

Generating Descriptions from Structured Data Using a Bifocal Attention Mechanism and Gated Orthogonalization

Preksha Nema^{†*} Shreyas Shetty^{†*} Parag Jain^{•*}
Anirban Laha^{•†} Karthik Sankaranarayanan[•] Mitesh M. Khapra^{†‡}

[†]IIT Madras, India [•]IBM Research

[‡] Robert Bosch Center for Data Science and Artificial Intelligence, IIT Madras

{preksha, shshett, miteshk}@cse.iitm.ac.in

{pajain34, anirlaha, kartsank}@in.ibm.com

Abstract

In this work, we focus on the task of generating natural language descriptions from a structured table of facts containing *fields* (such as nationality, occupation, *etc.*) and *values* (such as Indian, {actor, director}, *etc.*). One simple choice is to treat the table as a sequence of *fields* and *values* and then use a standard seq2seq model for this task. However, such a model is too generic and does not exploit task-specific characteristics. For example, while generating descriptions from a table, a human would attend to information at two levels: (i) the fields (macro level) and (ii) the values within the field (micro level). Further, a human would continue attending to a field for a few timesteps till all the information from that field has been rendered and then never return back to this field (because there is nothing left to say about it). To capture this behavior we use (i) a fused bifocal attention mechanism which exploits and combines this micro and macro level information and (ii) a gated orthogonalization mechanism which tries to ensure that a field is remembered for a few time steps and then forgotten. We experiment with a recently released dataset which contains fact tables about people and their corresponding one line biographical descriptions in English. In addition, we also introduce two similar datasets for French and German. Our experiments show that the proposed model gives 21% relative improvement over a recently proposed state of the art method and 10% relative improvement over basic seq2seq models. The code and the datasets developed as a part of this work are publicly available.¹

* The first three authors have contributed equally to this work.

¹https://github.com/PrekshaNema25/StructuredData_To_Descriptions

1 Introduction

Rendering natural language descriptions from structured data is required in a wide variety of commercial applications such as generating descriptions of products, hotels, furniture, *etc.*, from a corresponding table of facts about the entity. Such a table typically contains $\{field, value\}$ pairs where the field is a property of the entity (*e.g.*, *color*) and the value is a set of possible assignments to this property (*e.g.*, *color = red*). Another example of this is the recently introduced task of generating one line biography descriptions from a given Wikipedia infobox (Lebret et al., 2016). The Wikipedia infobox serves as a table of facts about a person and the first sentence from the corresponding article serves as a one line description of the person. Figure 1 illustrates an example input infobox which contains fields such as Born, Residence, Nationality, Fields, Institutions and Alma Mater. Each field further contains some words (*e.g.*, particle physics, many-body theory, *etc.*). The corresponding description is coherent with the information contained in the infobox.

Note that the number of fields in the infobox and the ordering of the fields within the infobox varies from person to person. Given the large size (700K examples) and heterogeneous nature of the dataset which contains biographies of people from different backgrounds (sports, politics, arts, *etc.*), it is hard to come up with simple rule-based templates for generating natural language descriptions from infoboxes, thereby making a case for data-driven models. Based on the recent success of data-driven neural models for various other NLG tasks (Bahdanau et al., 2014; Rush et al., 2015; Yao et al., 2015; Chopra et al., 2016; Nema et al., 2017), one simple choice is to treat the infobox as


	Born	1943
	Residence	India
	Nationality	Indian
	Fields	Particle physics, many-body theory, dynamical systems, stochastic processes, quantum dynamics, mechanical behavior of solids, and others
	Institutions	TIFR IIT Madras
	Alma mater	St. Stephens College Delhi University Brandeis University

Figure 1: Sample Infobox with description : V. Balakrishnan (born 1943 as Venkataraman Balakrishnan) is an Indian theoretical physicist who has worked in a number of fields of areas, including particle physics, many-body theory, the mechanical behavior of solids, dynamical systems, stochastic processes, and quantum dynamics.

a sequence of {field, value} pairs and use a standard seq2seq model for this task. However, such a model is too generic and does not exploit the specific characteristics of this task as explained below.

First, note that while generating such descriptions from structured data, a human keeps track of information at two levels. Specifically, at a macro level, she would first decide which field to mention next and then at a micro level decide which of the values in the field needs to be mentioned next. For example, she first decides that at the current step, the field *occupation* needs attention and then decides which is the next appropriate occupation to attend to from the set of occupations (*actor*, *director*, *producer*, etc.). To enable this, we use a bifocal attention mechanism which computes an attention over *fields* at a macro level and over *values* at a micro level. We then fuse these attention weights such that the attention weight for a field also influences the attention over the values within it. Finally, we feed a fused context vector to the decoder which contains both field level and word level information. Note that such two-level attention mechanisms (Nallapati et al., 2016; Yang et al., 2016; Serban et al., 2016) have been used in the context of unstructured data (as opposed to structured data in our case), where at a macro level one needs to pay attention to sentences and at a micro level to words in the sentences.

Next, we observe that while rendering the output, once the model pays attention to a field (say, occupation) it needs to stay on this field for a few timesteps (till all the occupations are produced in

the output). We refer to this as the *stay on* behavior. Further, we note that once the tokens of a field are referred to, they are usually not referred to later. For example, once all the occupations have been listed in the output we will never visit the occupation field again because there is nothing left to say about it. We refer to this as the *never look back* behavior. To model the *stay on* behaviour, we introduce a forget (or remember) gate which acts as a signal to decide when to forget the current field (or equivalently to decide till when to remember the current field). To model the *never look back* behaviour we introduce a gated orthogonalization mechanism which ensures that once a field is forgotten, subsequent field context vectors fed to the decoder are orthogonal to (or different from) the previous field context vectors.

