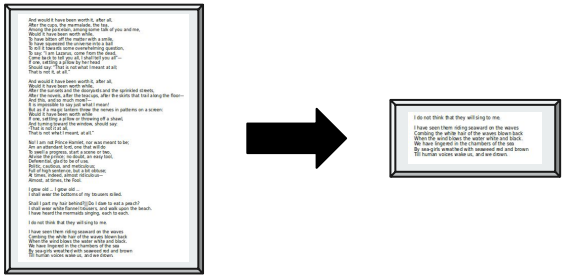
Abstractive Summarization

using Attentive Neural Techniques

Jacob Krantz
Aug 3, 2018

Why Summarize Automatically?



- News articles
- Tweets on Twitter
- Business meetings
- Microblogs
- Research papers
- Much more

Benefits

- Reduce reading time
- Remove bias
- Improve document selection efficiency

Method Overview: Extractive (Yao et al. 2017)

Process includes:

- Sentence Scoring
- Sentence Selection
- Sentence Reformulation

Techniques:

- HMMs
- CRFs
- Structural SVMs
- Integer Linear Programming (ILP)

Method Overview: Abstractive

1. Comprehend the meaning of the input text
2. Generate a summary using whatever words best fit the meaning

Paraphrase example:

The town was damaged by the cyclone. -> *The tornado hit the town.*

Project Goal

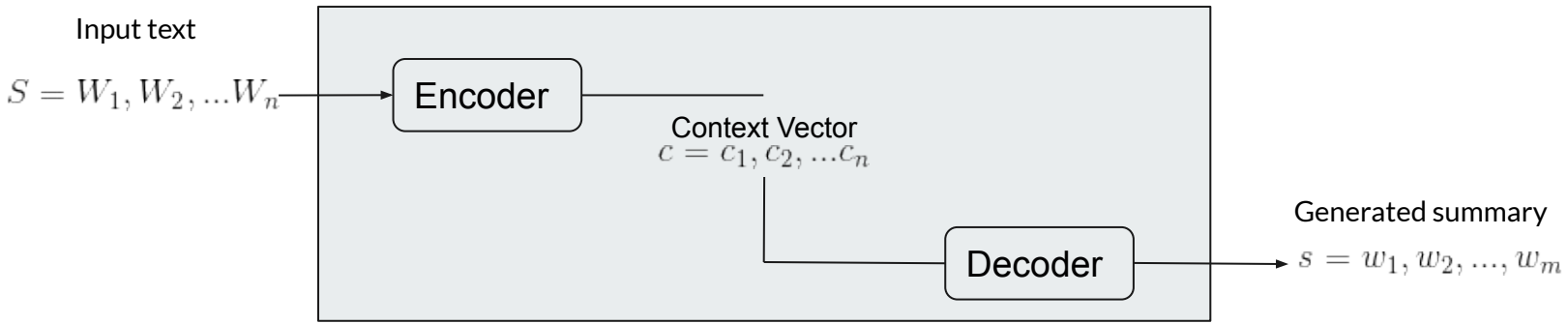
Create a novel neural network model that produces the best abstractive summaries of a single sentence.

Sentence	Angola is planning to refit its ageing soviet-era fleet of military jets in Russian factories, a media report said on Tuesday.
Summary	Angola to refit military fleet in Russia: report.

Example from DUC2004 taken from Zeng 2016

ANNs for Abstractive Summarization

Encoder-decoder networks are the modern architecture for developing summarization models.



End-to-end summary generation model

Related Work

- Koehn et al., 2007: a statistical machine translation model with confusion-matrix decoding.
  - Models: MOSES+
- Rush, et al. (2015): a standard **NNLM** with **attention** for encoding, **beam search** for decoding.
  - Models: ABS and ABS+
- Zeng et al. (2016): one **RNN** to reweight another with **attention**, **copy**, and **read-again** mechanisms.
  - Model: RA-C-LSTM
- Paulus et al. (2017): **RNN** model with **RL** and **teacher forcing**.
- Li et al. (2018): **RL** optimized directly on ROUGE for highest scores.
  - Model: AC-ABS

MT: The Transformer (Vaswani, et al. 2017)

In general:

- An encoder-decoder neural network built for machine translation.
- Uses multiheaded attention mechanisms in both the encoder and decoder networks.
- Includes positional encoding.

