Headline Generation Track

at Dialogue’2019

Organizers

VK.com:

- Valentin Malykh, Pavel Kalaidin
- Ivan Karabakin, Irina Shubina

vk.com/deepvk

With the help of:

- Ivan Smurov, ABBYY
- Ekaterina Artemova, HSE

Summarization Task

- Sentence Summarization
  - to produce more concise sentences

- Text Summarization
  - to produce shorter texts

Summarization Task

- Extractive Summarization
  - to take some phrases from a text

- Abstractive Summarization
  - to generate a new text basing on bigger one

Extractive Summarization

- To take some phrases from a text

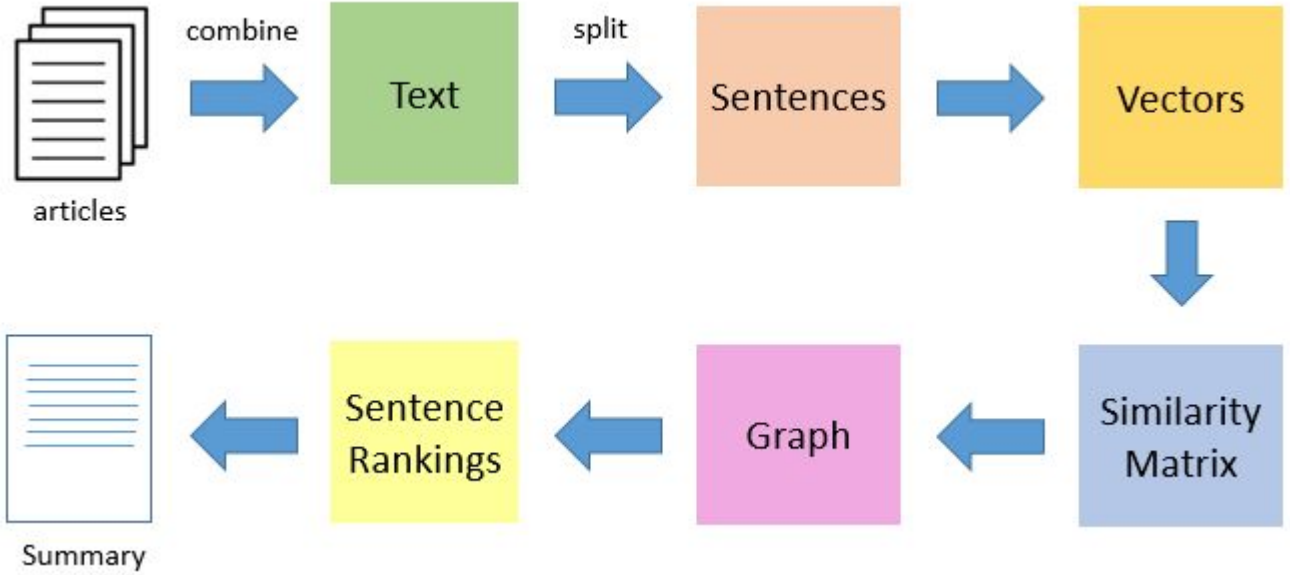
Supervised and Unsupervised:

- We have some gold markup of taken phrases.
- And there are no such markup.

Common Approaches to Ext. Sum.

