Word Saliency Methods for Non-Autoregressive Logical Data-to-Text Generation

Yale

Luke Benson

Department of Statistics & Data Science, New Haven, CT

LILY Lab

Introduction

Logical data-to-text generation (D2T) is an emerging task within the field of natural language generation (NLG). Given a structured representation of data, such as a table, logical D2T models are trained to produce human-readable summarizations which communicate logical operations over several data points. There are a myriad of real-world applications of these models, including generating detailed weather forecasts, sports game summaries, and business reports.

The current state-of-the-art logical D2T models utilize autoregressive (AR) decoding techniques: sentences are produced unidirectionally, and the generation of each word is conditioned on previously-generated words. The challenge that these AR techniques face, however, is that each generated word is unaware of the words that will appear after it. In this project, we assess the fluency and fidelity of D2T sentences which are produced through a non-autoregressive (NAR) generation scheme. To further improve upon these initial sentences, we inject word saliency priors into the training of our NAR decoders.

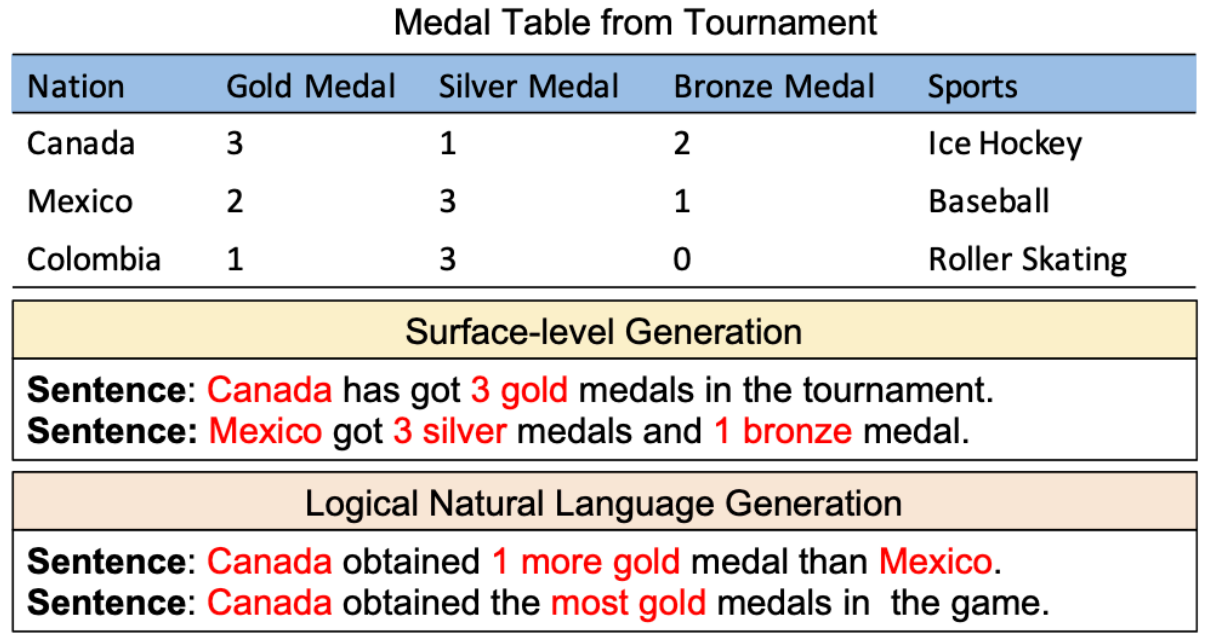
Materials and Methods

ToTTo (120k sentences) and LogicNLG (28k sentences) are two open-domain table-to-text data sets. The target sentences for both data sets utilize logical reasoning that go beyond surface-level summarizations. Our goal is to develop logical D2T models for LogicNLG, and so we first pre-train our models on ToTTo before refining the model parameters on LogicNLG.

The NAR decoding algorithm we utilize is called mask-predict. This algorithm generates all words in parallel and then “masks” the words that it is least confident in. It then generates the most probable words for those masked tokens. This process repeats over a pre-set number of iterations.

During training of the typical mask-predict decoder, words are masked at random. However, we hypothesize that masking words strategically during training may improve the performance of our models. If we mask the words that are most salient or important, our models may learn to predict more important words effectively when given less important words. Thus, we seek to incorporate a measure of each word’s salience to the data set into our NAR D2T training.

