section 4 and 5 respectively. The paper ends with a short discussion of the findings and an outlook for further work.

2 Dataset collection and Analysis

The task under study is to investigate the power of deep models on a specific task: the generation of company summaries. However, no large dataset corresponding to this task was available when the research started. Thus, a dataset about company descriptions has been collected, cleaned up, and organized for the task.

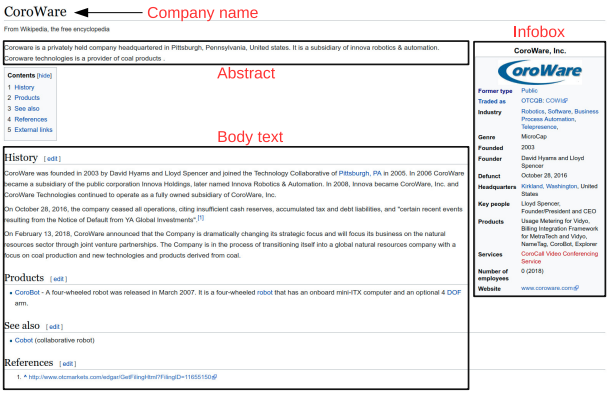

Figure 1: A Wikipedia page of a company. Abstract can be generated from the infobox (concept-to-text) or from the body text (text-to-txt).

2.1 Data source

Information about companies are typically be found in national company registers. However, they are not always accessible and are difficult to crawl. `dbpedia.org` also contains a semantically rich set of information about companies. However, the amount of companies is too small for a machine learning approach. A place were a large number of company descriptions can be found is Wikipedia. Wikipedia is a rich source of different types of data tackling a variety of topics, and the way articles are written in this source can be used for the task we address in this paper. As shown in Figure 1, an article contains an **abstract** followed by a table of content and then a **body** text. The abstract is a brief summary of the entire article containing the important ideas in the body text. Moreover, the top right side is the **infobox** (to be taken as MR), a panel containing semi-structured data about the company described in the article. As a result, English Wikipedia has been considered as a source for the dataset, and only articles about companies were collected.

2.2 Data collection method

The method to collect data was first to build a list of company name/id. This was performed from a dump of English Wikipedia (enwiki-20170820-pages-articles.xml.bz2) from which articles containing the terms “company” or “companies” in the “Category” section were retained. However a large number of articles were actually not company descriptions. Hence, we made use of the infobox attributes such as `Founded`, `Founder`, `Products`,

`Industry`, `Headquarters`, etc. which, in accordance to Wikipedia guidelines, should be found in company descriptions in Wikipedia. Then, articles which did not contain at least two company attributes in the infobox were dismissed from the list. At the end, we ended up with 64553 company links. The articles were then retrieved using the Wikipedia API. The abstract was directly extracted from the xml article as well as the infobox all stored into a json file as set of attribute value pairs. Articles that contained both an empty body and abstract were removed. Also those containing a too small amount of information were discarded leading to 51596 usable companies. Since the aim of the body text was to support single document summarization, information under the sections: `References`, `See also`, etc. were not needed. Thus the problem became to find out which section in each article indicates the end of the useful information. To do so, an analysis of the most frequent ending sections was performed. As a result, we end up with a list of 84 end headers (`reference`, `references`, `noteandreference`, etc.) chosen as a final maker of the body text. At the end of the process, 51k company were retrieved (excluded the empty articles).

For the infobox part, each attribute–value pair was represented as a sequence of string `attribute [value]`. Each attribute value which could contain a list was divided into at most 5 attributes (e.g., `attribute1 [value1] , attribute2 [value2] ... attribute5 [value5]`) using simple regex expression. Hence a string like *"Founder=[David Hyams and Lloyd Spencer]"* was converted into *"founder1[David Hyams], founder2[Lloyd Spencer]"*. At the end, the infobox is composed of 41 attributes with 4.5 attributes per article in average. The abstracts of the final dataset of 51k companies presents a vocabulary of size 158464 words.