Important takeaway:

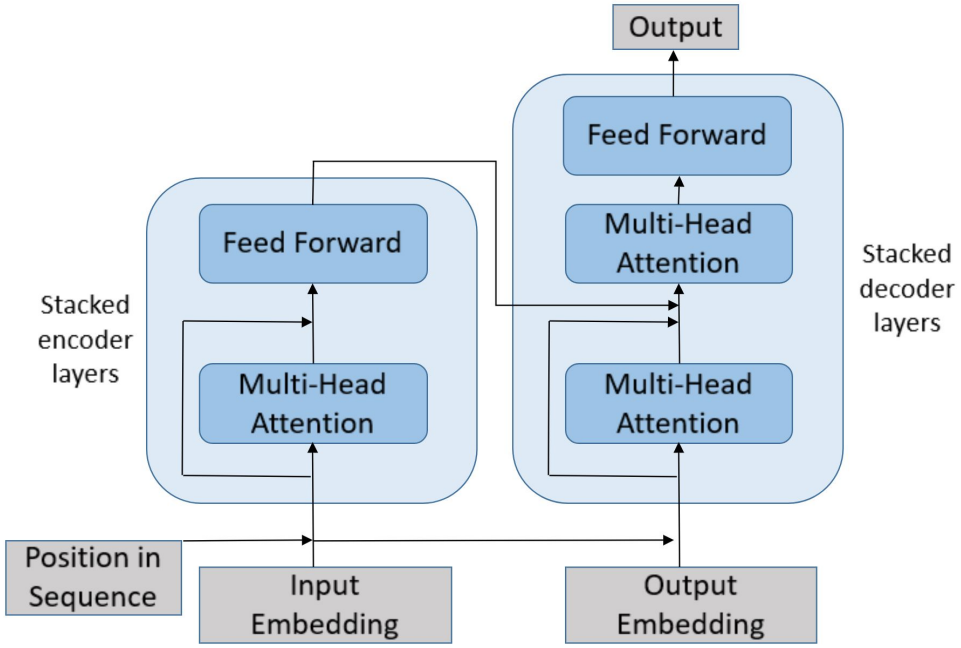
- Networks can solve sequence to sequence tasks using attention exclusively without the need for recurrence or convolutions.

Base Implementation

Attention-based encoder-decoder optimized for summarization. Involves positional encoding and standard feed forward layers.

Attention:

- Multiheaded
- self
- **scaled dot-product attention**

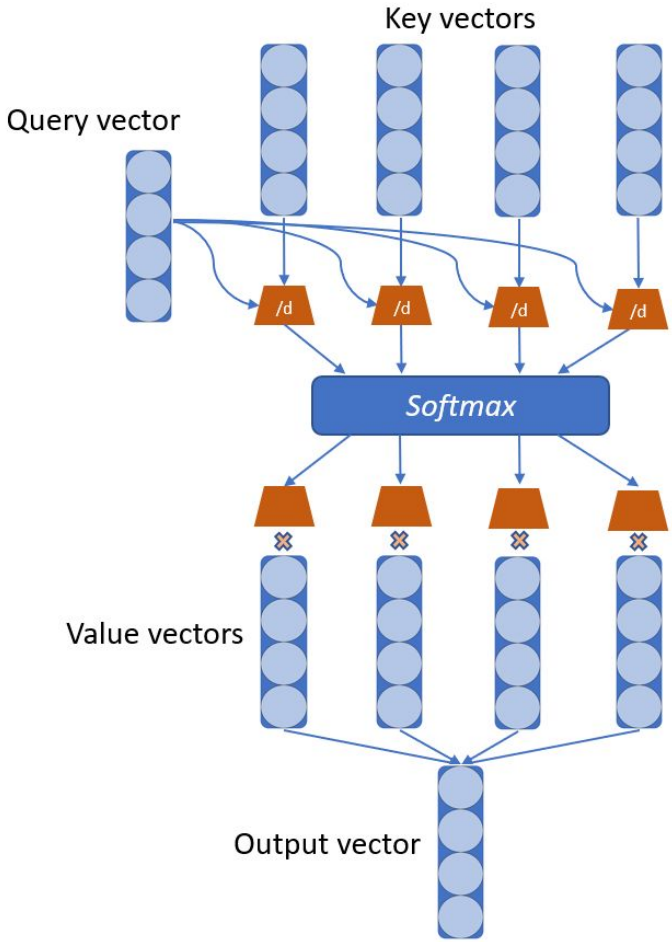


Attention Mechanisms

Scaled Dot-Product Attention

Maps a query (Q) to a given key-value pair (K,V).

$$attention = softmax(\frac{QK^T}{\sqrt{d}})V$$



Attention Mechanisms

Local Attention

Divides the key-value vectors into localized blocks. Each query position can see the corresponding block and past blocks.
Results of each block attention are merged.

Dilated Attention

Introduces gaps between the blocks of local attention. Query positions see a window of preceding and following blocks

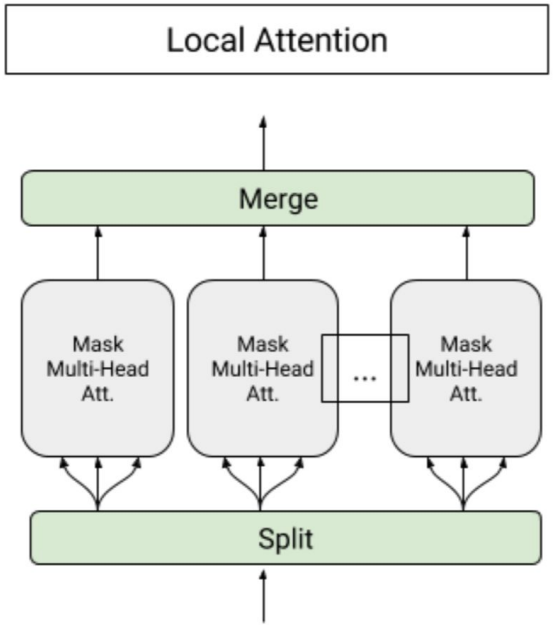


Image from Liu et al, 2018

Initial Results

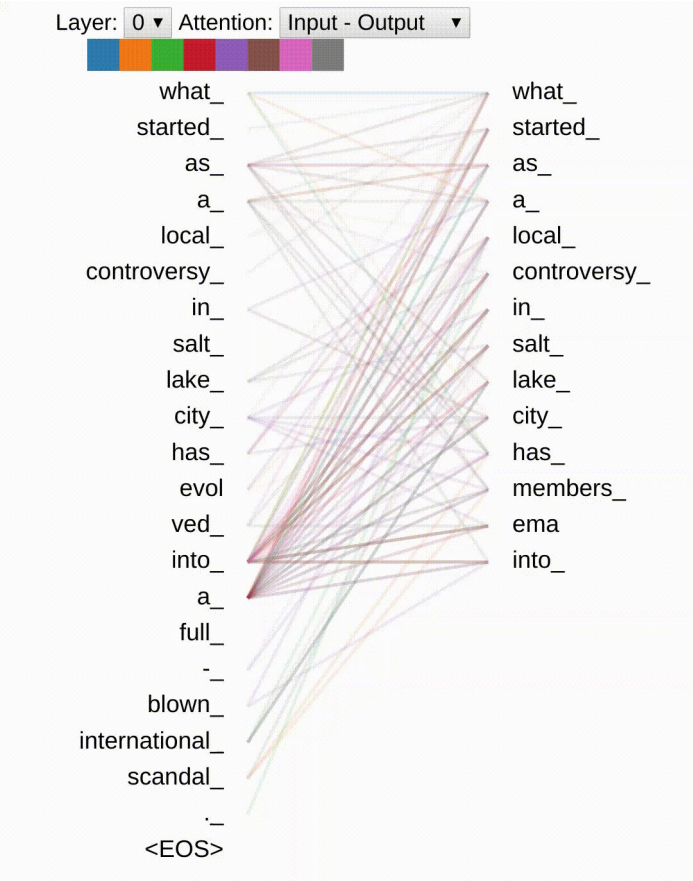
System	ROUGE-1	ROUGE-2	ROUGE-L
Transformer-Base	18.45	5.08	16.67
ABS+	28.18	8.49	23.81
RAS-Elman	28.97	8.26	24.06
AC-ABS	32.03	10.99	27.86

Input Sentence
the radical islamic group hamas on monday denounced u.s. president bill clinton 's upcoming visit to the gaza strip but carefully avoided making any threats against him .

