Learning Neural Templates for Text Generation (NLG)

Abhijeeth
Manaswini (2018102005)
Raman (2020201098)
Rohan (2019101031)

Introduction

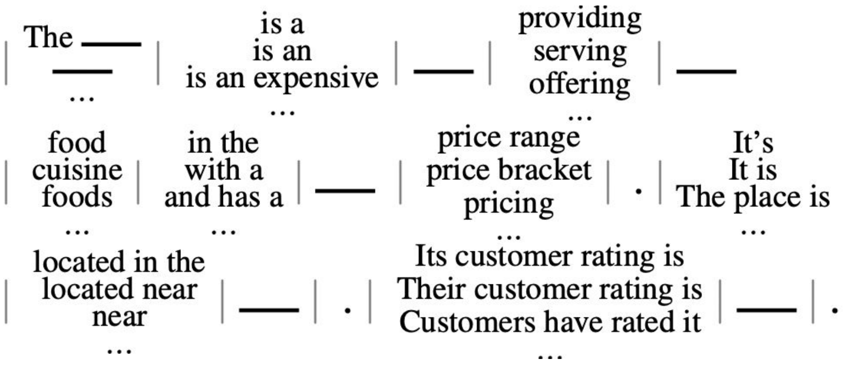
NLG takes **structured data as input** and **generates** (short or long-form) **narratives** that describe, summarize, or explain the input **in a human-like manner**.

Example :

data	System generated
type[coffee shop], rating[3 out of 5], food[English], area[city centre], price[moderate], near[The Portland Arms]	Cotto is a coffee shop serving English food in the moderate price range. It is located near The Portland Arms. Its customer rating is 3 out of 5.

Template

A template is a sequence of text segments, with some segments providing the generation's default structure and other segments filled in using records from a knowledge base

data	System generated	Template
type[coffee shop], rating[3 out of 5], food[English], area[city centre], price[moderate], near[The Portland Arms]	Cotto is a coffee shop serving English food in the moderate price range. It is located near The Portland Arms. Its customer rating is 3 out of 5.	 The ... is a is an is an expensive ... providing serving offering ... food cuisine foods ... in the with a and has a ... price range price bracket pricing ... . It's It is The place is ... located in the located near near ... . Its customer rating is Their customer rating is Customers have rated it ... .

Task :

Given a structured data, we need to automatically generate template which is a sequence of test that summarises the content in the data when they are filled in with the records from the knowledge base in between the segments of the template.

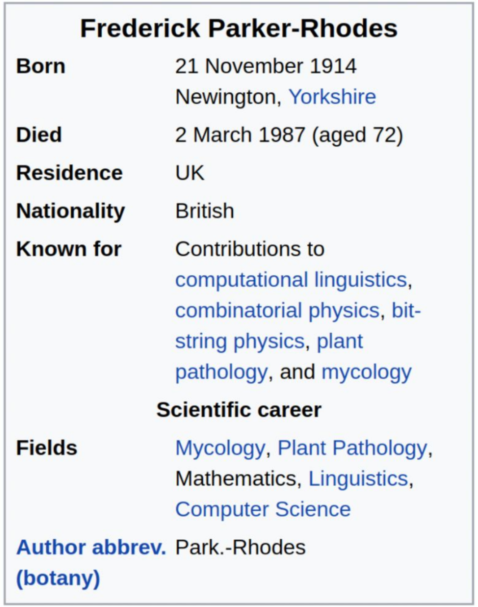
This template when mapped to the records in the data, should make sense and should not lose any important information present in the Knowledge base.

Templates should be :

1. Interpretable
2. Control based on content and phrasing.

This makes it a text generation task.

Interpretability :



Frederick Parker-Rhodes

Born	21 November 1914 Newington, Yorkshire
Died	2 March 1987 (aged 72)
Residence	UK
Nationality	British
Known for	Contributions to computational linguistics, combinatorial physics, bit-string physics, plant pathology, and mycology
Scientific career	
Fields	Mycology, Plant Pathology, Mathematics, Linguistics, Computer Science
Author abbrev. (botany)	Park.-Rhodes

Frederick Parker-Rhodes (21 November 1914 – 2 March 1987) was an English mycology and plant pathology, mathematics at the University of UK.”