2.3 Dataset characteristics

At the end of the process, although the dataset is faithful to the information found in Wikipedia, the dataset is not ideal for machine learning since the abstract, the body and the infobox are only loosely correlated. For instance, Figure 2 shows an abstract which is not based on information provided in the body text. Moreover, Figure 3 shows

Figure 2: Body text information is not correlated with the summary

that most of the abstract length is between 1 to 5 sentences while the body text size is much more spread with a peak at 1 sentence. The Pearson's correlation between abstract and body length (sentences) is very low $r = 0.275$ even when the body data of size 1 is removed ($r = 0.327$).

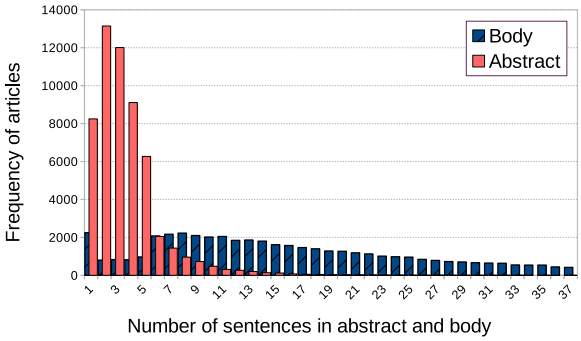

Figure 3: Distribution of abstract and body in terms of number of sentences.

In short, the problem found in the dataset can be summarized as follow. Given a article $a =< s, b, i >$ where s is the abstract b the body text, and i the infobox, the following problems exist:

- s is not guaranteed to be built from b.
- s and i does not always contains the same information.
- s, i and b vary greatly in terms of size and none of the size is correlated. Often, one or two of the sections are empty.

- there is only one version of s.
- there is no information about what was the objective of the writer(s) when producing s.

However, despites these problems, we believe this dataset represents a valuable resource since it represents the kind of data that can be found in real situations and that End2End systems must deal with in order to make a significant impact in society. The dataset is available for download[1].

3 E2E methods

3.1 Models

The basic model used for generating company description is based on the RNN seq2seq model architecture (Sutskever et al., 2014) which is divided into two main blocks: encoder which encodes the input sentence into fixed-length vector, and the decoder that decodes the vector into sequence of words. This model is able to treat sequence of words of variable size and has become the standard approach for many Natural Language Processing tasks. Briefly, a recurrent unit, at each step t takes an input x_t and a previous hidden state h_{t-1} and compute its hidden state and the output using:

$$h_t = \sigma_h(W_h x_t + U_h h_{t1} + b_h),$$

$$y_t = \sigma_y(W_y h_t + b_y),$$

where y_t is the output vector at each step; W, U, b are the parameters of the neural layer and σ_h and σ_y the activation functions of the neural layers. Once the encoder has read the entire input sequence of words (i.e., it read the special token $< EOS >$), the last hidden state h_t is passed to the decoder which begins to output a sequence of words using the previous hidden state and the previous predicted vector as input (using the special $< SOS >$ token as trigger) until it generates the end of a sequence (i.e., $< EOS >$). Numerous improvements have been made to this architecture such as using mono or multi layer of Long Short-Term Memory (LSTM) or Gated recurrent units (GRUs) to prevent the exploding/vanishing gradient problem and to model long dependencies in the sequence.

Another improvement is the attention mechanism introduced by (Bahdanau et al., 2014) which enables the decoder to attend on specific information in the input (encoder) to predict the next output. In that case, the decoder uses another information during the decoding which is the context vector c. At each step i and based on the sequence length T_x:

$$c_i = \sum_{j=1}^{T_x} \alpha_{ij}\, h_j.$$

The weight α_{ij} is computed as follows:

$$\alpha_{ij} = \frac{\exp(e_{ij})}{\sum_{k=1}^{T_x} \exp(e_{ik})},$$

where e_{ij} is computed as follows:

