Generation of Text from Structured Data

11-785 Final Project

Carnegie Mellon University

Anusha Prakash, Mansi Gupta, Raghuveer Chanda, Srividya Potharaju

Language Technologies Institute

Abstract

We wish to tackle text generation using structured data like tables, info-boxes, databases, etc to present meaningful information. We use the RotoWire dataset, that consists of basketball game statistics and corresponding document level textual description for each game. We use neural encoder-decoder model, and experiment with variations of encoder units, attention and copy mechanism..

Motivation and Contribution

Generating long form documents is significantly harder than generating short summaries. The generated text might not be able to capture all the important facts in the structured input and often contains factual errors or hallucinate factual statements. Additionally, our input data is structured and has a hierarchical representation that poses a unique challenge. We propose a hierarchical attention based pointer generator model to address the challenge

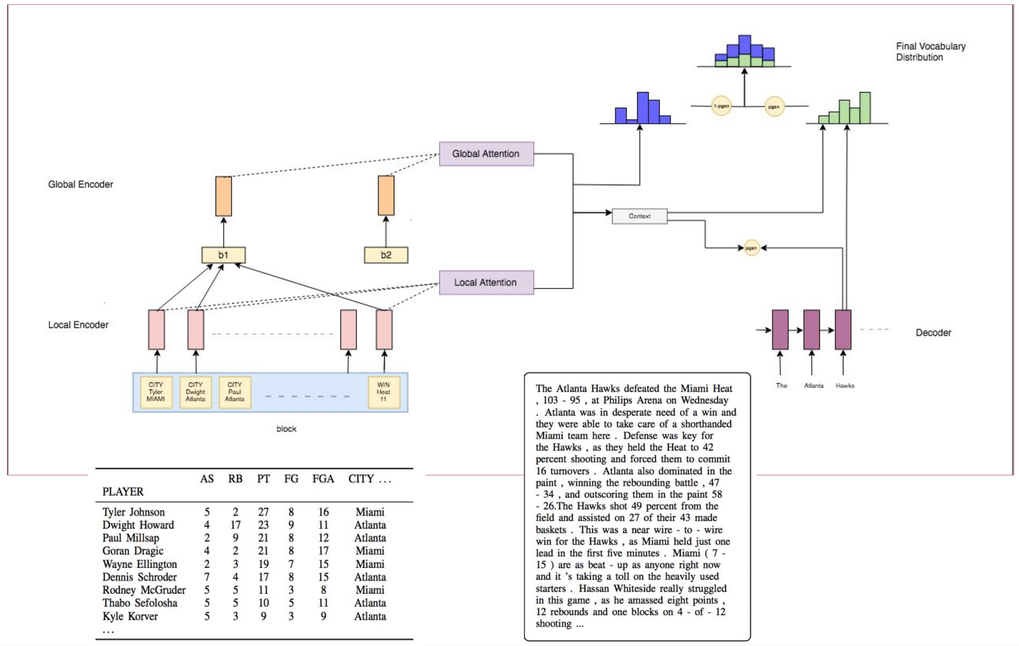
Hierarchical Attention - Pointer Generator Model

- Generation Probability p_{gen} is calculated from the context vector, decoder state and decoder input

$$p_{\text{gen}} = \sigma(w_h^T h_t^* + w_s^T s_t + w_x^T x_t + b_{\text{ptr}})$$

- It is used as a switch to choose between generating a word from vocabulary by sampling from P_{vocab} or copying a word from the input sequence by sampling from attention distribution

$$P(w) = p_{\text{gen}} P_{\text{vocab}}(w) + (1 - p_{\text{gen}}) \sum_{i:w_i = w} a_i^t$$



PLAYER	AS	RB	PT	FG	FGA	CITY ...
Tyler Johnson	5	2	27	8	16	Miami
Dwight Howard	4	17	23	9	11	Atlanta
Paul Millsap	2	9	21	8	12	Atlanta
Goran Dragic	4	2	21	8	17	Miami
Wayne Ellington	2	3	19	7	15	Miami
Dennis Schroder	7	4	17	8	15	Atlanta
Rodney McGruder	5	5	11	3	8	Miami
Thabo Sefolosha	5	5	10	5	11	Atlanta
Kyle Korver	5	3	9	3	9	Atlanta
...						

The Atlanta Hawks defeated the Miami Heat , 103 - 95 , at Philips Arena on Wednesday . Atlanta was in desperate need of a win and they were able to take care of a shorthanded Miami team here . Defense was key for the Hawks , as they held the Heat to 42 percent shooting and forced them to commit 16 turnovers . Atlanta also dominated in the paint , winning the rebounding battle , 47 - 34 , and outscoring them in the paint 58 - 26.The Hawks shot 49 percent from the field and assisted on 27 of their 43 made baskets . This was a near wire - to - wire win for the Hawks , as Miami held just one lead in the first five minutes . Miami (7 - 15) are as beat - up as anyone right now and it 's taking a toll on the heavily used starters . Hassan Whiteside really struggled in this game , as he amassed eight points , 12 rebounds and one blocks on 4 - of - 12 shooting ...