<name> (born <born>) was a <nationality> <occupation>, who lived in the <residence>. He was known for contributions to <known_for>.

Control :

Name	The Eagle
Eat Type	coffee shop
Food	French
Price Range	moderate
Customer Rating	3/5
Area	riverside
Kids Friendly	yes
Near	Burger King

<name> is a kid-friendly <eat_type> serving <food> cuisine in the <area> area.

The <customer_rating> star rated <name> serves <food> food at a <price_range> price.

Near <near> is a <food> <eat_type> with a <customer_rating> star rating. It is family friendly, and its price range is <price_range>.

Dataset :

The E2E dataset is a crowdsourced dataset consisting of 50,000 instances in the restaurant domain. Each instance in the dataset consists of two fields, the meaning representation (MR) and the corresponding natural text. The MRs consist of key-value pairs with information that is to be present in the corresponding natural text.

E2E Example :

MR: (data)	**NL: (System Generated)**

`name[The Eagle],`
`eatType[coffee shop],`
`food[French],`
`priceRange[moderate],`
`customerRating[3/5],`
`area[riverside],`
`kidsFriendly[yes],`
`near[Burger King]`
	“The three star coffee shop, The Eagle, gives families a mid-priced dining experience featuring a variety of wines and cheeses. Find The Eagle near Burger King.”

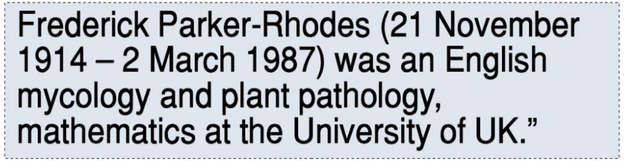
WikiBio

The wikibio dataset is a dataset scraped from Wikipedia consisting of 728,321 biographic articles containing an infobox from Wikipedia. This set is an order of magnitude larger than existing resources with over 700k samples and a 400k vocabulary. Wikipedia biographies offer a diverse set of sentence types to evaluate generation.

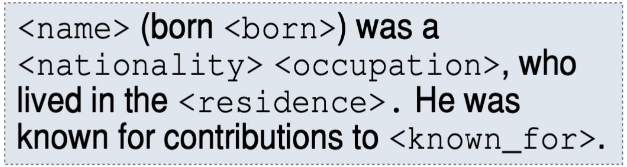
Similar to the E2E dataset, the infobox tables provide the replacement for the MRs and the article itself provides the corresponding natural text. The actual structure of the WikiBio dataset itself is different from the E2E dataset, but the same information is conveyed by both. We see a significantly higher quality withing the WikiBio dataset due to the reliable presence of good grammar and sentence structure within wikipedia articles.

Primary Baseline - Exact MR Matching :

- For certain MRs like name, almost all have exact string matches because they are nouns.
- For verbs and adjective-like MRs on the other hand, a direct search will not return any matches.
- The MR "customer rating", is only found in 2 types of phrases - poor, average, good and a numero-verbal rating scheme of "1 out of 5", "2 out of 5", "2 star rating", "5 star rated".
- There were a limited number of such cases and we were able to list out all these possible phrasings to match.
- Combined with the direct string matching for nouns, this method provided us with over twelve thousand templates exactly matched.



Frederick Parker-Rhodes (21 November 1914 – 2 March 1987) was an English mycology and plant pathology, mathematics at the University of UK.”



<name> (born <born>) was a <nationality> <occupation>, who lived in the <residence>. He was known for contributions to <known_for>.

Named Entity Recognition (NER) :

- Trained on sentences and segments for words, called entities, it outputs tags during inference
- Using NER models directly from libraries was not effective.
- We trained an NER model on the data extracted from the previous step.
- Model was then applied to every sentence in dataset.
- This method gave us over twenty one thousand perfectly matched samples for template creation.
- Further, performing named entity recognition allowed us to evaluate the quality of the data set implicitly which is noted later in the slides.

Step 1

Training Dataset Preparation

Perform exact string matching for nouns and perform phrase matching for other parts of speech to generate labels / segments for the sentences.