Generated Summary
the radical islamic group hamas on monday denounced u.s. president bill

Attention Analysis

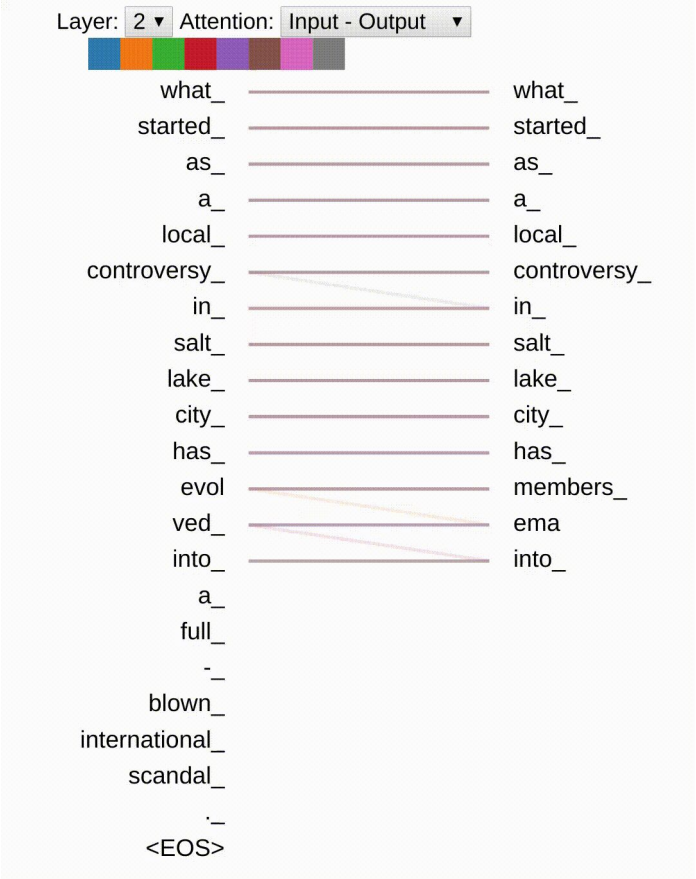
Layer 0 shows good diversity in what the decoder is attending over. Each color represents one attention head.



Attention Analysis

Layers 1-5 show poorly learned attention. Only the current position is considered from the input sequence.

1) Strong local minimum in training
2) Prepending inputs to targets



Training and Evaluation

Datasets

- Gigaword (3.8M)
- DUC2003 (625)
- DUC2004 (500)

Evaluation

- ROUGE scores
  - Recall-Oriented Understudy for Gisting Evaluation

Model Name	ROUGE-1	ROUGE-2	ROUGE-L
someModel1	Unigram %	Bigram %	LCS

Training and Evaluation

The Problem with ROUGE

There are many ways to say the same thing, but paraphrasing and synonymous concepts are not considered by ROUGE.

Generated summary:	technology companies win a case over copyright laws
Target summary:	Tech giants win a battle over copyright regulations

ROUGE-1	ROUGE-2
50.00	28.57

An abstractive summary requires an abstractive evaluation.