Training Instance : $(s, y_{1:T})$
Each tuple: $r \in s$

```
r (e, m) = {r.t : r ∈ s, r.e = e, r.m = m}
r.t -> City
r.e -> Tyler Johnson
r.m -> Miami
```

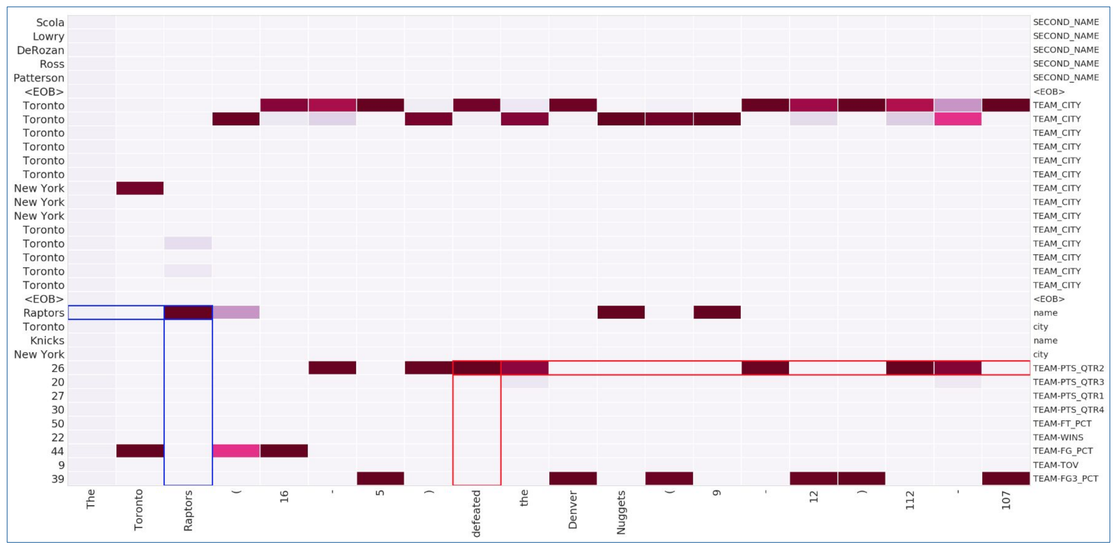
Results

Evaluation Metrics

- Content Selection (CS)
  - Measures how well generated text matches actual text in terms of unique relations.
- Relation Generation (RG)
  - Measures how well system generates text containing factual records.
- Content Ordering (CO)
  - Measures how well system orders the records it chooses to discuss.

Model	BLEU	RG (P% / #)	CS (P% / R%)	CO
Linear Encoder + Attention	8.87	0.29 / 0.08	0.13 / 0.20	0.23
RNN Encoder + Attention	6.95	0.41 / 0.08	0.24 / 0.30	0.28
BiLSTM Encoder + Attention	8.96	0.32 / 0.11	0.29 / 0.41	0.30
Joint Copy Mechanism*	13.61	62.7 / 16.5	21.4 / 31.8	10.4
Conditional Copy*	14.19	75.0 / 16.7	25.9 / 31.9	11.0
Hierarchical Attention + Pointer Networks	**11.56**	**61.07 / 12.12**	**20.2 / 29.5**	**9.98**

Global+Local Attention Visualization



- We show the visualization of combined global and local attention weights in this graph.
- Y-axis contains the input tuples with value on the left and type (block) on the right.
- Note that `'Raptors'` is copied from the same word, while `'defeated'` attends over `TEAM-PTS_QTR2`

Discussion and Future Work

Type	Text
Original Text	The Atlanta Hawks (46 - 12) beat the Orlando Magic (19 - 41) 95 - 88 on Friday. Al Horford had a good all - around game , putting up 17 points , 13 rebounds , four assists and two steals in a tough matchup against Nikola Vucevic .... They blew a seven point lead with 43 seconds remaining and they might have carried that with them into Friday 's contest against the Hawks .... The Magic will host the Charlotte Hornets on Sunday , and the Hawks with take on the Heat in Miami on Saturday .
Generated Text	The Atlanta Hawks (6 - 2) defeated the Orlando Magic (7 - 12) 104 - 98 on Wednesday at the Amway Center in Atlanta . The Hawks got off to a quick start in this one , out - scoring the Magic 36 - 18 in the first quarter alone .... The Hawks also secured a 43 - point advantage into the half , while the Magic went just 42 percent from the floor and a meager 26 percent from beyond the arc .... The Magic 's next game will be at home against the Atlanta Hawks on Wednesday , while the Magic will be at home against the Atlanta Hawks on Wednesday .

Comparison between gold output and generated text for given set of input records. The blue text shows the parts the generated output gets factually correct. Even after implementing copy, the network gets many facts wrong.

- Hierarchical copy models can be implemented with the hope of generating the facts correctly more often.
- On decoding front, instead of relying on a non-optimal greedy search algorithm that suffers from non-differentiability, Gumbel SoftMax reparameterization can be used.

References

Wiseman, S., Shieber, S.M. and Rush, A.M., 2017. Challenges in data-to-document generation.
Gu, J., Lu, Z., Li, H. and Li, V.O., 2016. Incorporating copying mechanism in sequence-to-sequence learning.
Yang, Zichao, et al. "Hierarchical attention networks for document classification."
See, Abigail, et. al. "Get to the point: Summarization with pointer-generator networks."