We propose a measure of word saliency based on the process of word erasure and calculated using GPT-2. For every sentence in ToTTo and LogicNLG, we use GPT-2 to first compute the probability of that sentence occurring. For a given sentence, we then remove each word and calculate the percentage change in the sentence probability when the word is erased. This process tries to capture how important a word is to the probability of a summary and results in a word saliency score for every word in every sentence. For each word, we then average every saliency score across all sentences to arrive at an average word saliency within each data set.



Medal Table from Tournament

Nation	Gold Medal	Silver Medal	Bronze Medal	Sports
Canada	3	1	2	Ice Hockey
Mexico	2	3	1	Baseball
Colombia	1	3	0	Roller Skating

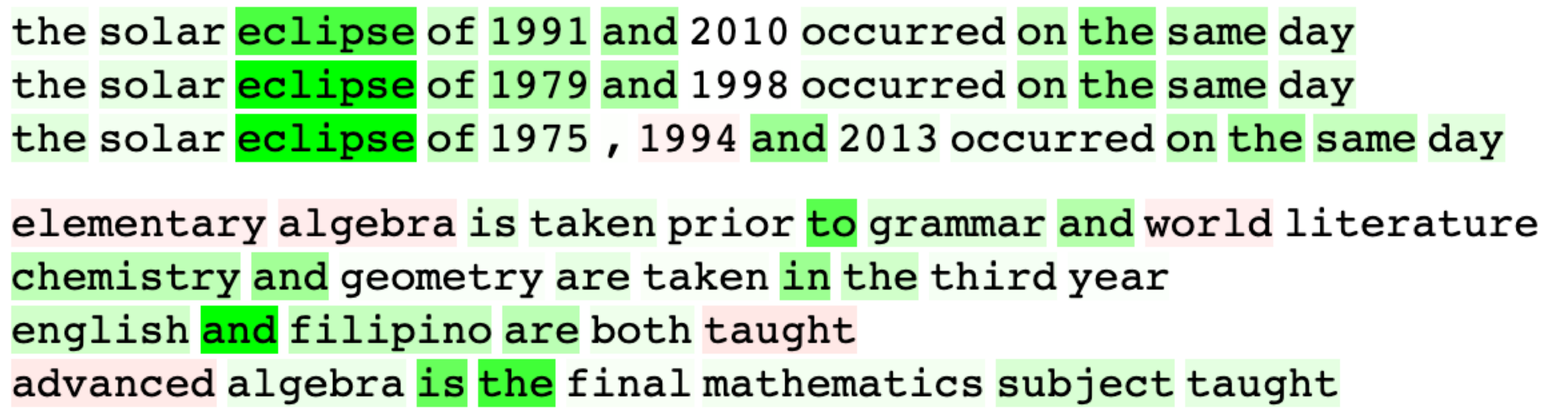
Surface-level Generation

Sentence: Canada has got 3 gold medals in the tournament.
Sentence: Mexico got 3 silver medals and 1 bronze medal.

Logical Natural Language Generation

Sentence: Canada obtained 1 more gold medal than Mexico.
Sentence: Canada obtained the most gold medals in the game.

Figure 1. The distinction between surface-level and logical generations. (*Source: LogicNLG)*



the solar eclipse of 1991 and 2010 occurred on the same day
the solar eclipse of 1979 and 1998 occurred on the same day
the solar eclipse of 1975 , 1994 and 2013 occurred on the same day

elementary algebra is taken prior to grammar and world literature
chemistry and geometry are taken in the third year
english and filipino are both taught
advanced algebra is the final mathematics subject taught

Figure 2. Word saliency scores as determined by GPT-2 for example sentences in the LogicNLG training set.