We experiment with the WIKIBIO dataset (Lebret et al., 2016) which contains around 700K {infobox, description} pairs and has a vocabulary of around 400K words. We show that the proposed model gives a relative improvement of 21% and 20% as compared to current state of the art models (Lebret et al., 2016; Mei et al., 2016) on this dataset. The proposed model also gives a relative improvement of 10% as compared to the basic seq2seq model. Further, we introduce new datasets for French and German on the same lines as the English WIKIBIO dataset. Even on these two datasets, our model outperforms the state of the art methods mentioned above.

2 Related work

Natural Language Generation has always been of interest to the research community and has received a lot of attention in the past. The approaches for NLG range from (i) rule based approaches (e.g., (Dale et al., 2003; Reiter et al., 2005; Green, 2006; Galanis and Androutsopoulos, 2007; Turner et al., 2010)) (ii) modular statistical approaches which divide the process into three phases (planning, selection and surface realization) and use data driven approaches for one or more of these phases (Barzilay and Lapata, 2005; Belz, 2008; Angeli et al., 2010; Kim and Mooney, 2010; Konstas and Lapata, 2013) (iii) hybrid approaches which rely on a combination of hand-crafted rules and corpus statistics (Langkilde and Knight, 1998; Soricut and Marcu, 2006; Mairesse and Walker, 2011) and (iv) the more recent neural network based models (Bahdanau et al., 2014).

Neural models for NLG have been proposed in the context of various tasks such as machine translation (Bahdanau et al., 2014), document summarization (Rush et al., 2015; Chopra et al., 2016), paraphrase generation (Prakash et al., 2016), image captioning (Xu et al., 2015), video summarization (Venugopalan et al., 2014), query based document summarization (Nema et al., 2017) and so on. Most of these models are data hungry and are trained on large amounts of data. On the other hand, NLG from structured data has largely been studied in the context of small datasets such as WEATHERGOV (Liang et al., 2009), ROBOCUP (Chen and Mooney, 2008), NFL RECAPS (Barzilay and Lapata, 2005), PRODIGY-METEO (Belz and Kow, 2009) and TUNA Challenge (Gatt and Belz, 2010). Recently Mei et al. (2016) proposed RNN/LSTM based neural encoder-decoder models with attention for WEATHERGOV and ROBOCUP datasets.

Unlike the datasets mentioned above, the biography dataset introduced by Lebre et al. (2016) is larger (700K {table, descriptions} pairs) and has a much larger vocabulary (400K words as opposed to around 350 or fewer words in the above datasets). Further, unlike the feed-forward neural network based model proposed by (Lebre et al., 2016) we use a sequence to sequence model and introduce components to address the peculiar characteristics of the task. Specifically, we introduce neural components to address the need for attention at two levels and to address the *stay on* and *never look back* behaviour required by the decoder. Kiddon et al. (2016) have explored the use of checklists to track previously visited ingredients while generating recipes from ingredients. Note that two-level attention mechanisms have also been used in the context of summarization (Nallapati et al., 2016), document classification (Yang et al., 2016), dialog systems (Serban et al., 2016), etc. However, these works deal with unstructured data (sentences at the higher level and words at a lower level) as opposed to structured data in our case.

3 Proposed model

As input we are given an infobox $\mathcal{I} = \{(g_i, k_i)\}_{i=1}^M$, which is a set of pairs (g_i, k_i) where g_i corresponds to *field* names and k_i is the sequence of corresponding *values* and M is the total number of fields in \mathcal{I} . For example, $(g =$

occupation, k = actor, writer, director) could be one such pair in this set. Given such an input, the task is to generate a description $y = y_1, y_2, \dots, y_m$ containing m words. A simple solution is to treat the infobox as a sequence of *fields* followed by the *values* corresponding to the field in the order of their appearance in the infobox. For example, the infobox could be flattened to produce the following input sequence (the words in bold are field names which act as delimiters)

[Name] John Doe **[Birth Date]** 19 March 1981 **[Nationality]** Indian

The problem can then be cast as a seq2seq generation problem and can be modeled using a standard neural architecture comprising of three components (i) an input encoder (using GRU/LSTM cells), (ii) an attention mechanism to attend to important values in the input sequence at each time step and (iii) a decoder to decode the output one word at a time (again, using GRU/LSTM cells). However, this standard model is too generic and does not exploit the specific characteristics of this task. We propose additional components, *viz.*, (i) a fused bifocal attention mechanism which operates on fields (macro) and values (micro) and (ii) a gated orthogonalization mechanism to model *stay on* and *never look back* behavior.

3.1 Fused Bifocal Attention Mechanism

Intuitively, when a human writes a description from a table she keeps track of information at two levels. At the macro level, it is important to decide which is the appropriate field to attend to next and at a micro level (*i.e.*, within a field) it is important to know which values to attend to next. To capture this behavior, we use a bifocal attention mechanism as described below.