Training and Evaluation: VERT

New metric for evaluating automatic summaries: Versatile Evaluation of Reduced Texts (VERT)

- ✅ Semantic abstractions
- ✅ Works at the sentence level and at the word level

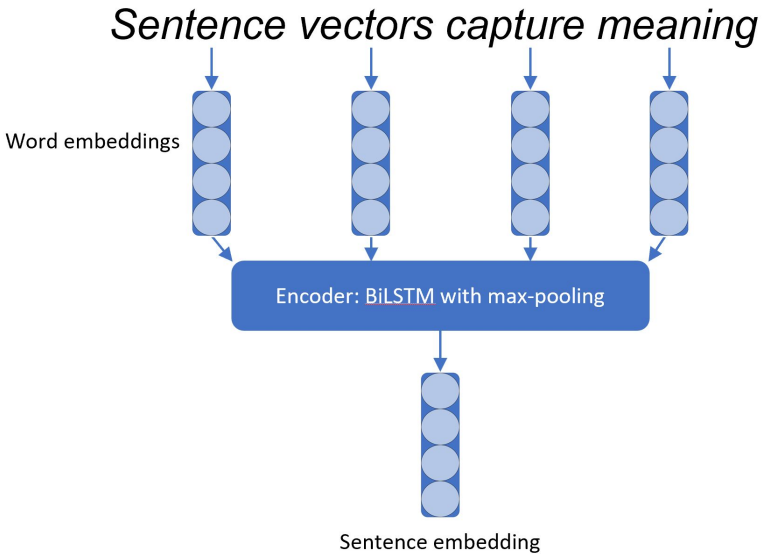
Target	*Endeavour astronauts join two segments of International Space Station.*
Gen1	Endeavour astronauts join two **sections** of International Space Station.
Gen2	Endeavour astronauts **remove** two segments of International Space Station.
Gen3	Endeavour astronauts join two segments of International Space Station.

Sentence	ROUGE-1	ROUGE-2	ROUGE-l	Cos-Sim	WMD	VERT
Gen1	88.89	75.00	88.89	0.979	0.418	94.77
Gen2	88.89	75.00	88.89	0.924	0.512	91.08
Gen3	100.00	100.00	100.00	1.000	0.000	100.00

Training and Evaluation: VERT

Similarity Sub-Score

Cosine similarity between sentence vectors generated by *InferSent* (Conneau, et al. 2017)



Training and Evaluation: VERT

Dissimilarity Sub-Score

Word Mover's Distance (Kusner, et al. 2015)

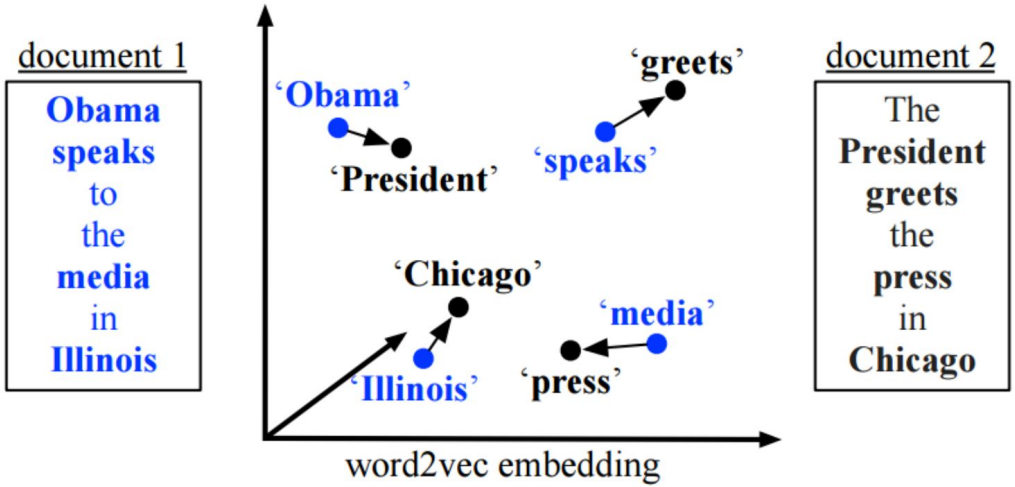


Image from Kusner, et al. 2015

Training and Evaluation: VERT

Similarity ranges [0,1] and *Dissimilarity* ranges [0, inf]

?

