Summary Construction Strategies for Headline Generation in the Russian Language

Valentin Malykh
valentin.malykh@phystech.edu

Kazan Federal University

MIPT
MOSCOW INSTITUTE OF PHYSICS AND TECHNOLOGY

Authors

- Valentin Malykh, PhD, Kazan Federal University
- Daniil Chernyavskiy, Moscow Institute of Physics and Technology
- Alex Valyukov, Moscow Institute of Physics and Technology

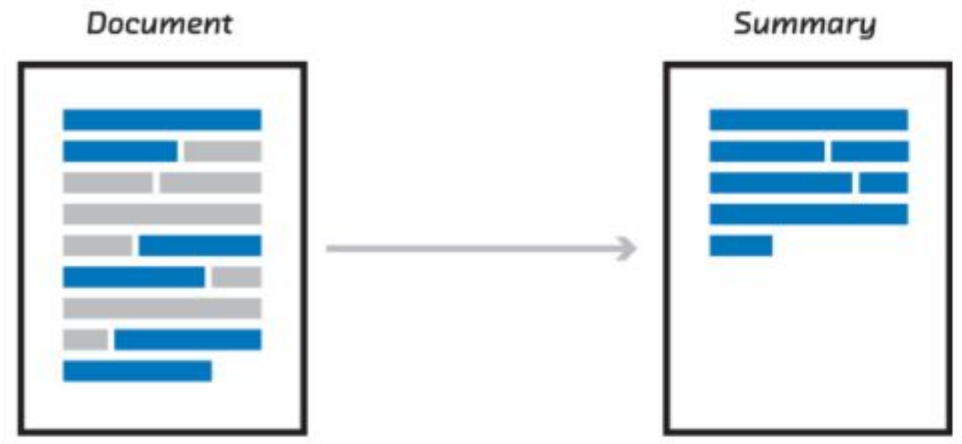
Acknowledgments

This work of the first author was funded by RFBR, project number 19-37-60027.

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

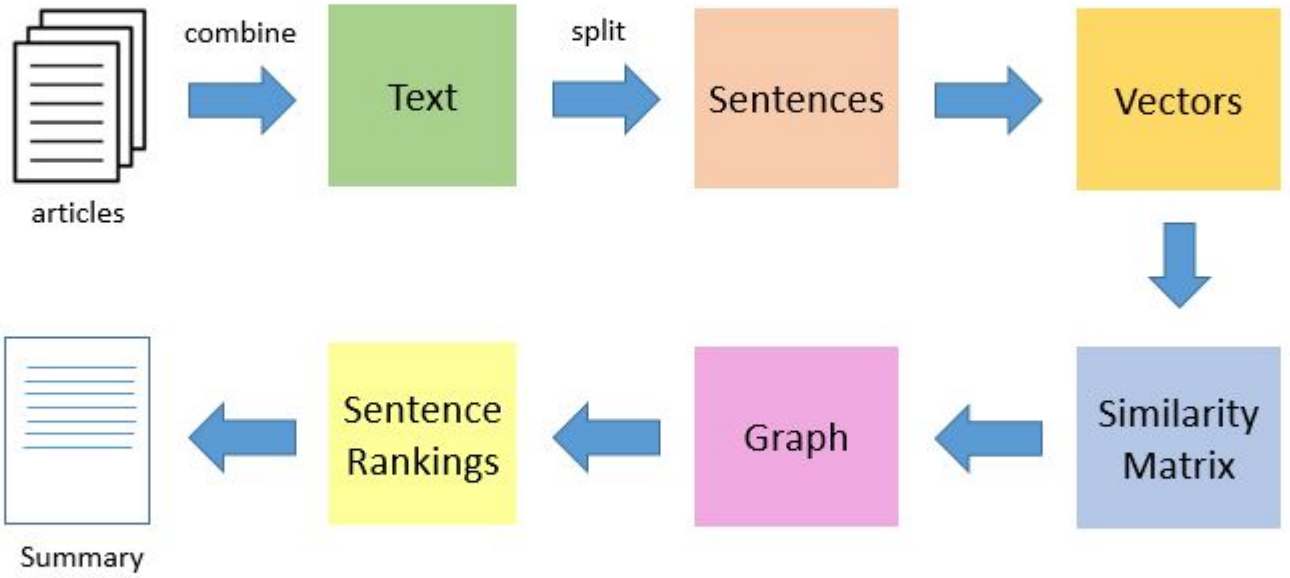
Supervised and Unsupervised:

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

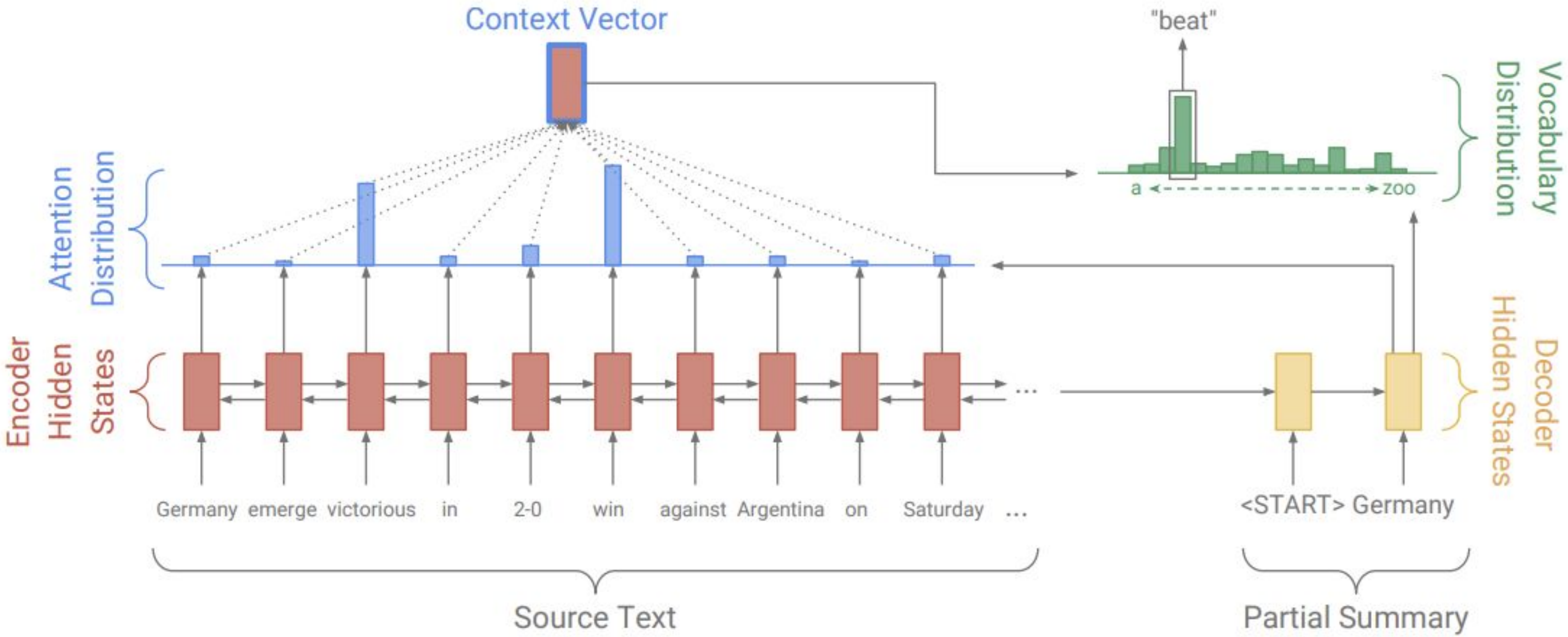
Common Approaches to Ext. Sum.