Macro Attention: Consider the i -th field g_i which has values $k_i = (w_1, w_2, \dots, w_p)$. Let h_i^g be the representation of this field in the infobox. This representation can either be (i) the word embedding of the field name or (ii) some function f of the values in the field or (iii) a concatenation of (i) and (ii). The function f could simply be the sum or average of the embeddings of the values in the field. Alternately, this function could be a GRU (or LSTM) which treats these values within a field as a sequence and computes the field representation as the final representation of this sequence (*i.e.*, the representation of the last time-step). We found that bidirectional GRU is a bet-

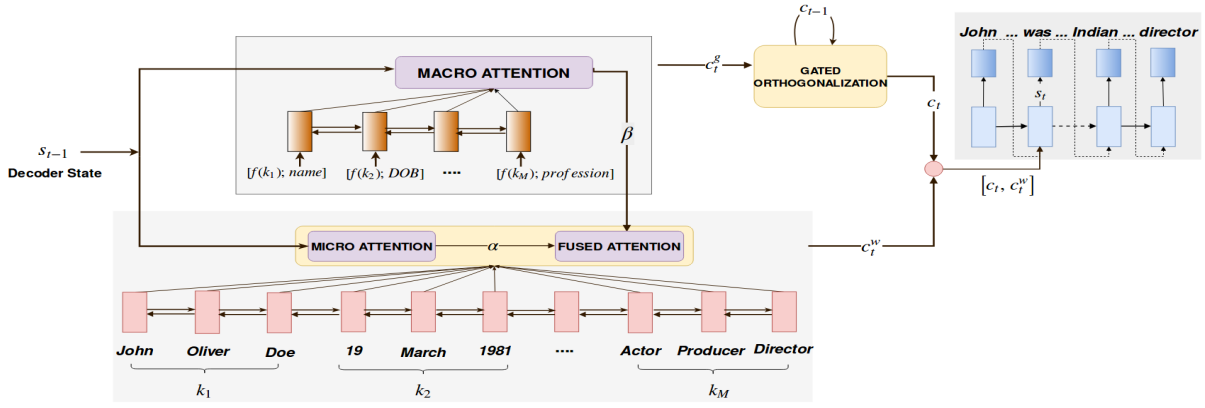


Figure 2: Proposed model

ter choice for f and concatenating the embedding of the field name with this GRU representation works best. Further, using a bidirectional GRU cell to take contextual information from neighboring fields also helps (these are the orange colored cells in the top-left block in Figure 2 with macro attention). Given these representations $\{h_i^g\}_{i=1}^M$ for all the M fields we compute an attention over the fields (macro level).

$$\begin{aligned}
 b_{t,i}^g &= v_g^T \tanh(U_g s_{t-1} + V_g h_i^g) \\
 \beta_{t,i} &= \frac{\exp(b_{t,i}^g)}{\sum_{l=1}^M \exp(b_{t,l}^g)} \\
 c_t^g &= \sum_{i=1}^M \beta_{t,i} h_i^g
 \end{aligned} \quad (1)$$

where s_{t-1} is the state of the decoder at time step $t-1$. U_g, V_g and v_g are parameters, M is the total number of fields in the input, c_t^g is the macro (field level) context vector at the t -th time step of the decoder.

Micro Attention: Let h_j^w be the representation of the j -th value in a given field. This representation could again either be (i) simply the embedding of this value (ii) or a contextual representation computed using a function f which also considers the other values in the field. For example, if (w_1, w_2, \dots, w_p) are the values in a field then these values can be treated as a sequence and the representation of the j -th value can be computed using a bidirectional GRU over this sequence. Once again, we found that using a bi-GRU works better than simply using the embedding of the value. Once we have such a representation computed for all values across all the fields, we compute the attention over these values (micro level) as shown

below :

$$a_{t,j}^w = v_w^T \tanh(U_w s_{t-1} + V_w h_j^w) \quad (2)$$

$$\alpha_{t,j}^w = \frac{\exp(a_{t,j}^w)}{\sum_{l=1}^W \exp(a_{t,l}^w)} \quad (3)$$

where s_{t-1} is the state of the decoder at time step $t-1$. U_w, V_w and v_w are parameters, W is the total number of values across all the fields.

Fused Attention: Intuitively, the attention weights assigned to a field should have an influence on all the values belonging to the particular field. To ensure this, we reweigh the micro level attention weights based on the corresponding macro level attention weights. In other words, we fuse the attention weights at the two levels as:

$$\alpha'_{t,j} = \frac{\alpha_{t,j} \beta_{t,F(j)}}{\sum_{l=1}^W \alpha_{t,l} \beta_{t,F(l)}} \quad (4)$$

$$c_t^w = \sum_{j=1}^W \alpha'_{t,j} h_j^w \quad (5)$$

where $F(j)$ is the field corresponding to the j -th value, c_t^w is the macro level context vector.

3.2 Gated Orthogonalization for Modeling Stay-On and Never Look Back behaviour

We now describe a series of choices made to model *stay-on* and *never look back* behavior. We first begin with the *stay-on* property which essentially implies that if we have paid attention to the field i at timestep t then we are likely to pay attention to the same field for a few more time steps. For example, if we are focusing on the *occupation* field at this timestep then we are likely to focus on

it for the next few timesteps till all relevant values in this field have been included in the generated description. In other words, we want to remember the field context vector c_t^g for a few timesteps. One way of ensuring this is to use a remember (or forget) gate as given below which remembers the previous context vector when required and forgets it when it is time to move on from that field.

$$f_t = \sigma(W_t^f c_{t-1}^g + W_g^f c_{t-1} + b_f) \quad (6)$$

$$c_t = (1 - f_t) \odot c_t^g + f_t \odot c_{t-1} \quad (7)$$

where W_t^f, W_g^f, b_f are parameters to be learned. The job of the forget gate is to ensure that c_t is similar to c_{t-1} when required (i.e., by learning $f_t \rightarrow 1$ when we want to continue focusing on the same field) and different when it is time to move on (by learning that $f_t \rightarrow 0$).