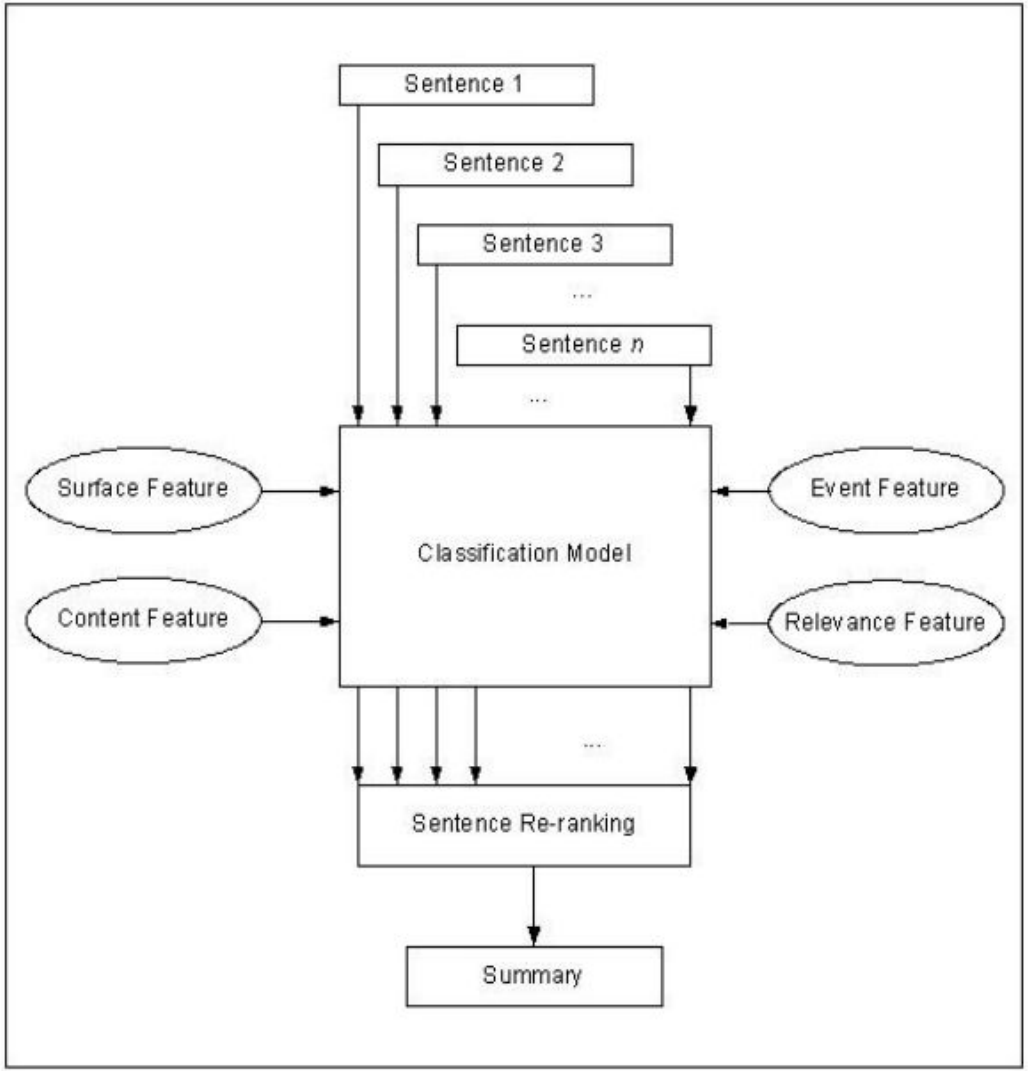
TextRank

LexRank



Common Approaches

Supervised Approaches



Wong KF, Wu M, Li W. Extractive summarization using supervised and semi-supervised learning. In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 2008 Aug 18 (pp. 985-992). Association for Computational Linguistics.

Abstractive Summarization

The bigger test is paraphrased to smaller one.

It is common, that bigger (original) and smaller (summary) texts are human-generated.