1. Want a range of [0,1]
2. Give equal weight to both sub-scores

$$VERT(s_1, s_2) = \begin{cases} 1.0 & \text{if } dis(s_1, s_2) \equiv 0.0 \\ \tanh(\frac{sim(s_1,s_2)}{[dis(s_1,s_2)]^k}) & otherwise \end{cases}$$

where

$$sim(s_1, s_2) = cos(encode(s_1), encode(s_2)),$$
$$dis(s_1, s_2) = wmd(s_1, s_2), \text{ and}$$
$$k = 1/3.$$

OR

$$VERT(s_1, s_2) = \frac{1}{2}(1 + (sim(s_1, s_2) - \frac{1}{\alpha} dis(s_1, s_2)))$$

Training and Evaluation: VERT

$$VERT(s_1, s_2) = \frac{1}{2}(1 + (sim(s_1, s_2) - \frac{1}{\alpha} dis(s_1, s_2)))$$

$$sim(s_1, s_2) = cos(encode(s_1), encode(s_2))$$

$$dis(s_1, s_2) = min(wmd(s_1, s_2), \alpha)$$

where $\alpha = 5.0$

WMD	Summary Count
$0 \rightarrow 1$	74
$1 \rightarrow 2$	860
$2 \rightarrow 3$	2858
$3 \rightarrow 4$	2150
$4 \rightarrow 5$	58
$5+$	0

Target summaries of DUC2004

Training and Evaluation: VERT

Human evaluation vs VERT?

Responsiveness assessment*:

50 generated summaries

Likert Scale:

1. Very Poor
2. Poor
3. Barely Acceptable
4. Good
5. Very Good

Metric	Pearson	P-Value
ROUGE-1	0.3039	0.0319
ROUGE-2	0.2577	0.0708
ROUGE-L	0.3071	0.0300
VERT	0.3681	0.0085

*https://duc.nist.gov/duc2007/responsiveness.assessment.instructions

Model Optimizations

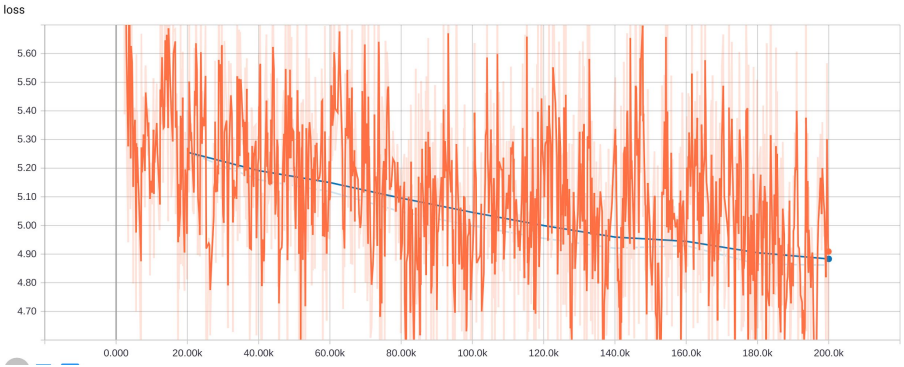
Decoding Problem:

The target summaries of Gigaword average 8 words long, but the target summaries of DUC2004 average 11.5 words long.

Solution:

1. Beam search decoding: beam size of 8
2. Alpha decoding parameter: control the change of generating <EOS>
3. Set a fixed token generation limit: 14 words

Convergence Problem:



Solution: delay gradient updates

Comparison of Attention Mechanisms

Mechanism	RG-1	RG-2	RG-L	VERT-S	VERT-D	VERT
s-dot-prod	25.72	8.51	23.08	0.73523	2.76307	59.13
rel-s-dot-prod	**27.05**	**9.54**	**24.44**	**0.73876**	**2.73907**	**59.55**
local	1.93	0.00	1.93	0.02084	5.00000	1.04
local-mask	25.72	8.54	23.30	0.73361	2.77857	58.89
local-blk-mask	14.13	2.75	12.63	0.67226	3.18881	51.73
dilated	0.01	0.00	0.01	0.09509	3.66543	18.10
dilated-mask	19.06	5.23	17.45	0.68682	3.04922	53.85

Trained on Gigaword. Tested on DUC2004. Each trained with 25000 steps.

