Abstractive Headline Generation: Statistical versus Deep Learning Approaches

Sujeong Cha & Emmy Phung

Agenda

1. Problem Statement
2. Dataset - NewSHead
3. Approach 1 - Statistical Model
4. Approach 2 - Deep Learning Model
5. Conclusion

Problem Statement

- Problem
  - Readers want to read less but get more
  - Informative headlines help guide readers to find the right content

- Text Summarization

❌ Extractive Summarization:
  - Does not work because: len(headline) < len(sentence)

✔ Abstractive Summarization:
  - Statistical Model (Banko et. al., 2000)
  - NHNet (Gu et. al., 2020)

- Evaluation Metrics
  - ROUGE 1-F (unigram, F1-based)
  - ROUGE L-F (longest common subsequence, F-1 based)

Dataset

- NewSHead Dataset
  - Released by Google Research [Gu et al., 2020]
  - Recent and the largest news dataset
    - articles published between May 2018 and May 2019
    - contains 369,940 stories with 932,571 articles (3-4 articles / story)
  - “Story headline” generated by crowd-sourced curators
- For our experiments, we picked 1 article from each story in the valid & test set

	Train	Valid	Test
Original NewSHead	359,940 stories	5,000 stories	5,000 stories
Our Experiments	X	5,000 articles	5,000 articles

└ Re-split into 80:10:10

Approach 1: Statistical Model [Banko et al., 2000]

- Find a headline of length n that maximizes $P(w_1, ..., w_n)$

Emission matrix

Transition matrix

$$P(w_1, ..., w_n) = \prod_{i=1}^{n} P(w_i \in H | w_i \in D) \cdot P(len(H) = n) \cdot \prod_{i=1}^{n} P(w_i | w_1, ..., w_{i-1})$$

Headline length

* H: headline, D: article body

- Our goal: 1) Implement the model given in the paper
  2) Replace transition matrix with neural LM (Open AI GPT-2)

Approach 1: Result

- Successfully implemented the model (Table 1)
  - Similar to POS tagging problem:
    1] Emission/transition from training set → 2] Decode with Viterbi
  - Marginal gap in performance due to preprocessing/splits ambiguity
- Fine-tuned neural LM improves headline generation (Table 2)
  - 33% improvement over the vanilla Banko in terms of ROUGE-1 score

Table 1. Replication Result

	Reuters
Banko et al.	0.1754
Our Replication	0.2220

Table 2. Banko + Neural LM Result on NewSHead

Model	ROUGE 1-F	ROUGE L-F
Banko et al. (Baseline)	0.1354	0.1345
w/ Pre-trained GPT-2	0.1644	0.1676
w/ Fine-tuned (on training set) GPT-2	0.1669	0.1604
w/ Fine-tuned (on test set) GPT-2	0.1805	0.1726

Approach 2: NHNet [Gu et al., 2020]

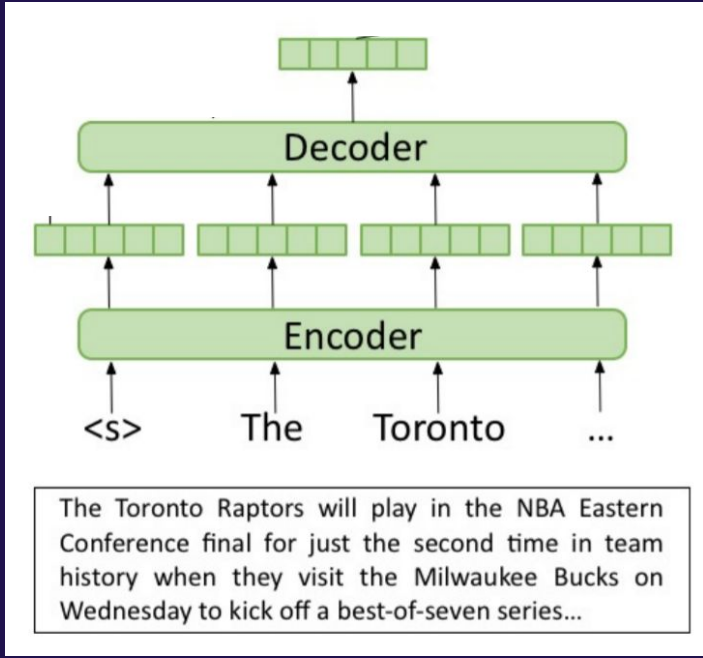
- Original NHNet model predicts a single story headline for a set of articles

  → Goal 1: Perform single-doc title prediction and compare this task with story headline generation

- Original NHNet's encoder: Bert-base-uncased

  → Goal 2: Pre-train NHNet's encoder on GLUE tasks to boost performance
  1. Pre-train on MNLI (Textual Entailment)
  2. Pre-train on MRPC (Paraphrase Detection)
  3. PRe-train on MNLI then on MRPC (sequential)

Figure 1: NHNet Encoder-Decoder for Single-Doc



Approach 2: Result

- Pre-train the encoder on MNLI gives the best performance
- Sequential pre-training leads to performance degradation → catastrophic forgetting
- Predicting Original Title is much more challenging than predicting Story Headline

Table 1: Impact of Transfer Learning

Model	ROUGE 1-F	ROUGE L-F
NHNet (Baseline)	0.1412	0.701
MNLI-Pre-trained	0.1823	0.1109
MRPC-Pretrained	0.1519	0.0825
MNLI & MRPC -Pretrained	0.1603	0.0915

Story headline: *human-annoted, fact-based*

Original title: *subjective to authors' writing style, may contain complex expression (ie. idioms)*

Table 2. Story vs. Title Labels

	ROUGE 1-F		ROUGE L-F	
	Story Headline	Original Title	Story Headline	Original Title
NHNet (Baseline)	0.2455	0.1412	0.0910	0.0701
MNLI-Pretrained	0.2220	0.1823	0.0888	0.1109

Conclusion

Statistical versus Deep-learning?

- Best performance from each: (Banko) 0.1805 vs. (NHNet) 0.1823
  ** NHNet was not ran on its full-blown capacity due to limited computing resources*
- *Given limited data and computing power*, a statistical model still equally perform well.
- A traditional statistical model can still play an important role in this era where only deep-learning models are valued and sought after.

Other lessons we learn:

- Abstractive headline generation, or abstractive text summarization in general, remains as a surprisingly challenging task in NLP.
- Building an abstractive headline generation model that produces a title that sounds as “fluent”l as a human writer is still likely to require decades of research.

Thank You