$$e_{ij} = a(s_{i-1}, h_j).$$

e_{ij} represents an alignment or attention model that tells the decoder at step i which part of hidden state of the input sequence to attend. The alignment model a can be a simple feed-forward neural network jointly trained with the rest of the architecture. The probability α_{ij}, reflects the importance of h_j with respect to the previous hidden state $i-1$ of the decoder in deciding the next state i and generating the output. Hence the decoder decides parts of the source sentence to pay attention to. This is particularly useful when the next word to output depends on an input word far apart in the input sequence. Note that this model, encoder-decoder with attention, is considered as a baseline model in almost all neural models in neural machine translation, text summarization, etc.

However, as pointed out by (See et al., 2017) the classical seq2seq models suffer from two commonly known problems: repetition of subsequences and wording off-topic (referred to as hallucination in the following).

Repetition is caused at the decoding stage, when the decoder relies too much on the previous output leading to infinite cycle. For instance if the decoder output 'to' then 'go' then 'to', it might happen that the next most probable word would be 'go' leading to an infinite 'to go to go to go to go to go to go...'. One way to deal with this problem is to use a coverage mechanism (Tu et al., 2016). This mechanism, used in machine translation, uses the attention weights to penalize the decoder for attending to input that has already been attended to previously. To do so, at each step t, the coverage vector $cov_t = \sum_{i=0}^{i=t-1} \alpha_i$ is computed, which is the sum of all the attention distributions until

[1] https://gricad-gitlab.univ-grenoble-alpes.fr/getalp/wikipediacompanycorpus

$t-1$. Which means that each source word coverage is the sum of the attentions it has received so far. Then, the coverage vector is added to the attention mechanism to have information of past attentions and then avoiding repetition. At the end, a new loss with a factor λ, specific to the coverage, can be computed and combined with the global loss

$$covloss = \lambda \sum_j min(\alpha_{ij}, cov_{i,j}).$$

Hallucination appears when some words are generated while there is no information related to these generated words in the input sequence. This can appear when the word to predict is infrequent in the training set and therefore has a poor word embedding making it close to a lot of other words. One way to deal with this problem would be to increase the training dataset but this is not always possible. Furthermore, any of such kind of systems will be likely to meet new unseen words such as company name and founder. Hence, a method to copy and past input word to the output has been developed in the translation domain and applied in the summarization community. The approach we have adopted is based on the Pointer-Generator Network (See et al., 2017) which computes a generation probability $p_{gen} \in [0, 1]$. This value evaluates the probability of 'generating' a word based on the vocabulary known by the model, versus copying a word from the source. The authors have implemented this pointing mechanism as:

$$P_{final}(w) = p_{gen}P_{vocab}(w) + (1 - p_{gen}) \sum_{i:w_i=w} a_i,$$

where $P_{final}(w)$ is the final probability of the word w, $P_{vocab}(w)$ is the probability of w as estimated by the model and $\sum_{i:w_i=w} a_i$ is the probability of w given the current attention it receives. In case w is the unknown word, then if the attention is high and p_{gen} sufficiently low then the input word will be used as output. It is important to note that in (See et al., 2017) p_{gen} is learned at the same time as the network.

3.2 concept-to-text approaches

The task of generating company description from a set of attribute–value pairs can be exemplified as going from the sequence

`name [Bodyarmor SuperDrink], founded [2011] , founder1 [Lance Collins] , founder2 [Mike Repole], headquarters [Queens, New York, United States], industry [Beverage manufacturing]`

to the sequence

Bodyarmor SuperDrink is an independently owned drink manufacturing company based in Queens, New York. It was founded in 2011 by Lance Collins and Mike Repole.