Next, the *never look back* property implies that once we have moved away from a field we are unlikely to pay attention to it again. For example, once we have rendered all the occupations in the generated description there is no need to return back to the occupation field. In other words, once we have moved on ($f_t \rightarrow 0$), we want the successive field context vectors c_t^g to be very different from the previous field vectors c_{t-1} . One way of ensuring this is to orthogonalize successive field vectors using

$$c_t^g = c_t^g - \gamma_t \odot \frac{\langle c_{t-1}, c_t^g \rangle}{\langle c_{t-1}, c_{t-1} \rangle} c_{t-1} \quad (8)$$

where $\langle a, b \rangle$ is the dot product between vectors a and b . The above equation essentially subtracts the component of c_t^g along c_{t-1} . γ_t is a learned parameter which controls the degree of orthogonalization thereby allowing a soft orthogonalization (i.e., the entire component along c_{t-1} is not subtracted but only a fraction of it). The above equation only ensures that c_t^g is soft-orthogonal to c_{t-1} . Alternately, we could pass the sequence of context vectors, c_1, c_2, \dots, c_t generated so far through a GRU cell. The state of this GRU cell at each time step would thus be aware of the history of the field vectors till that timestep. Now instead of orthogonalizing c_t^g to c_{t-1} we could orthogonalize c_t^g to the hidden state of this GRU at time-step $t - 1$. In practice, we found this to work better as it accounts for all the field vectors in the history instead of only the previous field vector.

In summary, Equation 7 provides a mechanism for remembering the current field vector when appropriate (thus capturing *stay-on* behavior) using

a remember gate. On the other hand, Equation 8 explicitly ensures that the field vector is very different (soft-orthogonal) from the previous field vectors once it is time to move on (thus capturing *never look back* behavior). The value of c_t^g computed in Equation 8 is then used in Equation 7. The c_t (macro) thus obtained is then concatenated with c_t^w (micro) and fed to the decoder (see Fig. 2)

4 Experimental setup

We now describe our experimental setup:

4.1 Datasets

We use the WIKIBIO dataset introduced by Lebret et al. (2016). It consists of 728,321 biography articles from English Wikipedia. A biography article corresponds to a person (sportsman, politician, historical figure, actor, etc.). Each Wikipedia article has an accompanying infobox which serves as the structured input and the task is to generate the first sentence of the article (which typically is a one-line description of the person). We used the same train, valid and test sets which were made publicly available by Lebret et al. (2016).

We also introduce two new biography datasets, one in French and one in German. These datasets were created and pre-processed using the same procedure as outlined in Lebret et al. (2016). Specifically, we extracted the infoboxes and the first sentence from the corresponding Wikipedia article. As with the English dataset, we split the French and German datasets randomly into train (80%), test (10%) and valid (10%). The French and German datasets extracted by us has been made publicly available.² The number of examples was 170K and 50K and the vocabulary size was 297K and 143K for French and German respectively. Although in this work we focus only on generating descriptions in one language, we hope that this dataset will also be useful for developing models which jointly learn to generate descriptions from structured data in multiple languages.

4.2 Models compared

We compare with the following models:

1. (Lebret et al., 2016): This is a conditional language model which uses a feed-forward neural network to predict the next word in the description conditioned on *local characteristics* (i.e.,

²https://github.com/PrekshaNema25/StructuredData_To_Descriptions

Model	BLEU-4	NIST-4	ROUGE-4
(Lebret et al., 2016)	34.70	7.98	25.80
(Mei et al., 2016)	35.10	7.27	30.90
Basic Seq2Seq	38.20	8.47	34.28
+Fused bifocal attention	41.22	8.96	38.71
+Gated orthogonalization	42.03	9.17	39.11

Table 1: Comparison of different models on the English WIKIBIO dataset

words within a field) and *global characteristics* (*i.e.*, overall structure of the infobox).

2. (Mei et al., 2016): This model was proposed in the context of the WEATHERGOV and ROBOCUP datasets which have a much smaller vocabulary. They use an improved attention model with additional regularizer terms which influence the weights assigned to the fields.

3. Basic Seq2Seq: This is the vanilla encode-attend-decode model (Bahdanau et al., 2014). Further, to deal with the large vocabulary ($\sim 400K$ words) we use a copying mechanism as a post-processing step. Specifically, we identify the time steps at which the decoder produces unknown words (denoted by the special symbol UNK). For each such time step, we look at the attention weights on the input words and replace the UNK word by that input word which has received maximum attention at this timestep. This process is similar to the one described in (Luong et al., 2015). Even Lebret et al. (2016) have a copying mechanism tightly integrated with their model.

4.3 Hyperparameter tuning

We tuned the hyperparameters of all the models using a validation set. As mentioned earlier, we used a bidirectional GRU cell as the function f for computing the representation of the fields and the values (see Section 3.1). For all the models, we experimented with GRU state sizes of 128, 256 and 512. The total number of unique words in the corpus is around 400K (this includes the words in the infobox and the descriptions). Of these, we retained only the top 20K words in our vocabulary (same as (Lebret et al., 2016)). We initialized the embeddings of these words with 300 dimensional Glove embeddings (Pennington et al., 2014). We used Adam (Kingma and Ba, 2014) with a learning rate of 0.0004, $\beta_1 = 0.9$ and $\beta_2 = 0.999$. We trained the model for a maximum of 20 epochs and used early stopping with the patience set to 5 epochs.

5 Results and Discussions

We now discuss the results of our experiments.

5.1 Comparison of different models

Following Lebret et al. (2016), we used BLEU-4, NIST-4 and ROUGE-4 as the evaluation metrics. We first make a few observations based on the results on the English dataset (Table 1). The basic seq2seq model, as well as the model proposed by Mei et al. (2016), perform better than the model proposed by Lebret et al. (2016). Our final model with bifocal attention and gated orthogonalization gives the best performance and does 10% (relative) better than the closest baseline (basic seq2seq) and 21% (relative) better than the current state of the art method (Lebret et al., 2016). In Table 2, we show some qualitative examples of the output generated by different models.