Supervised and Unsupervised

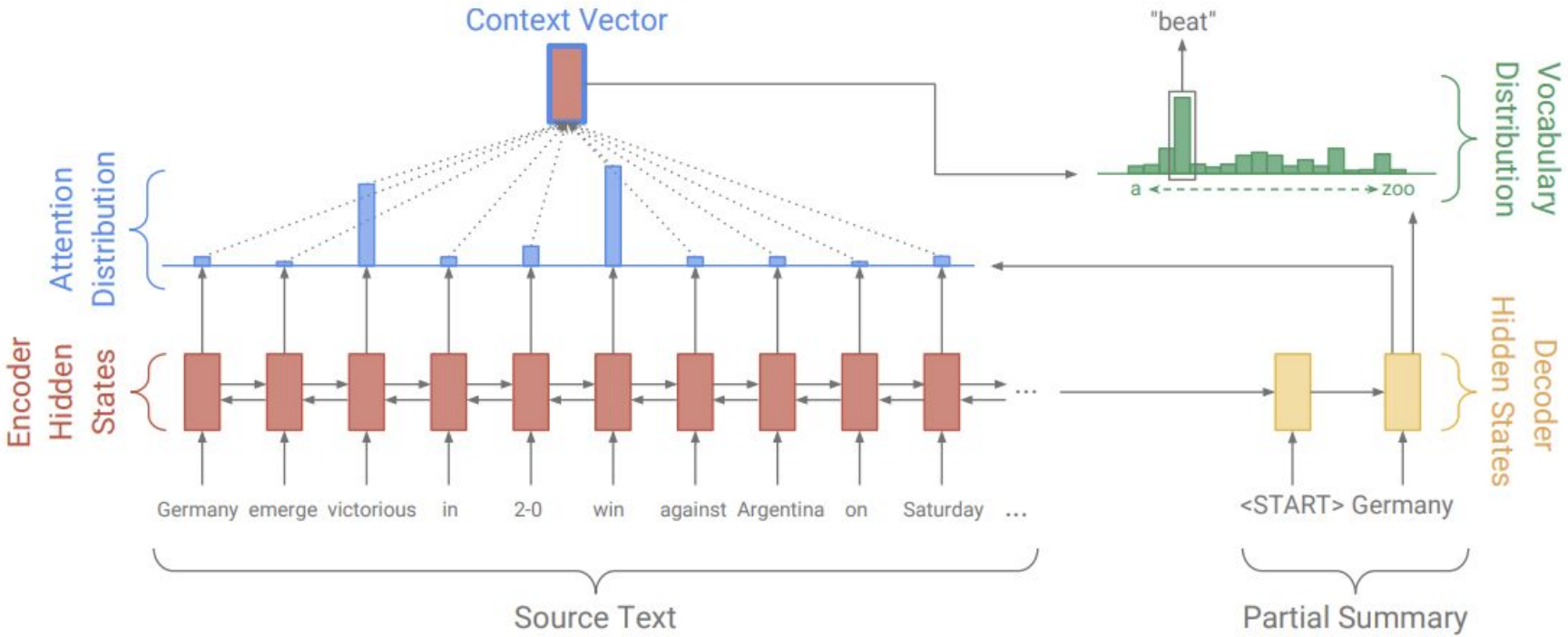
Direct Approaches to Abs. Sum.

Common approaches:

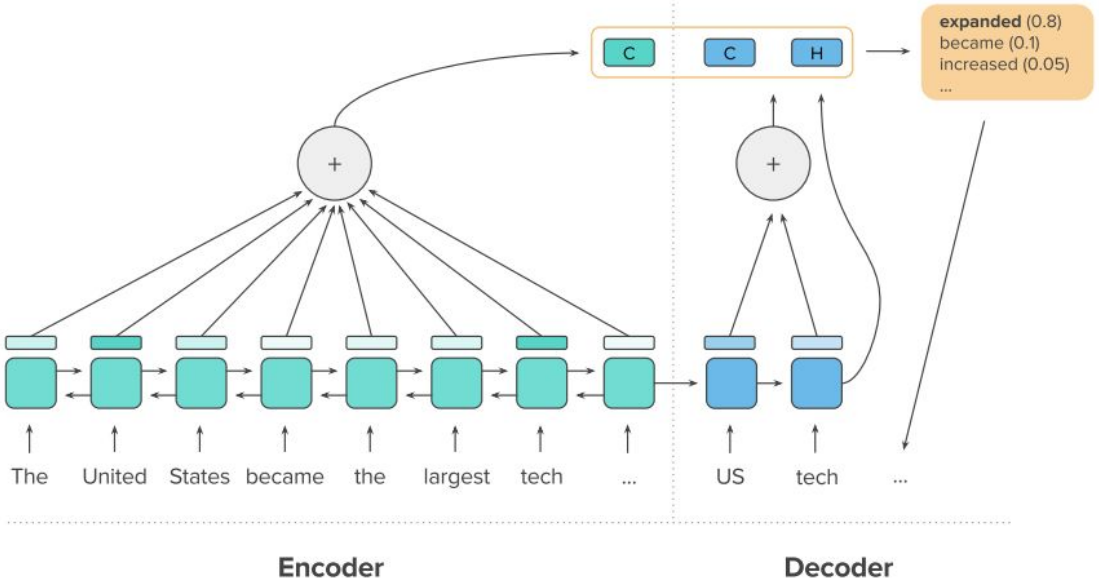
- BiRNN
- CNN
- Transformer
- Pointer-Generator