However, the Wikipedia company dataset that was collected contains a rather low rate of attribute–value pairs per article. Indeed, as mentioned in section 2.3 the target descriptions were often written using pieces of information that were not present in the infobox section. For this reason, an attribute–value augmentation step was considered using Natural Language Understanding (NLU) technique. Thus, the concept-to-text was done in two processing steps:

NLU : extraction of attribute–value pair from textual company descriptions

NLG : generation of company descriptions from a set of attribute–value pairs

The strategy employed to deal with the NLU part was to first create a reversed dataset where the company abstract is considered as the source sequence and the infobox attribute–value pairs as the target sequence, using a 5-fold cross-validation. More specifically, at each of the 5 turns, a 4/5 of the data was used to train a seq2seq character NLU model to infer the missing attributes in the unseen 1/5 of the data. During training, since ground truth was only partially available, inferred attribute–value pair was classified correct if:

- it corresponds to an attribute–value pair in the infobox ;
- it corresponds to an attribute–value pair whose similarity in the input text is high.

The similarity was computed using "difflib" library[2] of Python, which is an extension of the Ratcliff and Obershelp algorithm (Ratcliff and Metzener, 1988).

Using this method the dataset went from 304475 total attributes to 328682 in the augmented dataset. This is this augmented dataset that was used in all the experiments.

[2] `https://docs.python.org/2/library/difflib.html#difflib.SequenceMatcher`

For the NLG method, a basic seq2seq model with attention was used without any preprocessing. It is named C2T (Concept-to-Text). Also, a char to char model was used since it has been reported to be an effective model for the E2E challenge (Agarwal and Dymetman, 2017) despite a tendency to omit information or to repeat it on the E2E challenge data. The char to char model is an interesting way to deal with the rare word problem since the character vocabulary is very small and the network can learn to recompose unseen words. It is named C2T_char in the rest of the paper.

Another way to deal with rare word is to employ a pointing mechanism that is able to permit direct or indirect copy of input token in the output. Hence, the word seq2seq attention model was used with a pointer generator is called C2T+pg. Finally, to deal with repetitions, the coverage mechanism was added to the previous model to form the C2T+pg+cv system.

The recap of the systems under study:

1. C2T : the concept-to-text system word based seq2seq with attention model;
2. C2T_char : the concept-to-text system character based seq2seq with attention model;
3. C2T+pg : the concept-to-text system word based seq2seq with attention model with pointer generator;
4. C2T+pg+cv : same as above + coverage.

3.3 text-to-text approaches

Most text-to-text approaches require the ability to copy words or even sentences directly from the input to the output. Among the different models that we reviewed in Section 3, the Pointer-Generator Network has this capability, thus it was the only model used in the text-to-text experiments. The summary of the systems built to generate company description (abstract) from a source text (body text) are as the following :

5. Pointer-Generator Network (T2T+pg): deals with hallucination problem by having the ability to copy rare or unseen words during training while having the ability to generate words at the same time.
6. Coverage Model (T2T+pg+cv): deals with repetition problem by informing the decoder not to attend to input positions that have been repeatedly attended to. Note that this model is built on top of the Pointer-Generator Network.

4 Corpus based Experiment

4.1 Dataset formating

The original dataset has been through a limited amount of preprocessing for machine learning. For the C2T approaches, the dataset presented in Section 2 is filtered to contain only companies having abstracts of at least 7 words and at most 105 words. As a result of this process, 43681 companies are retained. Finally the dataset is partitioned to learning (35384), dev(3929) and test(4368) sets.

For the T2T approaches, the dataset is filtered at first to keep only the companies having abstracts with less than 105 tokens and bodies greater than 100 tokens while having the size of the abstract smaller than the size of the body text. As a result, 28034 are kept. The dataset is then splitted into three sets: training (21309), dev (2357) and test (4368).

In all the experiment the test set of 4368 companies is the same.

4.2 Corpus based evaluation

For the C2T and C2T_char experiments, we used the seq2seq model by Google[3], while for the C2T+pg, C2T+pg+cv, T2T+pg and T2T+pg+cv experiments, the Pointer-Generator Network implementation of (See et al., 2017)[4] was used. In addition, a baseline model called lead4 was also implemented. This baseline generates summaries by extracting the first 4 sentences from the article's body text.