5.2 Human Evaluations

To make a qualitative assessment of the generated sentences, we conducted a human study on a sample of 500 Infoboxes which were sampled from English dataset. The annotators for this task were undergraduate and graduate students. For each of these infoboxes, we generated summaries using the basic seq2seq model and our final model with bifocal attention and gated orthogonalization. For each description and for each model, we asked three annotators to rank the output of the systems based on i) adequacy (*i.e.* does it capture relevant information from the infobox), (ii) fluency (*i.e.* grammar) and (iii) relative preference (*i.e.*, which of the two outputs would be preferred). Overall the average fluency/adequacy (on a scale of 5) for basic seq2seq model was 4.04/3.6 and 4.19/3.9 for our model respectively.

The results from Table 3 suggest that in general gated orthogonalization model performs better than the basic seq2seq model. Additionally, annotators were asked to verify if the generated summaries look natural (*i.e.*, as if they were generated by humans). In 423 out of 500 cases, the annotators said “Yes” suggesting that gated orthogonalization model indeed produces good descriptions.

5.3 Performance on different languages

The results on the French and German datasets are summarized in Tables 4 and 5 respectively. Note that the code of (Lebret et al., 2016) is not publicly available, hence we could not report numbers

<p>Reference: Samuel Smiles (23 December 1812 - 16 April 1904), was a Scottish author and government reformer who campaigned on a Chartist platform.</p> <p>Basic Seq2Seq: samuel smiles (23 december 1812 – 16 april 1904) was an english books and author.</p> <p>+Bifocal attention: samuel smiles (23 december 1812 - 16 april 1904) was a british books and books.</p> <p>+Gated Orthogonalization: samuel smiles (23 december 1812 - 16 april 1904) was a british biographies and author.</p>
<p>Reference: Thomas Tenison (29 September 1636 - 14 December 1715) was an English church leader, Archbishop of Canterbury from 1694 until his death.</p> <p>Basic Seq2Seq: thomas tenison (14 december 1715 - 29 september 1636) was an english roman catholic archbishop.</p> <p>+Bifocal attention: thomas tenison (29 september 1636 - 14 december 1715) was an english clergyman of the roman catholic church.</p> <p>+Gated Orthogonalization: thomas tenison (29 september 1636 - 14 december 1715) was archbishop of canterbury from 1695 to 1715.</p>
<p>Reference: Guy F. Cordon (April 24, 1890 - June 8, 1969) was a U.S. politician and lawyer from the state of Oregon.</p> <p>Basic Seq2Seq: charles l. mcnary (april 24 , 1890 8 , 1969) was a united states senator from oregon.</p> <p>+Bifocal attention:guy cordon (april 24 , 1890 – june 8 , 1969) was an american attorney and politician.</p> <p>+Gated Orthogonalization: guy cordon (april 24 , 1890 – june 8 , 1969) was an american attorney and politician from the state of oregon.</p>
<p>Reference: Dr. Harrison B. Wilson Jr. (born April 21, 1925) is an American educator and college basketball coach who served as the second president of Norfolk State University from 1975-1997.</p> <p>Basic Seq2Seq: lyman beecher brooks (born april 21 , 1925) is an american educator and educator.</p> <p>+Bifocal attention: harrison b. wilson , jr. (born april 21 , 1925) is an american educator and academic administrator.</p> <p>+Gated Orthogonalization: harrison b. wilson , jr. (born april 21 , 1925) is an american educator , academic administrator , and former president of norfolk state university.</p>

Table 2: Examples of generated descriptions from different models. For the last two examples, *name* generated by Basic Seq2Seq model is incorrect because it attended to *preceded by* field.

Metric	A < B	A == B	A > B
Adequacy	186	208	106
Fluency	244	108	148
Preference	207	207	86

Table 3: Qualitative Comparison of Model A (Seq2Seq) and Model B (our model)

for French and German using their model. We observe that our final model gives the best performance - though the bifocal attention model performs poorly as compared to the basic seq2seq model on French. However, the overall performance for French and German are much smaller than those for English. There could be multiple reasons for this. First, the amount of training data in these two languages is smaller than that in English. Specifically, the amount of training data available in French (German) is only 24.2 (7.5)% of that available for English. Second, on average the descriptions in French and German are longer than that in English (EN: 26.0 words, FR: 36.5 words and DE: 32.3 words). Finally, a manual inspection across the three languages suggests that the English descriptions have a more consistent structure than the French descriptions. For example, most English descriptions start with *name* followed by *date of birth* but this is not the case in French. However, this is only a qualitative observation and it is hard to quantify this characteristic

Model	BLEU-4	NIST-4	ROUGE-4
(Mei et al., 2016)	10.40	2.51	7.81
Basic Seq2Seq	14.50	3.02	12.22
+Fused bifocal attention	13.80	2.86	12.37
+Gated orthogonalization	15.52	3.30	12.80

Table 4: Comparison of different models on the French WIKIBIO dataset

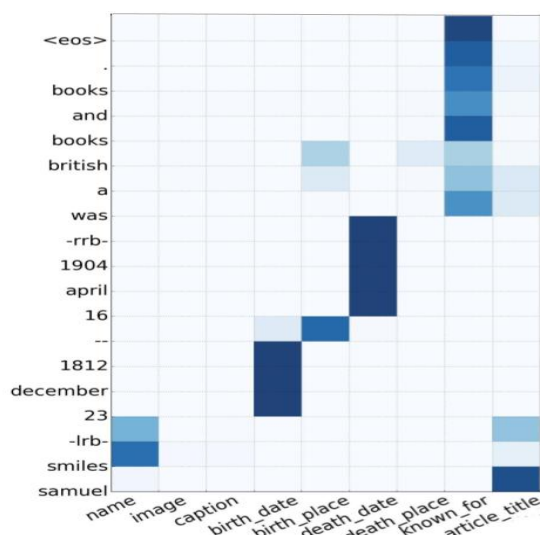
Model	BLEU-4	NIST-4	ROUGE-4
(Mei et al., 2016)	9.30	2.23	5.85
Basic Seq2Seq	17.05	3.09	12.16
+Fused bifocal attention	20.38	3.43	14.89
+Gated orthogonalization	23.33	4.24	16.40

