Neural Document Summarization by Jointly Learning to Score and Select Sentences

Qingyu Zhou[1*], Nan Yang[2], Furu Wei[2], Shaohan Huang[2], Ming Zhou[2], Tiejun Zhao[1]

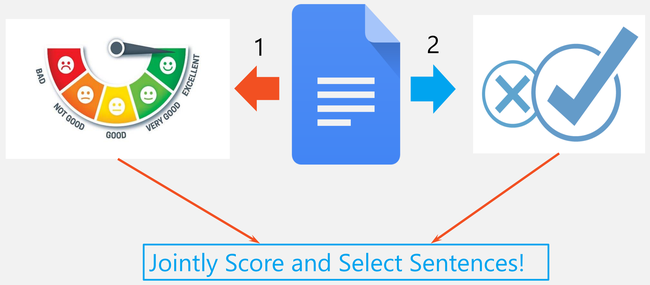
[1]Harbin Institute of Technology, Harbin, China [2]Microsoft Research, Beijing, China

* Work done during an internship at Microsoft Research.

Extractive Document Summarization

Sentence Scoring
- Feature based methods: word probability, TF-IDF, etc…
- Graph-based methods: TextRank, LexRank, etc…
- Neural Network based: CNN/RNN sentence encoding

Sentence Selection
- Simple greedy selection
- Maximal Marginal Relevance (MMR)
- Integer Linear Programming (ILP)
- Submodular functions



Experiments

Dataset: CNN/Daily Mail dataset
Evaluation Metric: ROUGE F1 (full length)
Human Evaluation: Rankings of NEUSUM and NN-SE (lower is better)

Models	ROUGE-1	ROUGE-2	ROUGE-L	
LEAD3	40.24	17.70	36.45	Rule-based
TEXTRANK	40.20	17.56	36.44	Unsupervised
CRSUM	40.52	18.08	36.81	Separated methods
NN-SE	41.13	18.59	37.40	Separated methods
PGN‡	39.53	17.28	36.38	Abstractive
LEAD3‡ *	39.2	15.7	35.5	
SUMMARUNNER‡ *	39.6	16.2	35.3	
NEUSUM	**41.59**	**19.01**	**37.98**	

* Systems trained and evaluated on the anonymized dataset, and so are not strictly comparable to our results on the original text.

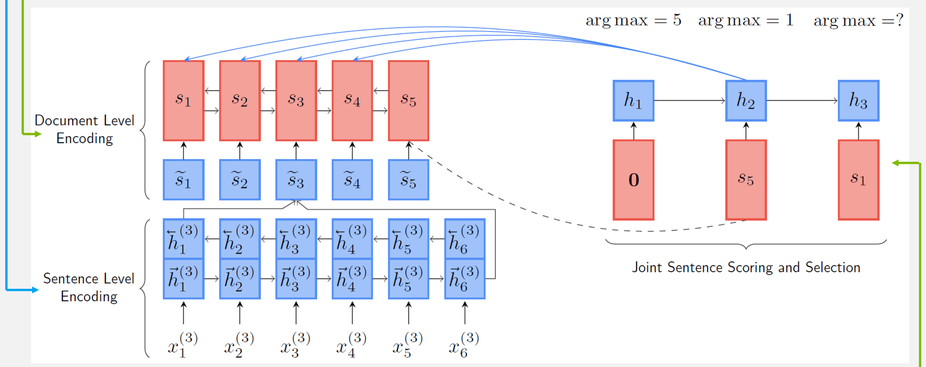
- Our NeuSum model beats the state-of-the-art systems

Models	Info	Rdnd	Overall
NN-SE	1.36	1.29	1.39
NEUSUM	**1.33**	**1.21**	**1.34**

Informativeness (Info), redundancy (Rdnd) and overall quality of NN-SE and NeuSum on a sampled set (lower is better)

Jointly Learning to Score and Select Sentences (NeuSum)

Hierarchical Document Encoding
- Sentence Level Encoding
  - Words to sentence (BiGRU, concat last forward and backward)
- Document Level Encoding
  - Context information of sentences (BiGRU, concat forward and backward)



Sentence Extractor
- GRU (maintain the extraction state)

$$h_t = \mathrm{GRU}(s_{t-1}, h_{t-1})$$

Extractor state; Last extracted sentence vector

Sentence Scoring
- Two-layer MLP (calculate score gain)

$$\delta(S_i) = W_s \tanh(W_q h_t + W_d s_i)$$

Score of sentence *i*; Representation of sentence *i*

Sentence Selection:
- Choose the sentence with maximum score gain

Objective function
- Kullback-Leibler (KL) divergence

$$g(S_i) = r(\mathbb{S}_{t-1} \cup \{S_i\}) - r(\mathbb{S}_{t-1})$$

$$\tilde{g}(S_i) = \frac{g(S_i) - \min(g(S))}{\max(g(S)) - \min(g(S))}$$

Score gain; ROUGE score; Min-Max Normalization

$$P(\hat{S}_t = S_i) = \frac{\exp(\delta(S_i))}{\sum_{k=1}^{L} \exp(\delta(S_k))}$$

Model predicted score

Loss: $J = D_{KL}(P \parallel Q)$

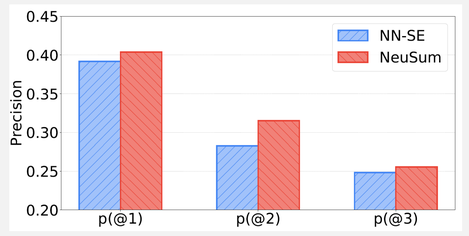
$$Q(S_i) = \frac{\exp(\tau \tilde{g}(S_i))}{\sum_{k=1}^{L} \exp(\tau \tilde{g}(S_k))}$$

Temperature

Advantages:
- Dynamic Sentence Scoring: Score each sentence every time when selecting sentences
- Selection strategy can be simple
- End-to-end training

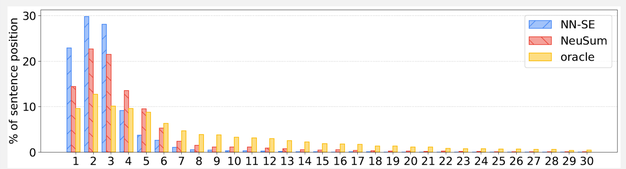
Discussion

1. Precision at Step-t



- NeuSum is slightly better for the first selection
  - Best sentence might be easy to find
- Large difference at the second step
  - NeuSum can remember the selection history
- Similar again for the last selection
  - Error propagation

2. Position of Selected Sentences



- Oracle is much more diverse
- Separated method (NN-SE) chooses lots of LEAD3 sentences (80.91%)
- NeuSum selects less LEAD3 sentences (58.64%)

Conclusion

- Joint sentence scoring and selection enables more accurate and diverse (position) selection
- Sentence scoring can leverage information from selection history
- Future work: adapt NeuSum for multi-document summarization

Acknowledgements:
i. National Key Research and Development Program of China (No. 2017YFB1002102)
ii. Project of National Natural Science Foundation of China (No. 91520204)
iii. Harbin Institute of Technology Scholarship Fund

Source code and related resources are available at https://res.qyzhou.me