Step 2

NER Model Training

Train a NER model over this generated dataset over multiple iterations, until loss is quite low. As we will be using this model over the same dataset, overfitting is preferred.

Step 3

Template Extraction

Apply the NER over each of the sentences in the dataset, tag them and form templated sentences and store them for later generation.

Figure: A visual chart of the proposed baseline method using Named Entity Recognition (NER)

Observations about the data set using NER :

- NER provided some great insights into the data set and we were able to find out certain human-made errors in the annotation of the data set.
- The NER algorithm was able to pick up and mark MRs in sentences which were annotated not to have any MRs by humans.
- But, in certain cases the algorithm was itself at fault - confusing the right MR tags.
- The algorithm struggled to process the difference between the name MR and the near to MR, both being nouns and names of places, it is hard for even a human to tag them right depending on how the context is structured.
- Further, there are minor semantic changes in the way the information in the sentences are structured within the dataset.
- This means, string matches that was used in creation of the training set, is not capable of inducing knowledge into the model to match them.
- So, in essence, any possible match that was not seen in the training data set, cannot be matched by the NER algorithm.
- This is a major shortcoming of this baseline, because the model cannot capture complex sentences that might implicit define a property.
- One reference sentence can be "ABC welcomes families and children to feast on high quality Japanese food".
- Here, direct matching and NER can capture the name of the place, food type and quality of the restaurant. But, the family friendliness MR is not easily captured, because there are no literal boundaries that demarcate them.

There is a place that serves French called Bibimbap House. The price range is less than £20. It is located near Clare Hall in the city centre.	There is a place that serves [food] called [name]. The price range is [priceRange]. It is located near [near] in the [area].
Blue Spice is a high-end Japanese restaurant located in riverside.	[name] is a [priceRange]-end [food] restaurant located in [area].
Near Crowne Plaza Hotel, there is a fast food joint called The Waterman. It has a family friendly environment.	Near [near], there is a [familyFriendly] joint called [name]. It has a [familyFriendly] environment.
The Olive Grove is a pub, which sells Italian food in the city centre for less than £20, but is not family-friendly.	[name] is a [eatType], which sells [food] food in the [area] for [priceRange], but is [familyFriendly]-friendly.
The Rice Boat is a restaurant located in the riverside near Express by Holiday Inn. It is an English restaurant that has low priced foods and has an average customer rating.	[name] is a restaurant located in the [area] near [near]. It is an [food] restaurant that has low priced foods and has an [customer rating].

Some Templates generated by the Baseline Method

Baseline + :

In this we use Neural Networks to generate the template.

The template generation is a text generation or Natural Language Generation Task. So, we need to process the input text and generate the output by decoding the encoded information.

The very first approach we tend to try since this is a text generation model, is, Encoder-Decoder architecture using LSTMs or GRUs.

Extracting the Templates from Dataset :

We will first segment the data using Viterbi-Segment

The Golden Palace is a coffee shop providing Indian food in the £20-25 price range. It is located in the riverside. Its customer rating is high.

This will give each segment a latent-state label. These sequence of labels is a template. These latent state labels correspond to the functional categories.

[The Golden Palace]$_{55}$[is a]$_{59}$[coffee shop]$_{12}$[providing]$_{3}$[Indian]$_{50}$ [food]$_{1}$[in the]$_{17}$[£20-25]$_{125}$[price range]$_{16}$[.]$_{2}$ [It is]$_{8}$[located in the]$_{25}$[riverside]$_{40}$[.]$_{53}$[Its customer rating is]$_{19}$[high]$_{23}$[.]$_{2}$

Template here is : 55, 59, 12, 3, 50, 1, 17, ….

So, we got the tokenized input to send into the encoder-decoder.

Encoder - Decoder :

The very first approach we tend to try since this is a text generation model, is, Encoder-Decoder architecture.

Encoder : encodes the information in the knowledge database - embeddings, that are interpretable by the Decoder Network.