The seq2seq model architecture has 2 layers of bidirectional LSTM trained using Adam optimization with learning rate of 0.001. As for the Pointer-Generator Network, it uses a single layer of bidirectional LSTM trained with AdaGrad and learning rate of 0.15. Both models have 256 hidden units for the encoder, decoders and embedding layers and a vocabulary size 50K (only for word models). The choice of hyper-parameters were determined by tunning the models on the dev set. Seq2seq models were trained until the loss on the dev set stops decreasing for several consecutive iterations. As for the Pointer-Generator Network

[3]https://github.com/google/seq2seq/

[4]https://github.com/abisee/pointer-generator

Table 1: Systems results on dev and test set using the E2E challenge metrics scripts provided with the baseline

	dev set					test set				
	BLEU	NIST	METEOR	ROUGE_L	CIDEr	BLEU	NIST	METEOR	ROUGE_L	CIDEr
lead4	0.0361	1.9599	**0.1282**	0.1645	0.0841	0.0364	2.0056	**0.1282**	0.1640	0.0908
C2T	0.0513	1.5784	0.0860	0.2032	0.1254	0.0608	1.9322	0.0906	0.2092	0.1872
C2T_char	**0.0648**	0.6390	0.1120	**0.2619**	**0.2351**	**0.0750**	1.0975	0.1159	0.2665	0.2731
C2T+pg	0.0327	0.0407	0.1014	0.2533	0.2198	0.0413	0.0893	0.1076	**0.2668**	**0.2836**
C2T+pg+cv	0.0400	0.2002	0.0975	0.2367	0.1888	0.0490	0.2349	0.1045	0.2589	0.2734
T2T+pg	0.0573	2.0101	0.1013	0.2232	0.2065	0.0567	1.9690	0.1002	0.2212	0.1992
T2T+pg+cv	0.0547	**2.1362**	0.1026	0.2214	0.1950	0.0558	**2.1188**	0.1024	0.2216	0.1974

and coverage models, we followed the strategy suggest in (See et al., 2017), i.e., to train the models with highly-truncated sequences then increase them during the training process until the maximum length is reached. Then the coverage mechanism is added and training is continued from the last training point of the Pointer-Generator Network.

Standard automatic measures BLEU (Papineni et al., 2002), ROUGE-L (Lin and Hovy, 2003), Meteor (Denkowski and Lavie, 2014) and CIDEr (Vedantam et al., 2015) were computed using the E2E challenge script. Table 1 shows evaluation results on both the dev and test sets for the lead4 (baseline), C2T and T2T tasks. The best system is difficult to extract from these results since there are close. However, C2T_char exhibits the best results for BLEU, ROUGE-L and CIDEr on the dev set while C2T+pg exhibits the best results for ROUGE-L and CIDEr on the test set. T2T+pg+cv shows the best NIST on both sets while lead4 is unbeatable from the METEOR perspective.

With respect to the results reported in the literature, such as the ones of the E2E challenge (for which the baseline system reaches: BLEU=0.6593; NIST=8.6094; METEOR=0.4483, ROUGE-L=0.6850, CIDEr=2.2338) these results are very low except for ROUGE-L. However, the main reason for such a large difference is that in the E2E challenge there are several references for each instance of the data. This leads to a higher ratio of match between the generated sentence words and the references ones, and thus, higher scores. In order to verify this, we conducted few tests using our C2T_char model on the E2E challenge data without any parameter tunning. The results showed that when only a single reference is counted, our model was able to achieve a score of 0.29 and 0.47 for BLEU and ROUGE-L respectively. However, when multiple references were included, the scores increased to 0.51 and 0.61 for BLEU and ROUGE-L. This clearly shows that the E2E challenge dataset is closer to the ideal case for machine learning than our case. Thus our results should not be directly compared with the E2E challenge.

