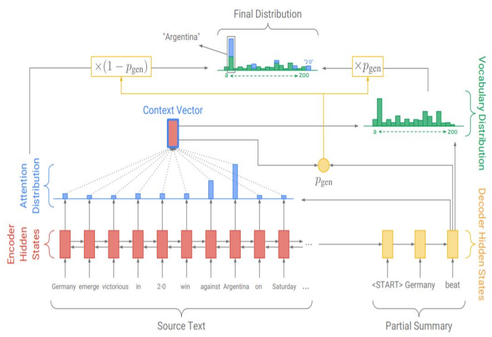
A More Abstractive Text Summarization Model

Satyaki Chakraborty, Xinya Li

Introduction

- Pointer-generator networks are the most popular non-RL based methods for summarizing large documents.
- Although it aims to learn **hybrid** summaries, the amount of abstraction in the summaries is significantly less. We use novelty as a metric to determine the level of abstraction following [1] and [2].
- We first conduct a survey where we ask individuals to label whether they prefer abstract or extractive summaries. From our survey, we observe humans mostly prefer abstract summaries over extractive ones.
- Hence, we intend to improve upon the level of abstraction of pointer-generator networks.

Model



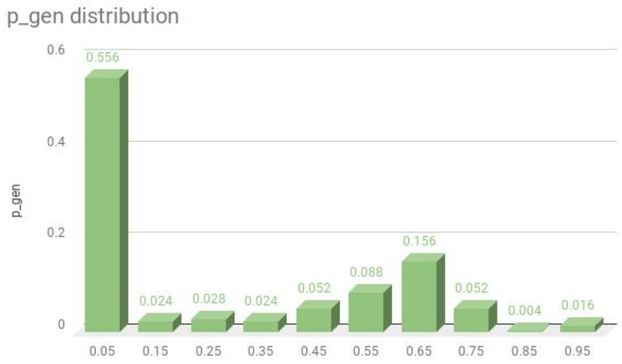
- p_{gen} is a switch that controls whether we are going to sample a new word from the vocab distribution or copy a word from the source text using the attention distribution
- Copy distribution helps us to produce words that are out-of-vocabulary (OOV) but in the source text.
- Following set of equations govern the decoding step of the model

$$a_i^t = softmax(v^T tanh(W_h h_i + W_s s_t + b_{attn}))$$
$$p_{attn} = \{a_0^t, a_1^t, .., a_N^t\}$$
$$h_t^* = \sum_i a_i^t h_i$$
$$p_{vocab} = softmax(V[s_t, h_t^*] + b)$$
$$p_{gen} = \sigma(w_{h^*}^T h_t^* + w_s^T s^t + w^T x^t + b_{ptr})$$
$$p_{final} = (1 - p_{gen}) * p_{attn} + p_{gen} * p_{vocab}$$

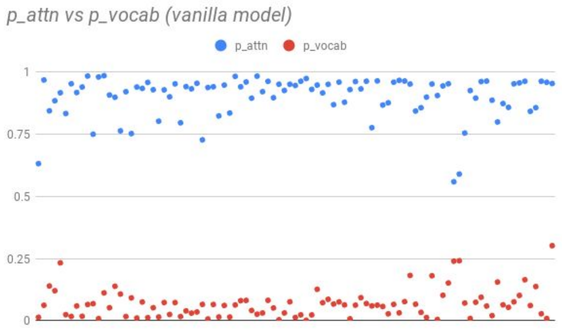
Survey

- We conducted a survey to investigate people's preferences between **abstractive** and **extractive** ground-truth summaries from the Document Understanding Conference 2003 and 2004 summarization dataset [3]. We received in total 15 responses, and on average across all choices, 66.7 % are in favor of the abstractive summaries.

Observations



- p_{gen} distribution over 1000 random samples of generated summaries by the vanilla model shows that, it has a tendency to copy words from the source text than generate new words.
- Above graph shows, p_{gen} is less than 0.1 for more than 55% of the time and more than 0.8 for much less than 1% of the time. This explains why novelty of the baseline model is so low.



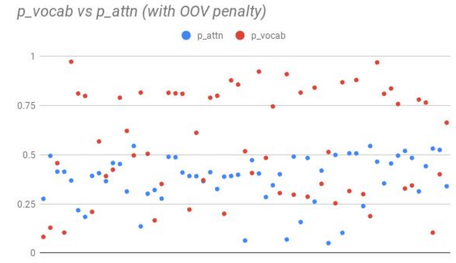
- Moreover, p_{vocab} of the words being sampled are significantly less compared to their corresponding p_{attn}.
- This could be because the attention distribution is predicted over a vector with dimensionality 400, where as in the vocab distribution probability mass is distributed across a vector of dimensionality ~50000.
- If attention probabilities are higher, p_{gen} will prefer the attention distribution over vocabulary distribution which explains the graph in first observation.

Modification

- Attention distribution should be definitely preferred when there is an OOV word, otherwise should we rely on the vocab distribution to generate words? If so, does that improve novelty?
- Adding OOV penalty term to penalize p_{gen}

$$L_{OOV} = -y_{oov} log(1 - p_{gen}) - (1 - y_{oov}) log(p_{gen})$$

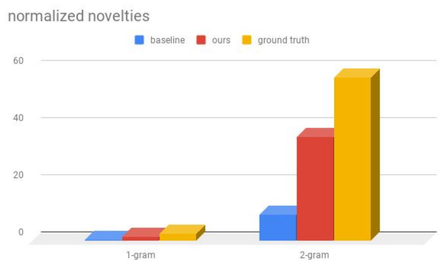
Results



- p_{vocab} of the words being sampled are much higher after adding the penalty term. .
- Our network learns to produce more abstract summaries shown below.
- We use n-gram novelty metric used in [2] to obtain level of abstraction.

$$N(x^{gen}, n) = \frac{||ng(x^{src}, n) - ng(x^{gen}, n)||}{||ng(x^{gen}, n)||}$$

- 1-gram and bigram normalized novelties for 1000 randomly chosen samples are shown below.



FUTURE WORK

- Even though Rouge scores are not suitable metrics for measuring the level of abstraction, we plan to obtain R1, R2, RL scores to see how they compare against those of the baseline model.

REFERENCES

[1] See, Abigail, Peter J. Liu, and Christopher D. Manning. "Get to the point: Summarization with pointer-generator networks." arXiv preprint arXiv:1704.04368 (2017).
[2] Kryściński, Wojciech, et al. "Improving abstraction in text summarization." arXiv preprint arXiv:1808.07913 (2018).
[3] PAST DATA, Natural Institute of Standards and Technology, www-nlpir.nist.gov/projects/duc/data.html.