TextRank

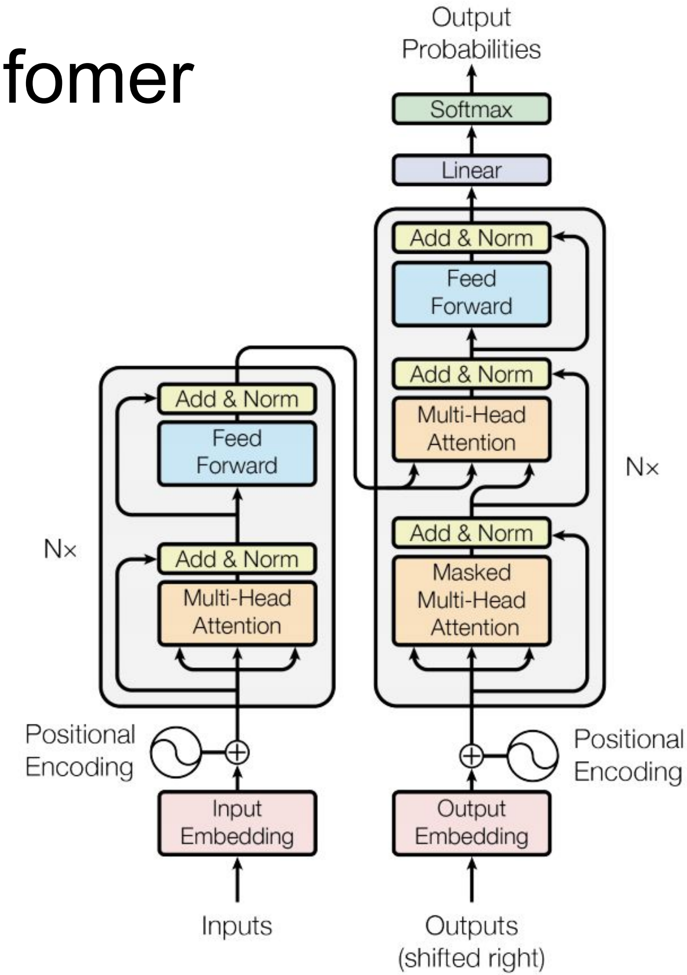
LexRank



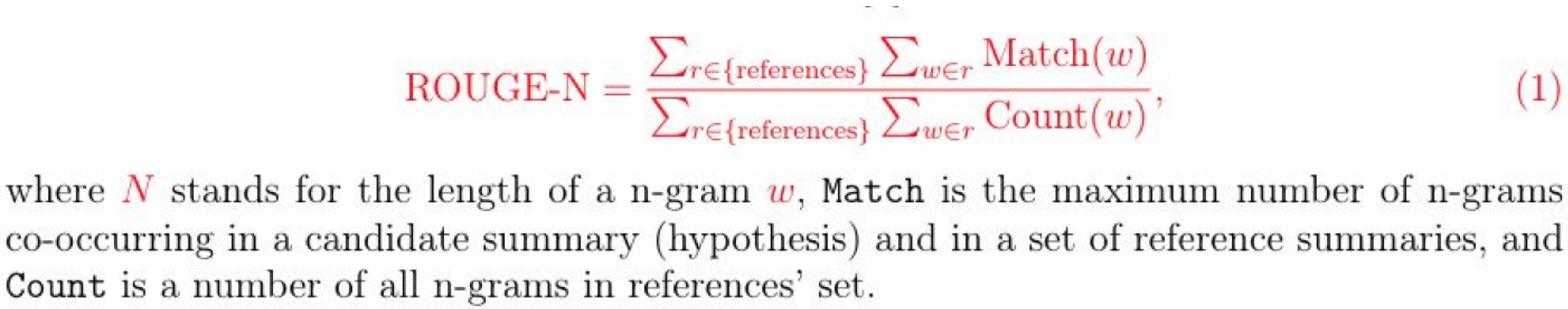
Baseline: RNN



Baseline: Transfomer



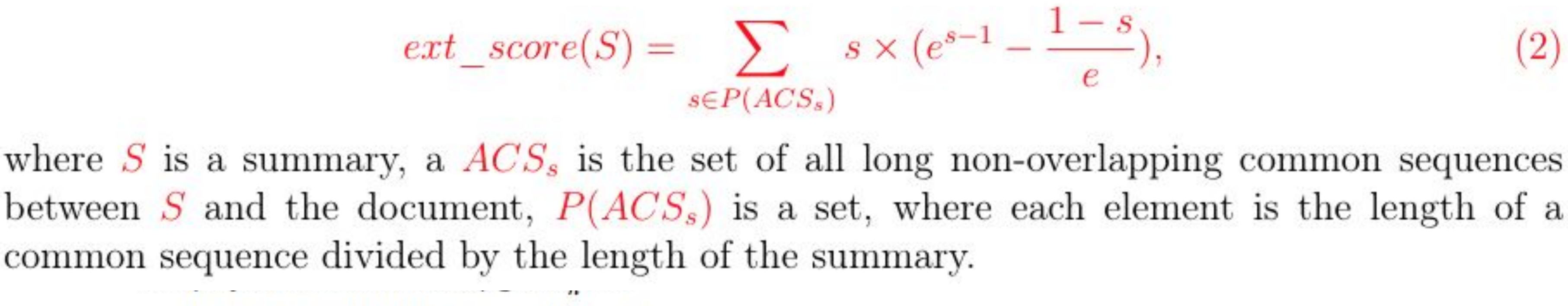
Metrics: ROUGE



$$\text{ROUGE-N} = \frac{\sum_{r \in \{\text{references}\}} \sum_{w \in r} \text{Match}(w)}{\sum_{r \in \{\text{references}\}} \sum_{w \in r} \text{Count}(w)}, \tag{1}$$

where N stands for the length of a n-gram w, `Match` is the maximum number of n-grams co-occurring in a candidate summary (hypothesis) and in a set of reference summaries, and `Count` is a number of all n-grams in references' set.