Generated Examples

S(1): exxon corp. and mobil corp. have held discussions about combining their business operations , a person involved in the talks said wednesday .
Target: exxon corp. and mobil corp. may combine business operations
S-ATT-REL: exxon and mobil discuss merger

S(2): prime minister rafik hariri , the business tycoon who launched lebanon 's multibillion dollar reconstruction from the devastation of civil war , said monday he was bowing out as premier following a dispute with the new president .
Target: prime minister hariri , claiming constitution violation , bows out
S-ATT-REL: lebanese prime minister resigns after dispute with new president

S(3): organizers of december 's asian games have dismissed press reports that a sports complex would not be completed on time , saying preparations are well in hand , a local newspaper said friday .
Target: bangkok says sports complex will be completed in time for asian games
S-ATT-REL: asian games organizers say sports complex will not be completed on time

S(4): a struggle for control of the house is under way , with rep. robert livingston conducting a telephone campaign that could lead to him running against newt gingrich as speaker .
Target: election of gingrich as house speaker in doubt as small group opposes him
S-ATT-REL: house speaker 's phone campaign could lead to gingrich

S(5): premier romano prodi battled tuesday for any votes freed up from a split in a far-left party , but said he will resign if he loses a confidence vote expected later this week .
Target: italian premier to resign if he loses pending confidence vote
S-ATT-REL: italy 's prodi says he will resign if he loses confidence vote

Comparison to Published Approaches

Model	RG-1	RG-2	RG-L	VERT
TOPIARY (Zajic, Dorr, and Schwartz 2004)	25.12	6.46	20.12	-
ABS (Rush, Chopra, and Weston 2015)	26.55	7.06	22.05	58.49
RAS-LSTM (Chopra, Auli, and Rush 2016)	27.41	7.69	23.06	-
MOSES+ (Koehn et al. 2007)	26.50	8.13	22.85	-
RAS-Elman (Chopra, Auli, and Rush 2016)	28.97	8.26	24.06	-
ABS+ (Rush, Chopra, and Weston 2015)	28.18	8.49	23.81	59.05
RA-C-LSTM (Zeng et al. 2016)	29.89	9.37	25.93	-
words-lvt5k-1sen (Nallapati et al. 2016)	28.61	9.42	25.24	-
S-ATT-REL (ours)	27.05	9.54	24.44	59.55
AC-ABS (Li, Bing, and Lam 2018)	32.03	10.99	27.86	-

Conclusion

Research questions answered:

- Can a self-attentive network be modified to perform sentence summarization?
  - Yes
- What is the effect of various attention mechanisms on summarization performance?
  - Relative dot-product self-attention performed the best
  - Local and dilated self-attention should be masked
- Is there a better way to judge abstractive summaries than ROUGE?
  - Proposed VERT

Acknowledgements

Advisor: Dr. Jugal Kalita

University of Colorado, Colorado Springs

Work supported by: NSF Grant No. 1659788

Citations

- Conneau, Alexis, et al. "Supervised Learning of Universal Sentence Representations from Natural Language Inference Data." *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*. 2017.
- Filippova, Katja, and Yasemin Altun. "Overcoming the lack of parallel data in sentence compression." *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 2013.
- Kusner, Matt, et al. "From word embeddings to document distances." *International Conference on Machine Learning*. 2015.
- Nallapati, Ramesh, et al. "Abstractive Text Summarization using Sequence-to-sequence RNNs and Beyond." *CoNLL 2016* (2016): 280.
- Paulus, Romain, Caiming Xiong, and Richard Socher. "A deep reinforced model for abstractive summarization." *arXiv preprint arXiv:1705.04304* (2017).
- Rush, Alexander M., Sumit Chopra, and Jason Weston. "A Neural Attention Model for Abstractive Sentence Summarization." *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. 2015.
- Vaswani, Ashish, et al. "Attention is all you need." *Advances in Neural Information Processing Systems*. 2017.
- Yao, Jin-ge, Xiaojun Wan, and Jianguo Xiao. "Recent advances in document summarization." *Knowledge and Information Systems* 53.2 (2017): 297-336.
- Zeng, Wenyuan, et al. "Efficient summarization with read-again and copy mechanism." *arXiv preprint arXiv:1611.03382*(2016).