- Reinforcement Learning

Direct Approaches: Pointer-Generation

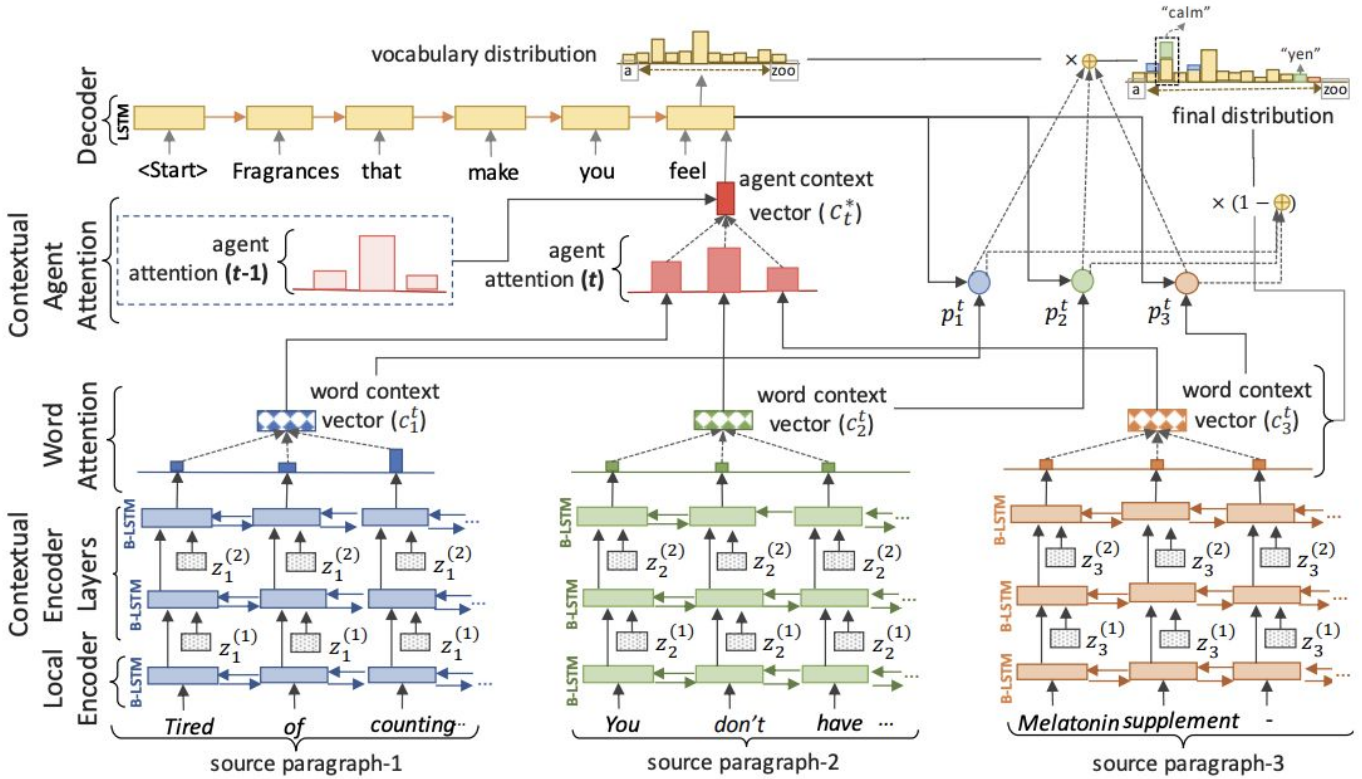


Direct Approaches: Reinforcement Learning

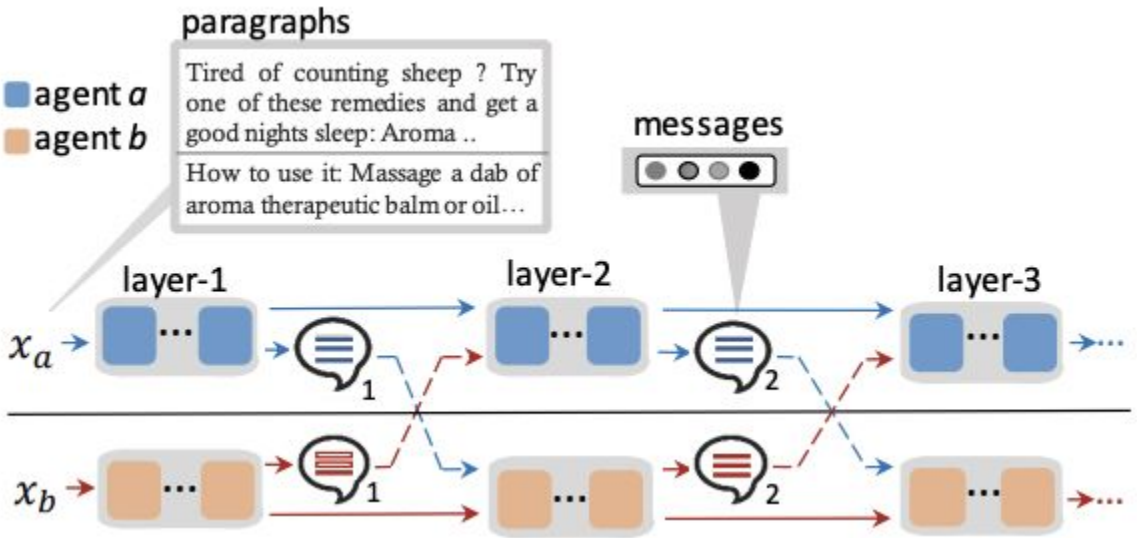


$$L_{rl} = (r(\hat{y}) - r(y^s)) \sum_{t=1}^{n'} \log p(y_t^s | y_1^s, \ldots, y_{t-1}^s, x)$$

Direct Approaches: Multi-Agent Sum.



Direct Approaches: Multi-Agent Sum.



Direct Approaches: Results

Model	ROUGE-1	ROUGE-2	ROUGE-L
SummaRuNNer (Nallapati et al., 2017)	39.60	16.20	35.30
graph-based attention (Tan et al., 2017)	38.01	13.90	34.00
pointer generator (See et al., 2017)	36.44	15.66	33.42
pointer generator + coverage (See et al., 2017)	39.53	17.28	36.38
controlled summarization with fixed values (Fan et al., 2017)	39.75	17.29	36.54
RL, with intra-attention (Paulus et al., 2018)	41.16	15.75	**39.08**
ML+RL, with intra-attention(Paulus et al., 2018)	39.87	15.82	36.90
(**m1**) MLE, pgen, no-comm (1-agent) (our baseline-1)	36.12	14.38	33.83
(**m2**) MLE+SEM, pgen, no-comm (1-agent) (our baseline-2)	36.90	15.02	33.00
(**m3**) MLE+RL, pgen, no-comm (1-agent) (our baseline-3)	38.01	16.43	35.49
(**m4**) DCA MLE+SEM, pgen, no-comm (3-agents)	37.45	15.90	34.56
(**m5**) DCA MLE+SEM, *mpgen*, with-comm (3-agents)	39.52	17.12	36.90
(**m6**) DCA MLE+SEM, *mpgen*, with-comm, with *caa* (3-agents)	41.11	**18.21**	36.03
(**m7**) DCA MLE+SEM+RL, *mpgen*, with-comm, with *caa* (3-agents)	**41.69**	**19.47**	37.92

Table 1: Comparison results on the **CNN/Daily Mail** test set using the **F1** variants of **Rouge**. Best model models are bolded.