Metrics: Extraction score

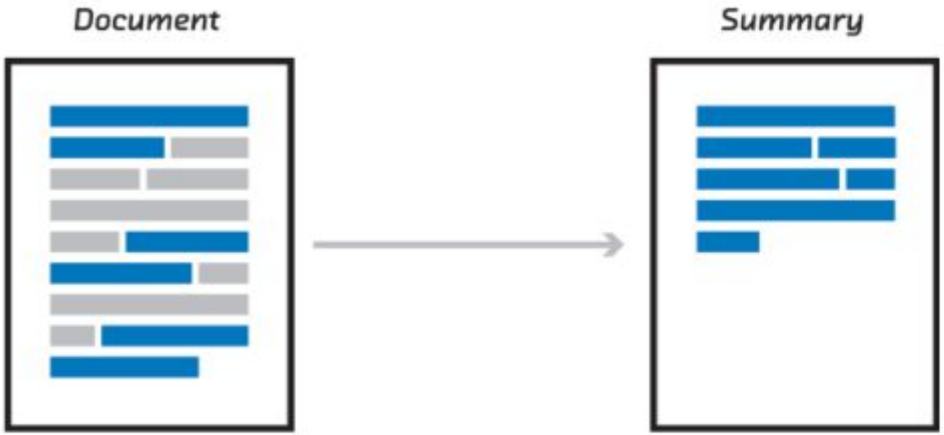


$$ext\_score(S) = \sum_{s \in P(ACS_s)} s \times (e^{s-1} - \frac{1-s}{e}), \tag{2}$$

where S is a summary, a ACS_s is the set of all long non-overlapping common sequences between S and the document, $P(ACS_s)$ is a set, where each element is the length of a common sequence divided by the length of the summary.

Our Hypothesis

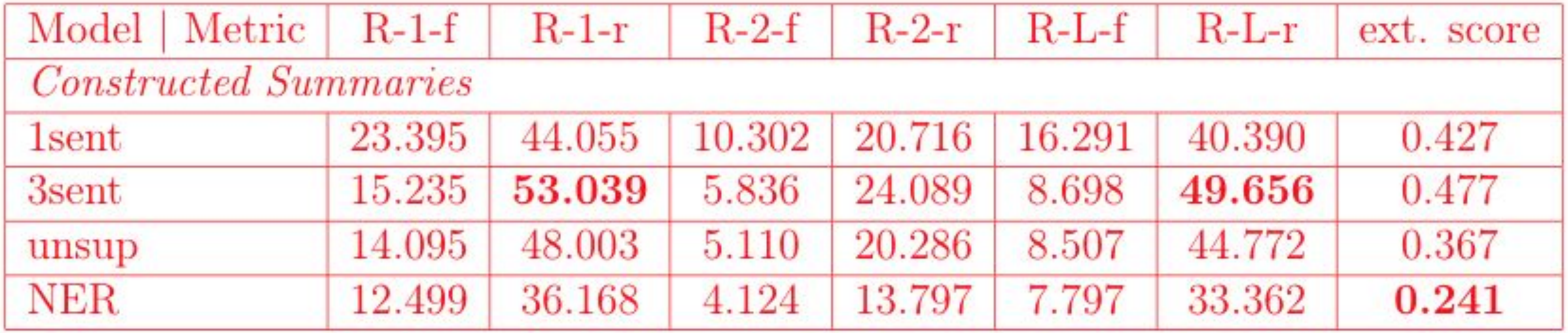
- We could improve the abstractive summarization results using the extractive summarization output.