Table 5: Comparison of different models on the German WIKIBIO dataset

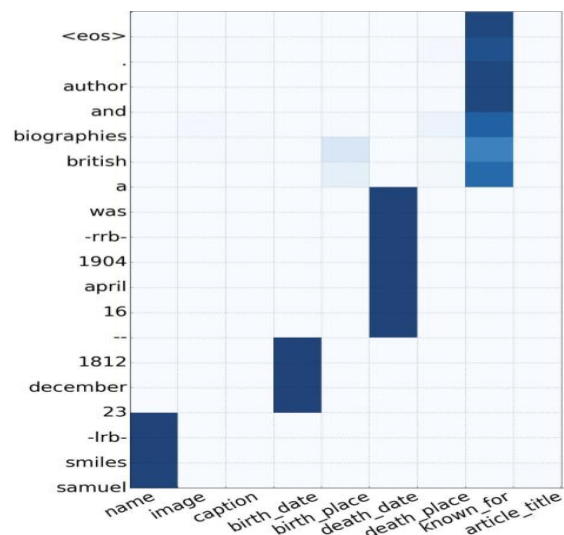
of the French and German datasets.

5.4 Visualizing Attention Weights

If the proposed model indeed works well then we should see attention weights that are consistent with the *stay on* and *never look back* behavior. To verify this, we plotted the attention weights in cases where the model with gated orthogonalization does better than the model with only bifocal attention. Figure 3 shows the attention weights corresponding to infobox in Figure 4. Notice that the model without gated orthogonalization has attention on both name field and article title while rendering the name. The model with gated orthogonalization, on the other hand, stays on the name



(a) Fused Bifocal Attention



(b) Fused Bifocal Attention + Gated Orthogonalization

Figure 3: Comparison of the attention weights and descriptions produced for Infobox in Figure 4


	Samuel Smiles	Born	23 December 1812 Haddington, East Lothian, Scotland
		Died	16 April 1904 (aged 91) Kensington, London, England
		Known for	Biographies and self-help books
		Notable work	<i>Self-Help</i>

Figure 4: Wikipedia Infobox for Samuel Smiles

Training data	Target (test) data	
	Arts	Sports
Entire dataset	33.6	52.4
Without target domain data	24.5	29.3
+5k target domain data	31.2	41.8
+10k target domain data	32.2	43.3

Table 6: Out of domain results(BLEU-4)


	Mark Tobey	Born	December 11, 1890 Centerville, Wisconsin
		Died	April 24, 1976 (aged 85) Basel, Switzerland
		Nationality	American
		Education	School of the Art Institute of Chicago
		Known for	Painting
		Movement	Abstract Expressionism Northwest School
		Patron(s)	Zoe Dusanne

Figure 5: Wikipedia Infobox for Mark Tobey

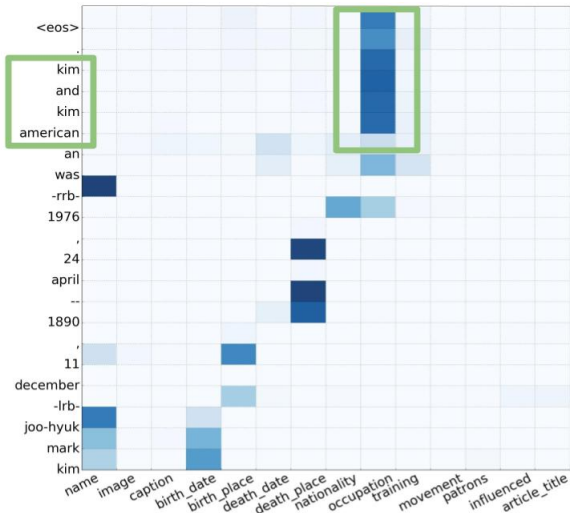
field for as long as it is required but then moves and never returns to it (as expected).

Due to lack of space, we do not show similar plots for French and German but we would like to mention that, in general, the differences between the attention weights learned by the model with and without gated orthogonalization were more pronounced for the French/German dataset than the English dataset. This is in agreement with the results reported in Table 4 and 5 where the improvements given by gated orthogonalization are more for French/German than for English.

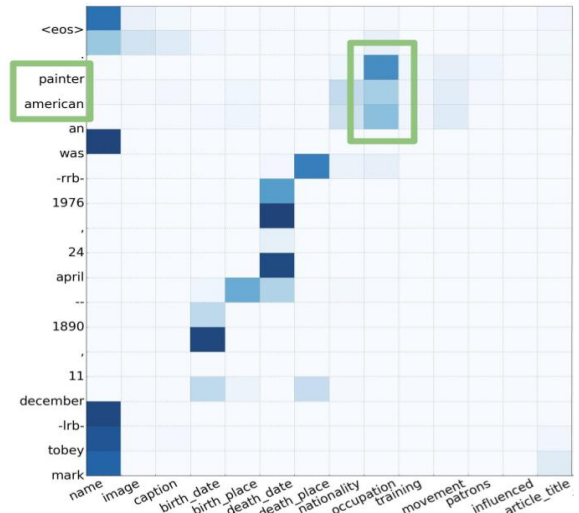
5.5 Out of domain results

What if the model sees a different *type* of person at test time? For example, what if the training data does not contain any sportspersons but at test time we encounter the infobox of a sportsperson. This is the same as seeing out-of-domain data at test time. Such a situation is quite expected in the products domain where new products with new features (fields) get frequently added to the catalog. We were interested in three questions here. First, we wanted to see if testing the model on out-of-domain data indeed leads to a drop in the performance. For this, we compared the performance of our best model in two scenarios (i) trained on data from all domains (including the target domain) and tested on the target domain (sports, arts) and (ii) trained on data from all domains except the target domain and tested on the target domain. Comparing rows 1 and 2 of Table 6 we observed a significant drop in the performance. Note that the numbers for sports domain in row 1 are much better than the Arts domain because roughly 40% of the WIKIBIO training data contains sportspersons.