However, we also computed the F1 of ROUGE 1, 2 and L score using the pyrouge package[5] and compared to recent summarization methods[6]. In that case C2T+pg exhibits the best results for ROUGE-1 (.3346) ROUGE-2 (.1701) and ROUGE-L (.3132) on the test set. These results are comparable to the abstractive method of (Nallapati et al., 2016) (ROUGE-1=.3546, ROUGE-2=.1330, ROUGE-L=.3265) and the pointer generation approach of (See et al., 2017) (ROUGE-1=.3644, ROUGE-2=.1728, ROUGE-L=.3342). However, they were both tested on the CNN/Daily Mail test set, for which no problem of content mismatch between documents and summaries were reported while it is a difficulty of our dataset. Furthermore, the difference with the E2E challenge is important since our dataset contains a large vocabulary, a large number of named entities and only one –not always reliable– reference summary. The few number of reference summaries give fewer opportunity for the models output to match n-grams in the references than when multiple references are available.

Although such metrics suggest our models are far from achieving satisfying results, they give in fact little insight about the actual weakness of the models. Moreover correlation between automatic and human-based metrics in NLG is still debatable (Gatt and Krahmer, 2018). That is why we conducted a human evaluation as well.

[5] https://pypi.python.org/pypi/pyrouge/0.1.0

[6] Also the ROUGE-L value given by the two scripts were not the same due to different parameters, we checked a very high (>.96) and significant Spearman correlation between all ROUGE value.

5 Human Evaluation

In order to gain more insight about the generation properties of each model a human evaluation with 19 human subjects was performed. We set up a web-based experiment which was circulated inside the lab but to people who were not involved in this project. The 4 questions below were asked on a 5-point Lickert scale:

Q1 How do you judge the Information Coverage of the company summary : 1 no information, 5 contains everything

Q2 How do you judge the Non-Redundancy of Information in the company summary. 1: means lots of repeated information, 5: no repetition.

Q3 How do you judge the Semantic Adequacy of the company summary? 1: lots of semantical mistakes, 5: semantically very correct.

Q4 How do you judge the Grammatical Correctness of the company summary? 1: very incorrect, 5: very good

We did not include fluency in the question since it is often correlated with grammar and because participants have difficulty to judge this property. Q1 to 4 were specifically designed to measure the recurrent weakness of seq2seq models: content selection, repetition, hallucination and bad segment connection.

Participants were exposed to a screen where a background (extract of the original Wikipedia body text, cut to 400 max), an infobox and a summary were visible all together in the screen. After reading the background, the infobox and the summary, the participant could answer the question by scrolling down. Not limit of time was imposed. A first example was given for training, then each participant had to treat 10 summaries. The participant could not go to the next step without explicitly answering all questions. In average one session last 15 minutes. At no time participants have been aware that one of the summary was human generated (i.e., the Wikipedia abstract).

30 companies were selected from the 4368 companies of the test set. They were selected based on the number of views during the month preceding the experiment. The less viewed one were retained to avoid participants judging well known companies.

Results of the human experiment are reported in Table 2. The first line report the result of the reference (i.e., the Wikipedia abstract) for comparison. It is clear from the coverage metric that no system nor the reference was seen as doing a good job at conveying the information. It is a known problem of the Wikipedia dataset and the systems were not able to do better than the reference. Non-redundancy metric gives a more contrasted view of the systems. C2T+pg was judged to be the least repetitive after the reference, while C2T_char to be the most repetitive. Regarding semantic correctness, C2T+pg is clearly above the others again including the reference. Same observation can be made for grammatical correctness.

Table 2: Results of the human evaluation per system.

	cover.	non-redun.	semant.	gramm.
reference	3.1	4.6	3.9	4.2
C2T	2.9	2.9	3.3	3.6
C2T_char	2.3	3.9	2.8	3.0
C2T+pg	2.3	4.5	4.0	4.3
C2T+pg+cv	2.7	3.9	3.6	4.2
T2T+pg	1.8	3.3	2.9	3.7
T2T+pg+cv	2.3	3.8	2.4	3.5

These results of human evaluation were compared to those of the automatic metrics (excluding the reference one). The correlation matrix is given in Figure 4. It can be seen that among automatic metrics, METEOR, ROUGE-L and CIDEr are highly correlated. When it comes to human vs automatic metrics, it is obvious that CIDEr has a highest correlation with semantic and grammar. It is worth noting that ROUGE-L is is also highly correlated to semantic and grammar.