Model	Rouge-1	Rouge-2	Rouge-L
ML, no intra-attention (Paulus et al., 2018)	44.26	27.43	40.41
RL, no intra-attention (Paulus et al., 2018)	47.22	30.51	**43.27**
ML+RL, no intra-attention(Paulus et al., 2018)	47.03	30.72	43.10
(**m1**) MLE, pgen, no-comm (1-agent) (our baseline-1)	44.28	26.01	37.87
(**m2**) MLE+SEM, pgen, no-comm (1-agent) (our baseline-2)	44.50	28.04	38.80
(**m3**) MLE+RL, pgen, no-comm (1-agent) (our baseline-3)	46.15	29.50	39.38
(**m4**) DCA MLE+SEM, pgen, no-comm (3-agents)	45.84	28.23	39.32
(**m5**) DCA MLE+SEM, *mpgen*, with-comm (3-agents)	46.20	30.01	40.65
(**m6**) DCA MLE+SEM, *mpgen*, with-comm, with *caa* (3-agents)	**47.30**	30.50	41.06
(**m7**) DCA MLE+SEM+RL, *mpgen* with-comm, with *caa* (3-agents)	**48.08**	**31.19**	42.33

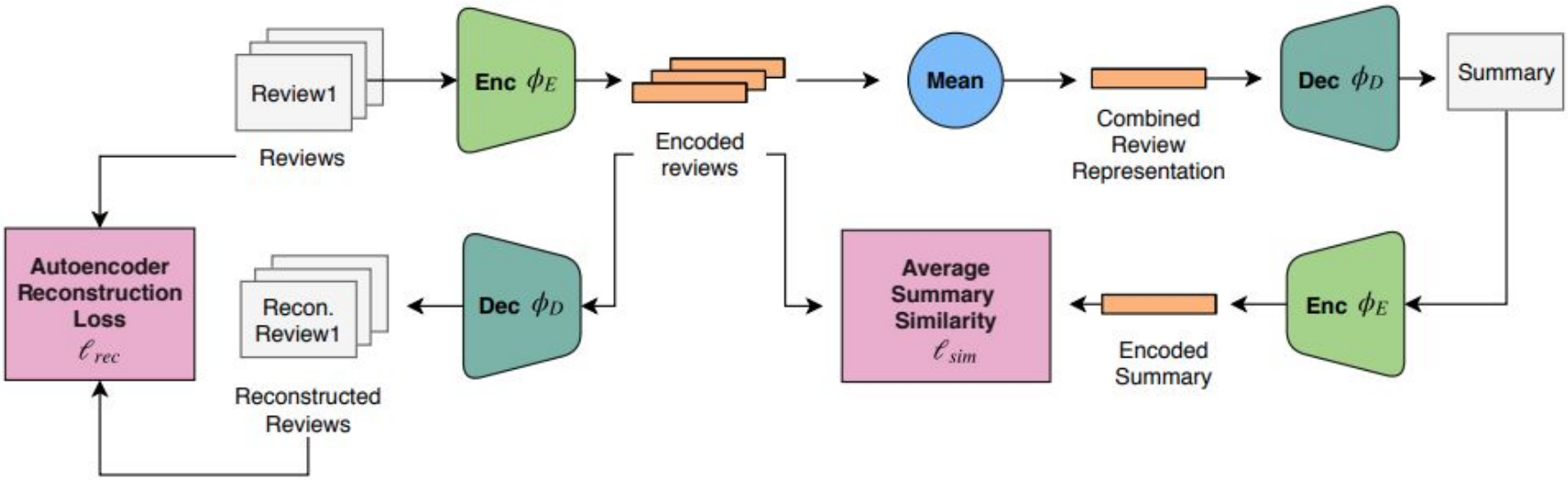
Table 2: Comparison results on the **New York Times** test set using the **F1** variants of **Rouge**. Best model models are bolded.

Common Approaches to Abs. Sum.

Indirect approaches:

- 5W1H
- First Sentence
- Topic-sentence
- Unsupervised Extraction Summary-based
- etc.