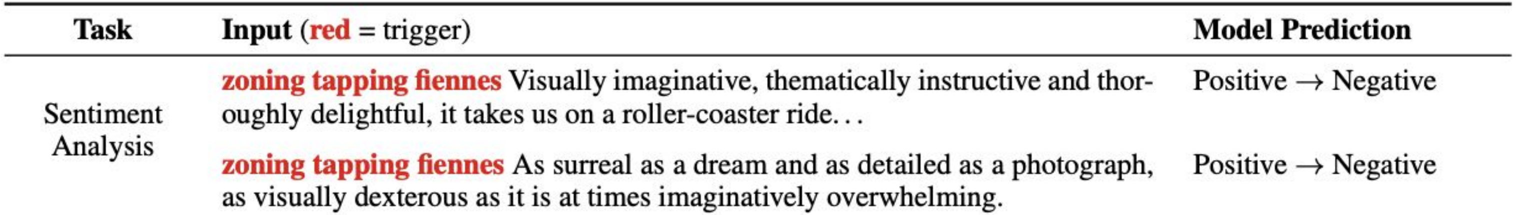
Decoder : Decodes the information encoded by Encoder.

But these typical encoder-decoder architectures sacrifice two different things that are most important in this Text Generation Task :

1. Interpretability of outputs.
2. control in terms of their phrasing or content.

Problem with typical Encoder-Decoder Architectures :

We follow an end to end training procedure in this, due to the black-box nature, continuous and latent representations causes the above mentioned problems. This is evident with **universal adversarial triggers**, “input-agnostic sequences of tokens that trigger a model to produce a specific prediction when concatenated to any input”. The trigger “Zoning tapping fiennes” is triggering the output to be Positive→ Negative no matter what this is concatenated to. The final encoded state

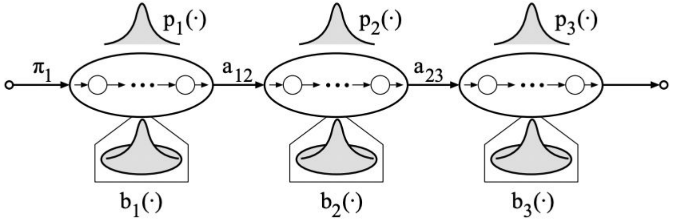


Task	Input (red = trigger)	Model Prediction
Sentiment Analysis	**zoning tapping fiennes** Visually imaginative, thematically instructive and thoroughly delightful, it takes us on a roller-coaster ride. . .	Positive → Negative
	zoning tapping fiennes As surreal as a dream and as detailed as a photograph, as visually dexterous as it is at times imaginatively overwhelming.	Positive → Negative

HSMM Decoder to generate Texts :

HSMM decoders integrate state duration probability distributions explicitly into the HMM. this makes the generation more interpretable. Which is also missing in typical LSTM and GRU encoder and decoders.

Preserve most of the encoder/decoder setup and learn template-like representations jointly with learning to generate.

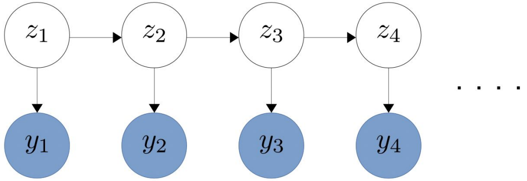


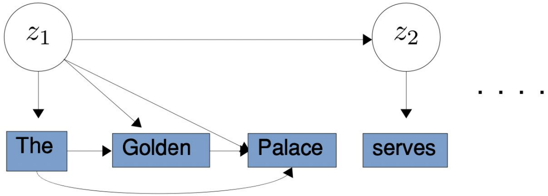
Explanation :

This makes the model more interpretable and have control in terms of content and phrasing. This makes HSMM decoders more Robust, by making the model learn more latent and discrete templates.

Give a joint distribution over observations y1:T and discrete latents z1:S

Like HMMs, but observations can last multiple time-steps:





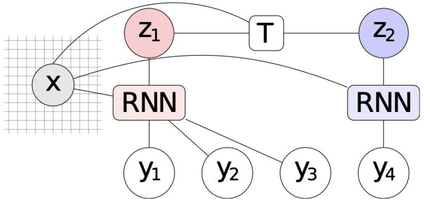
The outputs are converter to text by argmax with the embeddings of the vocab.