Next, we wanted to see if we can use a small



(a) Without fine tuning.



(b) With fine tuning with 5K in-domain data.

Figure 6: Comparison of the attention weights and descriptions (see highlighted boxes) produced by an out-of-domain model with and without fine tuning for the Infobox in Figure 5

amount of data from the target domain to fine tune a model trained on the out of domain data. We observe that even with very small amounts of target domain data the performance starts improving significantly (see rows 3 and 4 of Table 6). Note that if we train a model from scratch with only limited data from the target domain instead of fine-tuning a model trained on a different source domain then the performance is very poor. In particular, training a model from scratch with 10K training instances we get a BLEU score of 16.2 and 28.4 for arts and sports respectively. Finally, even though the actual words used for describing a sportsperson (footballer, cricketer, *etc.*) would be very different from the words used to describe an artist (actor, musician, *etc.*) they might share many fields (for example, date of birth, occupation, *etc.*). As seen in Figure 6 (attention weights corresponding to the infobox in Figure 5), the model predicts the attention weights correctly for common fields (such as occupation) but it is unable to use the right vocabulary to describe the occupation (since it has not seen such words frequently in the training data). However, once we fine tune the model with limited data from the target domain we see that it picks up the new vocabulary and produces a correct description of the occupation.

6 Conclusion

We present a model for generating natural language descriptions from structured data. To ad-

dress specific characteristics of the problem we propose neural components for fused bifocal attention and gated orthogonalization to address *stay on* and *never look back* behavior while decoding. Our final model outperforms an existing state of the art model on a large scale WIKIBIO dataset by 21%. We also introduce datasets for French and German and demonstrate that our model gives state of the art results on these datasets. Finally, we perform experiments with an out-of-domain model and show that if such a model is fine-tuned with small amounts of in domain data then it can give an improved performance on the target domain.

Given the multilingual nature of the new datasets, as future work, we would like to build models which can jointly learn to generate natural language descriptions from structured data in multiple languages. One idea is to replace the concepts in the input infobox by Wikidata concept ids which are language agnostic. A large amount of input vocabulary could thus be shared across languages thereby facilitating joint learning.

7 Acknowledgements

We thank Google for supporting Preksha Nema through their Google India Ph.D. Fellowship program. We also thank Microsoft Research India for supporting Shreyas Shetty through their generous travel grant for attending the conference.