Summary Construction Strategies

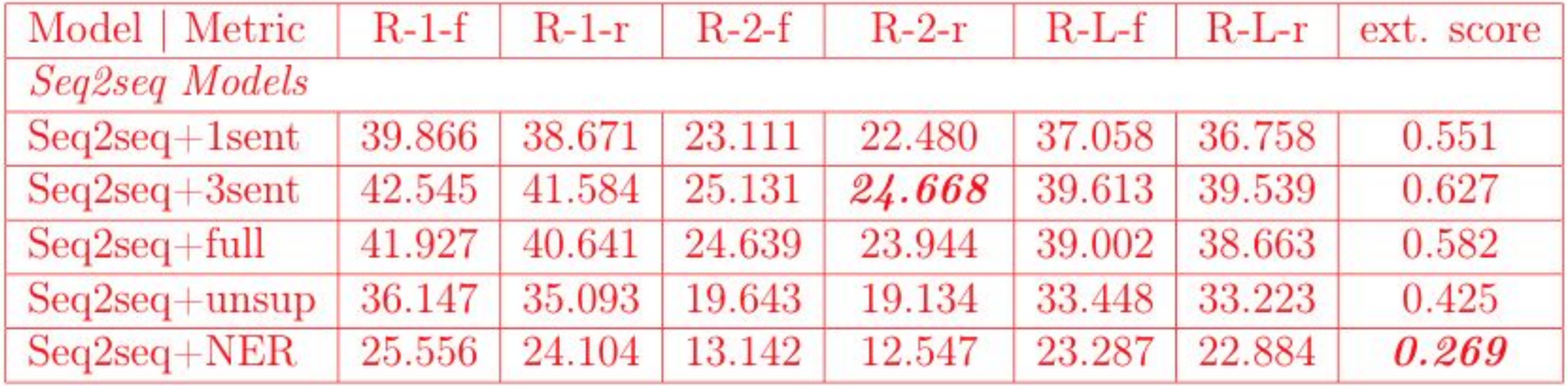
- Whole body
- 1st sentence
- Three first sentences
- Unsupervised summary
- NER Summary

Results: Summaries



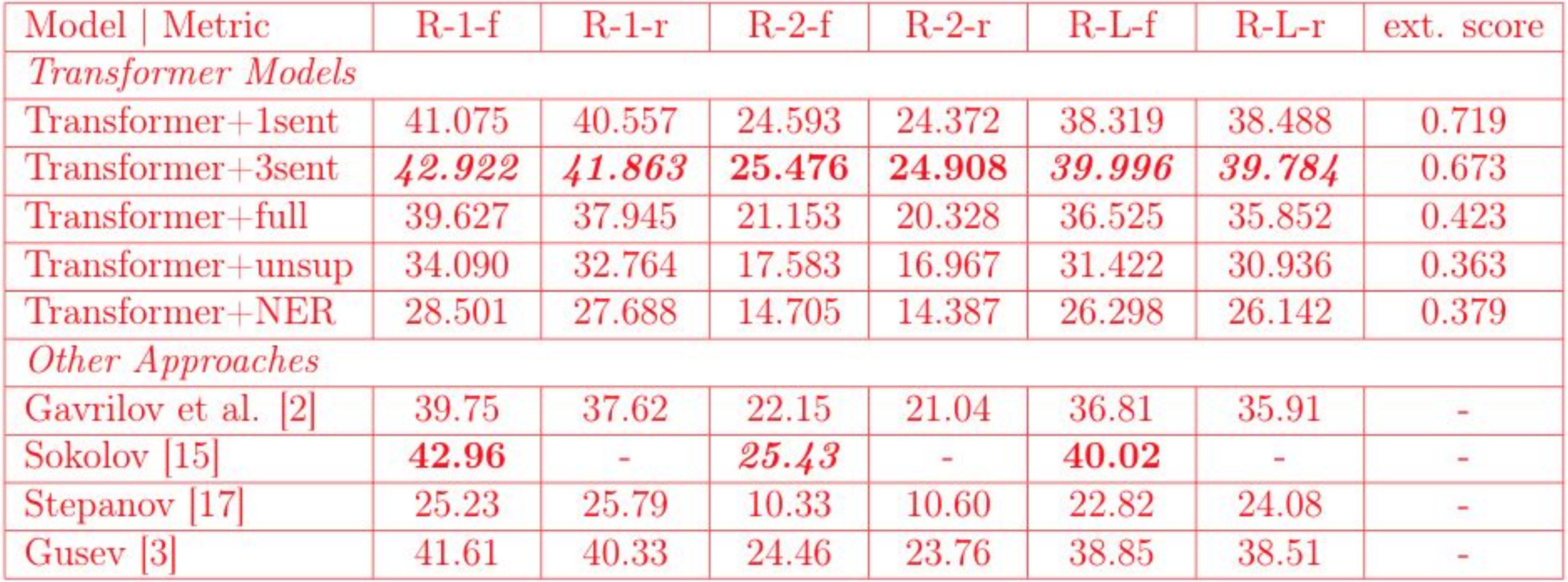
Model | Metric	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r	ext. score
Constructed Summaries							
1sent	23.395	44.055	10.302	20.716	16.291	40.390	0.427
3sent	15.235	**53.039**	5.836	24.089	8.698	**49.656**	0.477
unsup	14.095	48.003	5.110	20.286	8.507	44.772	0.367
NER	12.499	36.168	4.124	13.797	7.797	33.362	**0.241**