Conditional HSMM

$$p(y, z \mid x) = \prod_{s=1}^{S} \underbrace{p(z_s \mid z_{s-1}, x)}_{\text{transition prob}} \; \underbrace{p(l_s \mid z_s)}_{\text{length prob}} \; \underbrace{p(y_{t_0(s):t_1(s)} \mid z_s, l_s, x)}_{\text{segment prob}}$$

HSMM decoder gives us the text segments by using argmax to the vocabulary with the output. This introduces output probability. The RNN we used here is LSTM to generate the probabilities and attention is calculated by the relevance with the final hidden state to the embeddings. Copy attention is included to increase the attention weights of the tokens that are exactly the same.

This parameterizes the output probabilities with neural components to include them.



Segment probability by RNN + attention + copy attention

Training

We get the input x and y pairs to train, z is observed during training. We need to maximize the probability of z to match with y.

$$\ln p(y_{1:T} \mid x) = \ln \sum_{z} p(y_{1:T}, z \mid x)$$

We used the same dynamic program analogous to the forward or backward algorithm used in learning HMMs.

I.e., we will generate output for a given x by argmax p(y, z|x)

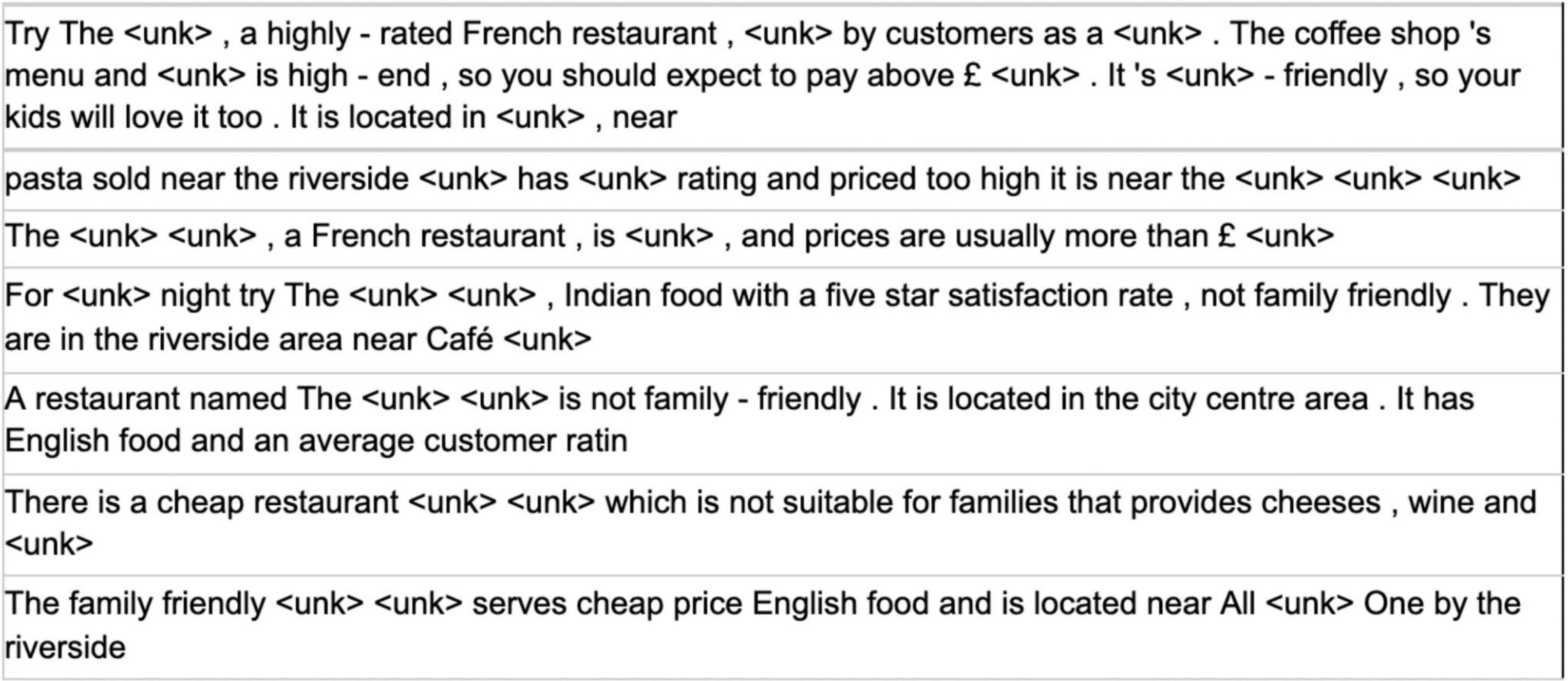
We condition RNNs on latent state labels by concatenating state embedding to the RNN input. This doesnot allow the splitting up of segments labels.