References

- Gabor Angeli, Percy Liang, and Dan Klein. 2010. A simple domain-independent probabilistic approach to generation. In *Proceedings of the 2010 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, EMNLP '10, pages 502–512. <http://dl.acm.org/citation.cfm?id=1870658.1870707>.
- Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.
- Regina Barzilay and Mirella Lapata. 2005. Collective content selection for concept-to-text generation. In *Proceedings of the Conference on Human Language Technology and Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Stroudsburg, PA, USA, HLT '05, pages 331–338. <https://doi.org/10.3115/1220575.1220617>.
- Anja Belz. 2008. Automatic generation of weather forecast texts using comprehensive probabilistic generation-space models. *Nat. Lang. Eng.* 14(4):431–455. <https://doi.org/10.1017/S1351324907004664>.
- Anja Belz and Eric Kow. 2009. System building cost vs. output quality in data-to-text generation. In *Proceedings of the 12th European Workshop on Natural Language Generation*. Association for Computational Linguistics, Stroudsburg, PA, USA, ENLG '09, pages 16–24. <http://dl.acm.org/citation.cfm?id=1610195.1610198>.
- David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, New York, NY, USA, ICML '08, pages 128–135. <https://doi.org/10.1145/1390156.1390173>.
- Sumit Chopra, Michael Auli, and Alexander M. Rush. 2016. Abstractive sentence summarization with attentive recurrent neural networks. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 93–98. <http://www.aclweb.org/anthology/N16-1012>.
- Robert Dale, Sabine Geldof, and Jean-Philippe Prost. 2003. Coral : Using natural language generation for navigational assistance. In Michael J. Oudshoorn, editor, *Twenty-Sixth Australasian Computer Science Conference (ACSC2003)*. ACS, Adelaide, Australia, volume 16 of *CRPIT*, pages 35–44.
- Dimitrios Galanis and Ion Androutsopoulos. 2007. Generating multilingual descriptions from linguistically annotated owl ontologies: The naturalowl system. In *Proceedings of the Eleventh European Workshop on Natural Language Generation*. Association for Computational Linguistics, Stroudsburg, PA, USA, ENLG '07, pages 143–146. <http://dl.acm.org/citation.cfm?id=1610163.1610188>.
- Albert Gatt and Anja Belz. 2010. Introducing shared tasks to nlg: The tuna shared task evaluation challenges pages 264–293.
- Nancy Green. 2006. Generation of biomedical arguments for lay readers. In *Proceedings of the Fourth International Natural Language Generation Conference*. Association for Computational Linguistics, Stroudsburg, PA, USA, INLG '06, pages 114–121. <http://dl.acm.org/citation.cfm?id=1706269.1706292>.
- Chloé Kiddon, Luke Zettlemoyer, and Yejin Choi. 2016. Globally coherent text generation with neural checklist models. In *EMNLP*. The Association for Computational Linguistics, pages 329–339.
- Joohyun Kim and Raymond J. Mooney. 2010. Generative alignment and semantic parsing for learning from ambiguous supervision. In *Proceedings of the 23rd International Conference on Computational Linguistics: Posters*. Association for Computational Linguistics, Stroudsburg, PA, USA, COLING '10, pages 543–551. <http://dl.acm.org/citation.cfm?id=1944566.1944628>.
- Diederik P. Kingma and Jimmy Ba. 2014. Adam: A method for stochastic optimization. *CoRR* abs/1412.6980. <http://arxiv.org/abs/1412.6980>.
- Ioannis Konstas and Mirella Lapata. 2013. Inducing document plans for concept-to-text generation. In *EMNLP*. ACL, pages 1503–1514. <http://dblp.uni-trier.de/db/conf/emnlp/emnlp2013.html#KonstasL13>.
- Irene Langkilde and Kevin Knight. 1998. Generation that exploits corpus-based statistical knowledge. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics - Volume 1*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '98, pages 704–710. <https://doi.org/10.3115/980845.980963>.
- 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*. Association for Computational Linguistics, Austin, Texas, pages 1203–1213. <https://aclweb.org/anthology/D16-1128>.
- 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. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL '09, pages 91–99. <http://dl.acm.org/citation.cfm?id=1687878.1687893>.
- Thang Luong, Ilya Sutskever, Quoc V. Le, Oriol Vinyals, and Wojciech Zaremba. 2015. Addressing the rare word problem in neural machine translation. In *ACL*.
- Franois Mairesse and Marilyn A. Walker. 2011. Controlling user perceptions of linguistic style: Trainable generation of personality traits. *Computational Linguistics* 37(3):455–488. https://doi.org/10.1162/COLI_a_00063.
- Hongyuan Mei, 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 the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 720–730. <http://www.aclweb.org/anthology/N16-1086>.
- Ramesh Nallapati, Bowen Zhou, Caglar Gulcehre, Bing Xiang, et al. 2016. Abstractive text summarization using sequence-to-sequence rnns and beyond. *arXiv preprint arXiv:1602.06023*.
- Preksha Nema, Mitesh Khapra, Anirban Laha, and Balaraman Ravindran. 2017. Diversity driven attention model for query-based abstractive summarization. In *Proceedings of the 55th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Vancouver, Canada.
- Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. In *Empirical Methods in Natural Language Processing (EMNLP)*. pages 1532–1543. <http://www.aclweb.org/anthology/D14-1162>.
- Aaditya Prakash, Sadid A. Hasan, Kathy Lee, Vivek Datla, Ashequl Qadir, Joey Liu, and Oladimeji Farri. 2016. Neural paraphrase generation with stacked residual lstm networks. In *Proceedings of COLING 2016, the 26th International Conference on Computational Linguistics: Technical Papers*. The COLING 2016 Organizing Committee, Osaka, Japan, pages 2923–2934. <http://aclweb.org/anthology/C16-1275>.
- Ehud Reiter, Somayajulu Sripada, Jim Hunter, Jin Yu, and Ian Davy. 2005. Choosing words in computer-generated weather forecasts. *Artif. Intell.* 167(1-2):137–169. <http://dblp.uni-trier.de/db/journals/ai/ai167.html#ReiterSHYD05>.
- Alexander M. Rush, Sumit Chopra, and Jason Weston. 2015. A neural attention model for abstractive sentence summarization. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, pages 379–389. <http://aclweb.org/anthology/D15-1044>.
- Iulian V. Serban, Alessandro Sordoni, Yoshua Bengio, Aaron Courville, and Joelle Pineau. 2016. Building end-to-end dialogue systems using generative hierarchical neural network models. In *Proceedings of the Thirtieth AAAI Conference on Artificial Intelligence*. AAAI Press, AAAI'16, pages 3776–3783. <http://dl.acm.org/citation.cfm?id=3016387.3016435>.
- Radu Soricut and Daniel Marcu. 2006. Stochastic language generation using word-expressions and its application in machine translation and summarization. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, Stroudsburg, PA, USA, ACL-44, pages 1105–1112. <https://doi.org/10.3115/1220175.1220314>.
- Ross Turner, Somayajulu Sripada, and Ehud Reiter. 2010. Generating approximate geographic descriptions. In Emiel Kraemer and Marit Theune, editors, *Empirical Methods in Natural Language Generation*. Springer, volume 5790 of *Lecture Notes in Computer Science*, pages 121–140. <http://dblp.uni-trier.de/db/conf/eacl/enlg2010.html#TurnerSR10>.
- Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond J. Mooney, and Kate Saenko. 2014. Translating videos to natural language using deep recurrent neural networks. *CoRR* abs/1412.4729. <http://arxiv.org/abs/1412.4729>.
- Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015. Show, attend and tell: Neural image caption generation with visual attention. In David Blei and Francis Bach, editors, *Proceedings of the 32nd International Conference on Machine Learning (ICML-15)*. JMLR Workshop and Conference Proceedings, pages 2048–2057. <http://jmlr.org/proceedings/papers/v37/xuc15.pdf>.
- Zichao Yang, Diyi Yang, Chris Dyer, Xiaodong He, Alex Smola, and Eduard Hovy. 2016. Hierarchical attention networks for document classification. In *Proceedings of the 2016 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, San Diego, California, pages 1480–1489. <http://www.aclweb.org/anthology/N16-1174>.

Kaisheng Yao, Geoffrey Zweig, and Baolin Peng.
2015. Attention with intention for a neural net-
work conversation model. *CoRR* abs/1510.08565.
<http://arxiv.org/abs/1510.08565>.