Results: RNN



Model | Metric	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r	ext. score
Seq2seq Models							
Seq2seq+1sent	39.866	38.671	23.111	22.480	37.058	36.758	0.551
Seq2seq+3sent	42.545	41.584	25.131	***24.668***	39.613	39.539	0.627
Seq2seq+full	41.927	40.641	24.639	23.944	39.002	38.663	0.582
Seq2seq+unsup	36.147	35.093	19.643	19.134	33.448	33.223	0.425
Seq2seq+NER	25.556	24.104	13.142	12.547	23.287	22.884	***0.269***

Results: Transformers



Model | Metric	R-1-f	R-1-r	R-2-f	R-2-r	R-L-f	R-L-r	ext. score
Transformer Models							
Transformer+1sent	41.075	40.557	24.593	24.372	38.319	38.488	0.719
Transformer+3sent	***42.922***	***41.863***	**25.476**	**24.908**	***39.996***	***39.784***	0.673
Transformer+full	39.627	37.945	21.153	20.328	36.525	35.852	0.423
Transformer+unsup	34.090	32.764	17.583	16.967	31.422	30.936	0.363
Transformer+NER	28.501	27.688	14.705	14.387	26.298	26.142	0.379
Other Approaches							
Gavrilov et al. [2]	39.75	37.62	22.15	21.04	36.81	35.91	-
Sokolov [15]	**42.96**	-	***25.43***	-	**40.02**	-	-
Stepanov [17]	25.23	25.79	10.33	10.60	22.82	24.08	-
Gusev [3]	41.61	40.33	24.46	23.76	38.85	38.51	-

Original text, truncated:	пожар, произошедший в среду в ресторане в центре москвы, ликвидирован, пострадавших нет, сообщил риа новости источник в правоохранительных органах столицы. "пожар в ресторане "эль гаучо" на садовой-триумфальной улице в двухэтажном здании ликвидирован. по предварительным данным, горели жировые отложения в вентиляции. возгорание произошло в вентиляционной системе", - сказал собеседник агентства. в настоящее время причины пожара устанавливаются. по данным представителя мчс, сообщение о пожаре поступило на пульт дежурного "01" в 21.25 мск. он отметил, что, благодаря своевременной эвакуации, никто из посетителей и сотрудников ресторана не пострадал.
Original headline:	пожар в ресторане в центре москвы ликвидирован, никто не пострадал
Transformer+1sent:	пожар в ресторане в центре москвы потушен
Transformer+3sent:	пожар в ресторане в центре москвы ликвидирован, пострадавших нет
Transformer+full:	пожар в ресторане в центре москвы ликвидирован
Transformer+ner:	пожар в центре москвы потушен

Thank you for your attention!

Valentin Malykh
valentin.malykh@phystech.edu