Results :

All the generated templates for E2E Data is in the segs folder in our submission with file name - seg-e2e-300-60-1-far.txt. All the pretrained models are in the folder models.

Example templates we generated :

Try The <unk> , a highly - rated French restaurant , <unk> by customers as a 5 out of 5 . The coffee shop 's menu and <unk> is high - end , so you should expect to pay above £ <unk> . It 's children - friendly , so your kids will love it too . It is located in Riverside , near|57



Try The <unk> , a highly - rated French restaurant , <unk> by customers as a <unk> . The coffee shop 's menu and <unk> is high - end , so you should expect to pay above £ <unk> . It 's <unk> - friendly , so your kids will love it too . It is located in <unk> , near
pasta sold near the riverside <unk> has <unk> rating and priced too high it is near the <unk> <unk> <unk>
The <unk> <unk> , a French restaurant , is <unk> , and prices are usually more than £ <unk>
For <unk> night try The <unk> <unk> , Indian food with a five star satisfaction rate , not family friendly . They are in the riverside area near Café <unk>
A restaurant named The <unk> <unk> is not family - friendly . It is located in the city centre area . It has English food and an average customer ratin
There is a cheap restaurant <unk> <unk> which is not suitable for families that provides cheeses , wine and <unk>
The family friendly <unk> <unk> serves cheap price English food and is located near All <unk> One by the riverside

Information about the Templates Generated

We are able to successfully generate the templates that are to mapped with the records to get the summary. All The <unk> words should be mapped.

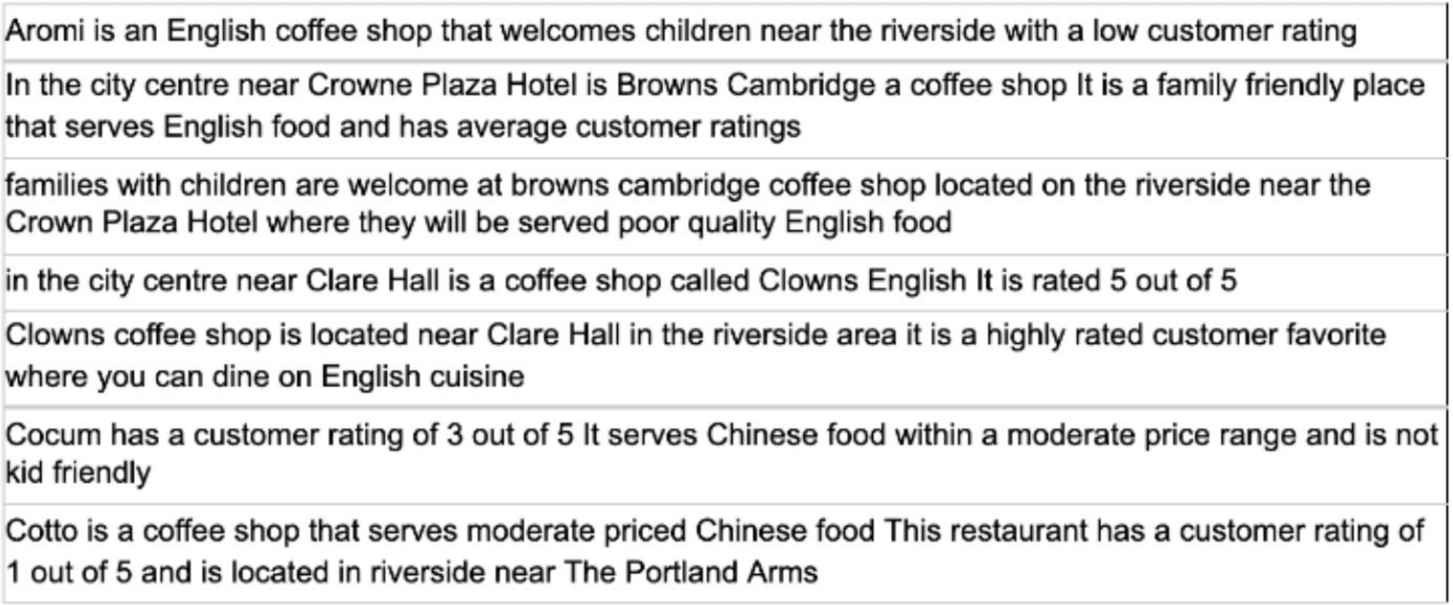
The summary from our intuition also makes sense.