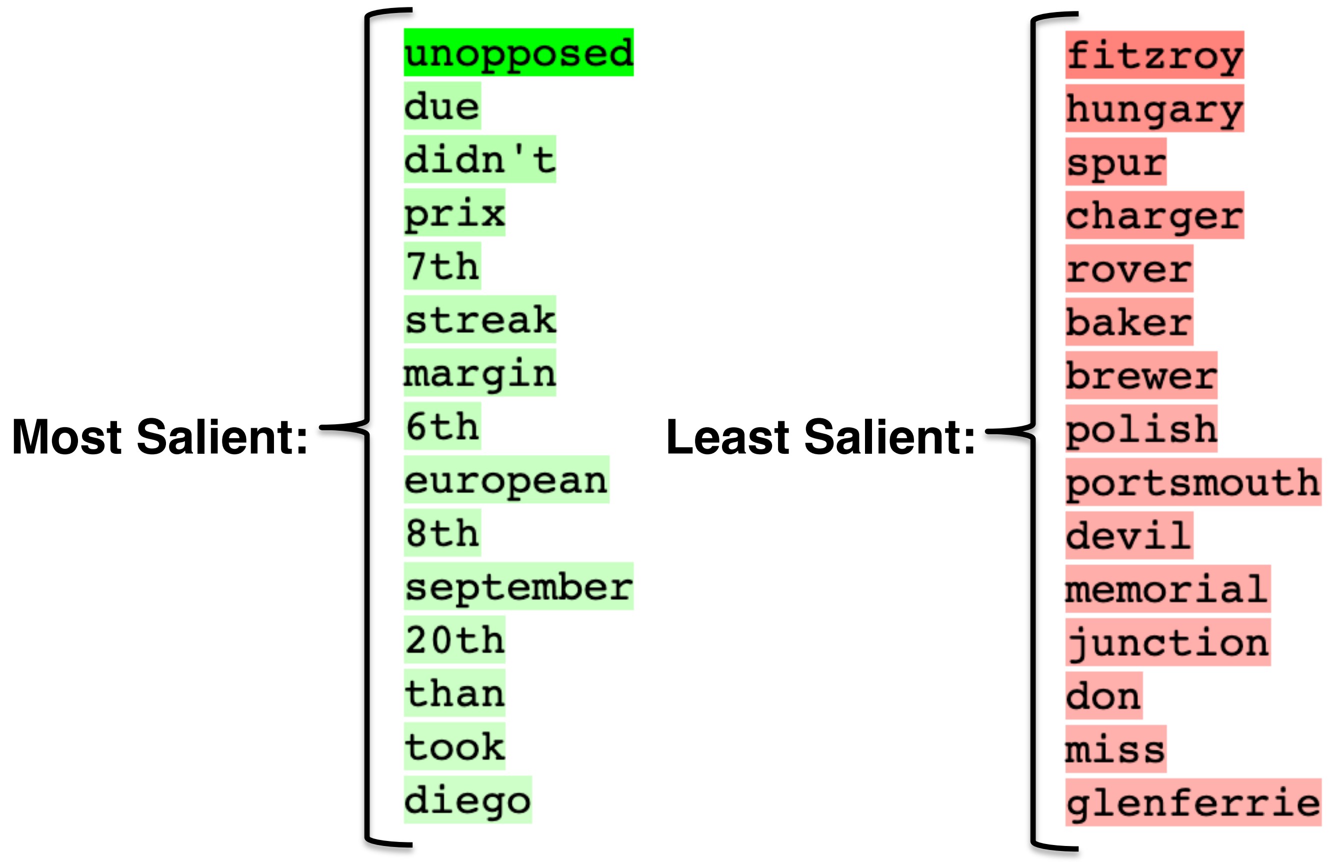


Figure 3. The most and least salient frequent words (25+ instances throughout 28k sentences) for the LogicNLG training set.

Results

1) *Word Saliency*

For both data sets, word saliency is calculated for each word that appears. Through qualitative evaluation of these score, we can see that many of the least salient words are named entities directly contained within the data set (‘hungary’, ‘charger’, ‘rover’). Some of the most salient words either reflect the relationships between entities (‘unopposed’, ‘7th’, ‘6th’) or link phrases (‘due’, ‘than’). These observations align with the fact that word saliency is measured through the effect of word erasure.

2) *D2T Generations*

Our standard NAR model trained through the mask-predict algorithm performs comparably, but consistently worse, than the standard AR decoder model across all fluency and fidelity metrics. We are currently training the NAR decoder which incorporates our word saliency priors, and hypothesize that the inclusion of this additional information will only improve baseline model performance.

Decoding Technique	Fluency		Fidelity	
	sacreBLEU	BERTScore	NLI-Acc	SP-Acc
AR	15.40	87.75	69.69	41.02
Standard NAR	13.20	86.98	54.52	39.51
Saliency-based NAR	*In Progress*	*In Progress*	*In Progress*	*In Progress*

Table 1. Initial fluency and fidelity metric results for autoregressive, standard non-autoregressive, and saliency-based non-autoregressive generations.

Conclusion and Future Work

The contribution of this project is twofold: 1) we propose a measure of word saliency within a given data set through word erasure techniques, and 2) we incorporate these saliency scores into an investigation of an alternative training scheme for the task of NAR logical D2T generation. Further quantitative research into the word saliency scores could better illuminate how these scores could be used effectively for this task. Either 1) additional model training or 2) greater pre-training are also needed to fully evaluate the potential of standard and saliency-based NAR methods here.

Acknowledgement

I would like to thank Professor Dragomir Radev for advising me on this project. I would also like to thank Linyong Nan for his guidance and assistance throughout the process.