Unsupervised Abstractive Summarization



Datasets

- DUC 2001-2007
  - hundreds of documents each
- CNN / Dailymail
  - 287226 articles for training, 13368 for validation, and 11490 for test
  - 781 token on average for article, 56 tokens for a summary

summary

- New York Times Annotated
  - 1444919 articles
  - 708 tokens for an article, 8 tokens for a headline
- Rossiya Segodnya News
  - 1003869 articles
  - 316 tokens for an article, 10 tokens for a headline

headline

Track Datasets

- Rossiya Segodnya News Dataset
  - 1m of news documents
  - 1 news agency

- ROMIP News Collection
  - 32k of news documents
  - 16k has been used to compute public score, and the rest - the private one
  - 25 different news agencies

Metrics: METEOR

- Weighted combination of the unigram precision and recall
- A fragmentation penalty to address fluency
  - A “chunk” is a monotonic sequence of aligned words
- Final Score

$$F_{mean} = \frac{P \cdot R}{(\alpha \cdot P + (1-\alpha) \cdot R)}$$

$$frag = \frac{(no.\ of\ chunks)}{(no.\ of\ unigram\ matches)}$$

$$score = F_{mean} \times (1 - \gamma \times frag^{\beta})$$

Metrics: ROUGE

$$\text{ROUGE-N} = \frac{\sum_{S \in \{ReferemceSummaries\}} \sum_{gram_n \in S} Count_{match}(gram_n)}{\sum_{S \in \{ReferenceSummaries\}} \sum_{gram_n \in S} Count(gram_n)}$$

$$R_{lcs} = \frac{LCS(X,Y)}{m}$$

$$P_{lcs} = \frac{LCS(X,Y)}{n}$$

$$F_{lcs} = \frac{(1+\beta^2) R_{lcs} P_{lcs}}{R_{lcs} + \beta^2 P_{lcs}}$$

Track Metric

There are 9 different variants of ROUGE. We take F-score ones and mean them.

$$score(r, h) = \frac{1}{3N} \sum_{i=1}^{N} (\text{ROUGE-1}(r_i, h_i) + \text{ROUGE-2}(r_i, h_i) + \text{ROUGE-L}(r_i, h_i))$$

Platform

- Docker
- 1 GPU
- 16 Gb RAM
- 2 vCPU
- private docker registry

A solution has been run on private test set of size 16k.

Competition Statistics

- 15 registered participants
- 6 participants who made at least 1 submit
- 258 submits in total (~100 testing submits)
- 3 participants who beat the baseline

Results

Team	Score
Black and Yellow	**23.142**
Burning Headlines	20.293
Symmetrical potato	20.268
DreamTeam	20.267
L&M	20.267
Зульфат Мифтахутдинов	20.267

References

Paulus, R., Xiong, C. and Socher, R., 2017. A deep reinforced model for abstractive summarization. arXiv preprint arXiv:1705.04304.

Wong KF, Wu M, Li W. Extractive summarization using supervised and semi-supervised learning. In Proceedings of the 22nd International Conference on Computational Linguistics - Volume 1 2008 Aug 18 (pp. 985-992). Association for Computational Linguistics.

See, A., Liu, P. J., & Manning, C. D. (2017). Get to the point: Summarization with pointer-generator networks. arXiv preprint arXiv:1704.04368.

Chu, E., & Liu, P. J. (2018). Unsupervised Neural Multi-Document Abstractive Summarization of Reviews.

Mihalcea, R. and Tarau, P., 2004. Textrank: Bringing order into text. In Proceedings of the 2004 conference on empirical methods in natural language processing.

Celikyilmaz, A., Bosselut, A., He, X., & Choi, Y. (2018). Deep communicating agents for abstractive summarization. arXiv preprint arXiv:1803.10357.

Grusky, M., Naaman, M., & Artzi, Y. (2018). Newsroom: A dataset of 1.3 million summaries with diverse extractive strategies. arXiv preprint arXiv:1804.11283.

H.T. Dang. Overview of DUC 2006. National Institute of Standards and Technology (NIST)

Thank you for your attention!

Applied Research @ VK.com vk.com/deepvk
Valentin Malykh, val.maly.hk