Baseline ++ - Our Approach :

The HSMM decoder incorporated by LSTMs and attention, which have an excellent track record of effective neural text generation, while keeping the HSMM structure. The LSTMs are very good at predicting, and is explicitly designed to avoid long term dependency problems using 3 different gates. Remembering the long sequences for a long period of time is its way of working.

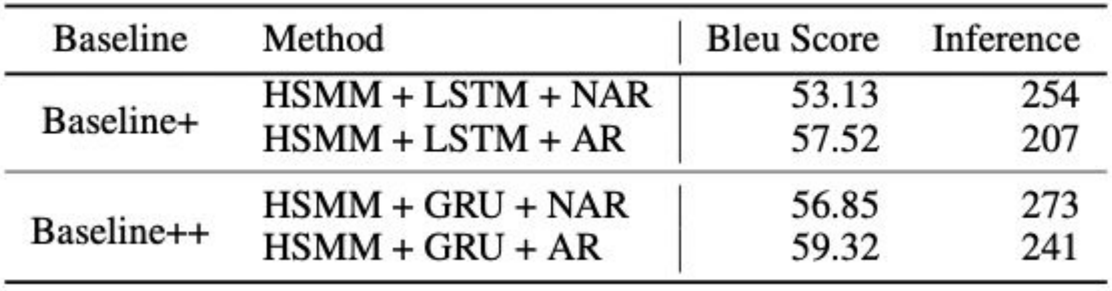
The templates in our datasets are not too long but long enough that GRUs can handle in a pretty good way. The GRUs do not have 3 different gates like LSTMs, has 2 gates, update and reset gate. This makes it faster than LSTM. GRUs are computationally efficient and requires lesser number of parameters. We used this as our approach on top of baseline.

Results - E2E



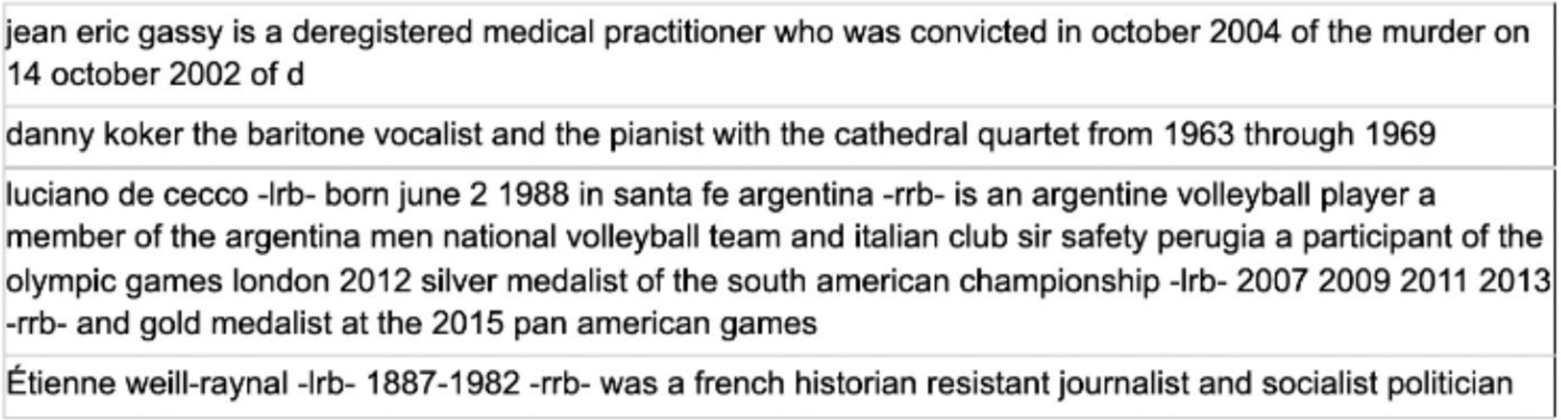
Aromi is an English coffee shop that welcomes children near the riverside with a low customer rating
In the city centre near Crowne Plaza Hotel is Browns Cambridge a coffee shop It is a family friendly place that serves English food and has average customer ratings
families with children are welcome at browns cambridge coffee shop located on the riverside near the Crown Plaza Hotel where they will be served poor quality English food
in the city centre near Clare Hall is a coffee shop called Clowns English It is rated 5 out of 5
Clowns coffee shop is located near Clare Hall in the riverside area it is a highly rated customer favorite where you can dine on English cuisine
Cocum has a customer rating of 3 out of 5 It serves Chinese food within a moderate price range and is not kid friendly
Cotto is a coffee shop that serves moderate priced Chinese food This restaurant has a customer rating of 1 out of 5 and is located in riverside near The Portland Arms