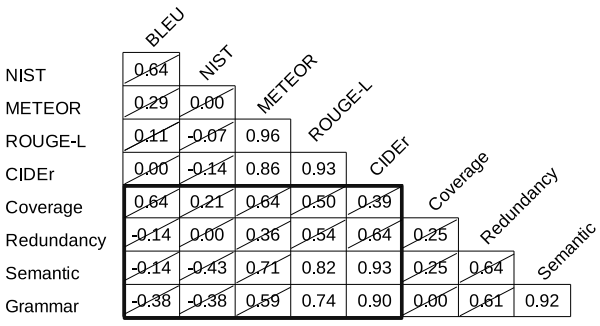

Figure 4: Correlation values based on Spearman's ρ. Human vs automatic metric correlations are in the black square. Crossed area are not significant correlation ($p>.05$).

Table 3: Sample of generated summaries from the test set using our systems along with the reference infobox, abstract and body text. Green color indicates repeated information and red color indicates factual errors.

Infobox: name1[rgb entertainment], headquarters1[argentina], founded1[2000], industry1[television production], type1[production company], owner1[gustavo yankelevich y victor gonzalez]
Body text (truncated): it was created in 2000 by gustavo yankelevich, former telefe director, and victor gonzalez, with headquarters in buenos aires, argentina and sao paulo, brazil. it include creation and production of television shows, films, cds, live events and multitudinous events. the company co-produces all cris morena productions...
Reference abstract: rgb entertainment is a production company from argentina. it was established in the year 2000.
C2T_char: rgb entertainment is an argentine television production company based in argentina. the company was founded in 2000 by gustavo yankelevich **yankelevich yankelevich** y victor gonzalez **in 2000**.
C2T+pg: rgb entertainment is a television production company in argentina .
T2T+pg: the argentine channel productions is an **american** film production and distribution company . the company was founded in 2000 by gustavo yankelevich and victor gonzalez in **buenos** , argentina . **it is owned by ideas group**.

6 Discussion and further work

Participants did not always understand the first question. They use English daily as working language but they were not native English speakers, that might have had an influence on the grammatical evaluation. However, English Wikipedia content is not always written by English natives and the level of English employed in the summary was quite standard.

C2T+pg capability is more emphasized by human evaluation than automatic metrics. Once again it shows that not all the automatic metrics are correlated with the human evaluation and that both evaluations are necessary to understand strengths and weaknesses of models. Despite this, some surprising correlation between semantics, grammar, CIDEr and ROUGE-L can be observed. However this findings are not in line with what was observed in (Shimorina, 2018), as they report only one significant correlation which is between semantics and METEOR. However, CIDEr and ROUGE-L are themselves highly correlated with METEOR. Nevertheless, this difference might be from the fact that our human evaluation questions are not exactly the same, thus, the answer of the subjects for certain questions might have been influenced by the other questions.

In order to better analyze the results, in Table 3 we show samples of generated summaries by some of our systems. The first remark that can be noticed is that the reference abstract does not contain some of the information given in the infobox, e.g., owners. This mismatch between the reference abstract and the infobox can be observed throughout all the corpus. This obviously poses a limitation on the models to learn to generate all the information given in the infobox. Then when it comes to our models, it can be seen that the C2T_char manages to generate all the infobox information but it has repetition problem. The T2T+pg, on the other hand, is not so behind when it comes to information coverage, however this models suffers from the problem of hallucination as it can be seen in its last sentence. Finally the C2T+pg manages to generate a correct but too short sentence which is lacking some information of the infobox.