Evaluation - E2E :



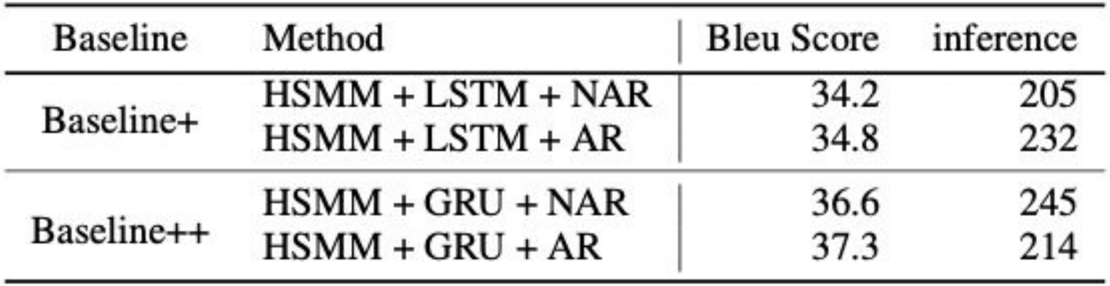
Baseline	Method	Bleu Score	Inference
Baseline+	HSMM + LSTM + NAR	53.13	254
	HSMM + LSTM + AR	57.52	207
Baseline++	HSMM + GRU + NAR	56.85	273
	HSMM + GRU + AR	59.32	241

Results - WikiBio :



jean eric gassy is a deregistered medical practitioner who was convicted in october 2004 of the murder on 14 october 2002 of d
danny koker the baritone vocalist and the pianist with the cathedral quartet from 1963 through 1969
luciano de cecco -lrb- born june 2 1988 in santa fe argentina -rrb- is an argentine volleyball player a member of the argentina men national volleyball team and italian club sir safety perugia a participant of the olympic games london 2012 silver medalist of the south american championship -lrb- 2007 2009 2011 2013 -rrb- and gold medalist at the 2015 pan american games
Étienne weill-raynal -lrb- 1887-1982 -rrb- was a french historian resistant journalist and socialist politician

Evaluation - WikiBio :



Baseline	Method	Bleu Score	inference
Baseline+	HSMM + LSTM + NAR	34.2	205
	HSMM + LSTM + AR	34.8	232
Baseline++	HSMM + GRU + NAR	36.6	245
	HSMM + GRU + AR	37.3	214

Future work for Final Deliverable :

Whole Code is available at :

Baseline:
https://colab.research.google.com/drive/1jTZoXthTO5vuTWxDDLZWfveLeXzbQ0K4?usp=sharing

Baseline+ and Baseline++ :

https://drive.google.com/drive/folders/1t5AFgTuboCSEtuBEyIyngtIKb6IUbaoS?usp=sharing

Thank You