Some weakness of the current C2T approaches may be due to the NLU model. A possible future work might be to deal with the weakness of the database and to perform more joined learning of NLU/NLG and to evaluate models on the real company database provided by the company that we are working with on a research project. In addition, we could also force the models to generate more guided summaries by taking both the infobox and the body text as input. In this way, the model can learn to do text-to-text and concept-to-text at the same time by giving more priority to sentences of the body text containing infobox values.

Acknowledgments

This project was partly funded by the IDEX Université Grenoble Alpes innovation grant (AI4I-2018-2019) and the Région Auvergne-Rhône-Alpes (AISUA-2018-2019).

References

Shubham Agarwal and Marc Dymetman. 2017. A surprisingly effective out-of-the-box char2char model on the E2E NLG challenge dataset. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue, Saarbrücken*, pages 158–163.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *Proceedings of the International Conference on Learning Representations (ICLR)*.

Andrew Chisholm, Will Radford, and Ben Hachey. 2017. Learning to generate one-sentence biographies from wikidata. In *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics*, pages 633–642.

Michael Denkowski and Alon Lavie. 2014. Meteor universal: Language specific translation evaluation for any target language. In *Proceedings of the Ninth Workshop on Statistical Machine Translation*, pages 376–380.

Ondrej Dusek and Filip Jurcícek. 2016. Sequence-to-sequence generation for spoken dialogue via deep syntax trees and strings. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 45–51.

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

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

Chin-Yew Lin and Eduard Hovy. 2003. Automatic evaluation of summaries using n-gram co-occurrence statistics. In *Proceedings of the North American Chapter of the Association for Computational Linguistics on Human Language Technology*, pages 71–78.

François Mairesse and Steve J. Young. 2014. Stochastic language generation in dialogue using factored language models. *Computational Linguistics*, pages 763–799.

Ramesh Nallapati, Bowen Zhou, Cicero dos Santos, Ça glar Gulçehre, and Bing Xiang. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. In *Proceedings of the SIGNLL Conference on Computational Natural Language Learning (CoNLL)*, pages 280–290.

Jekaterina Novikova, Ondrej Dušek, and Verena Rieser. 2017. The E2E dataset: New challenges for end-to-end generation. In *Proceedings of the Annual SIGdial Meeting on Discourse and Dialogue*, pages 201–206.

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

Yevgeniy Puzikov and Iryna Gurevych. 2018. E2e nlg challenge: Neural models vs. templates.

John W Ratcliff and David E Metzener. 1988. Pattern-matching-the gestalt approach. *Dr Dobbs Journal*, pages 46–51.

Ehud Reiter and Anja Belz. 2009. An investigation into the validity of some metrics for automatically evaluating natural language generation systems. *Computational Linguistics*, pages 529–558.

Ehud Reiter and Robert Dale. 2000. *Building Natural Language Generation Systems*. Studies in Natural Language Processing. Cambridge University Press.

Abigail See, Peter J Liu, and Christopher D Manning. 2017. Get to the point: Summarization with pointer-generator networks. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*, pages 1073–1083.

Anastasia Shimorina. 2018. Human vs automatic metrics: on the importance of correlation design. *CoRR*, abs/1805.11474.

Ilya Sutskever, Oriol Vinyals, and Quoc V Le. 2014. Sequence to sequence learning with neural networks. In *Proceedings of the Advances in Neural Information Processing Systems*, pages 3104–3112.

Zhaopeng Tu, Zhengdong Lu, Yang Liu, Xiaohua Liu, and Hang Li. 2016. Modeling coverage for neural machine translation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics, (ACL)*, pages 76–85.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 4566–4575.

Tsung-Hsien Wen, Milica Gasic, Nikola Mrksic, Pei-hao Su, David Vandyke, and Steve J. Young. 2015. Semantically conditioned lstm-based natural language generation for spoken dialogue systems. In *Proceedings of the Empirical Methods in Natural Language Processing, EMNLP 2015, Lisbon, Portugal, September 17-21, 2015*, pages 1711–1721.