Phrase-Based Attentional Transformer for Headline Generation

Sokolov Andrey

31 may 2019 г.

St Petersburg University

Baseline

Dataset $\mathcal{D}$

Pair news-title $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$

Baseline

Dataset $\mathcal{D}$

Pair news-title $(\mathbf{x}, \mathbf{y}) \in \mathcal{D}$

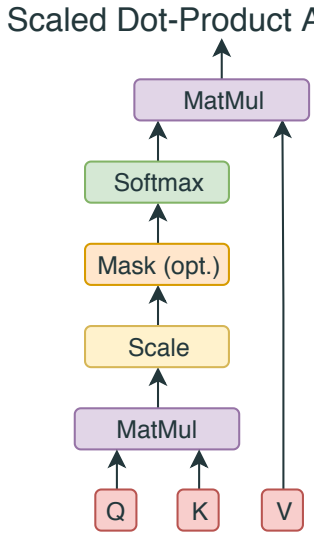
Language modelling:

$$p_\theta(y_1, \ldots, y_m | x_1 \ldots, x_n) = \prod_j p_\theta(y_j | y_1, \ldots, y_{j-1}, x_1, \ldots, x_n) \quad (1)$$

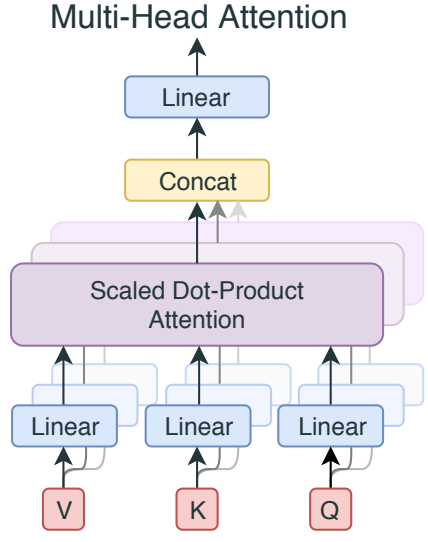
$$\hat{\theta}_{MLE} = \arg\max_\theta \prod_{(\mathbf{x}^i, \mathbf{y}^i) \in \mathcal{D}} p_\theta(\mathbf{y}^i | \mathbf{x}^i) \quad (2)$$

Transformer architecture

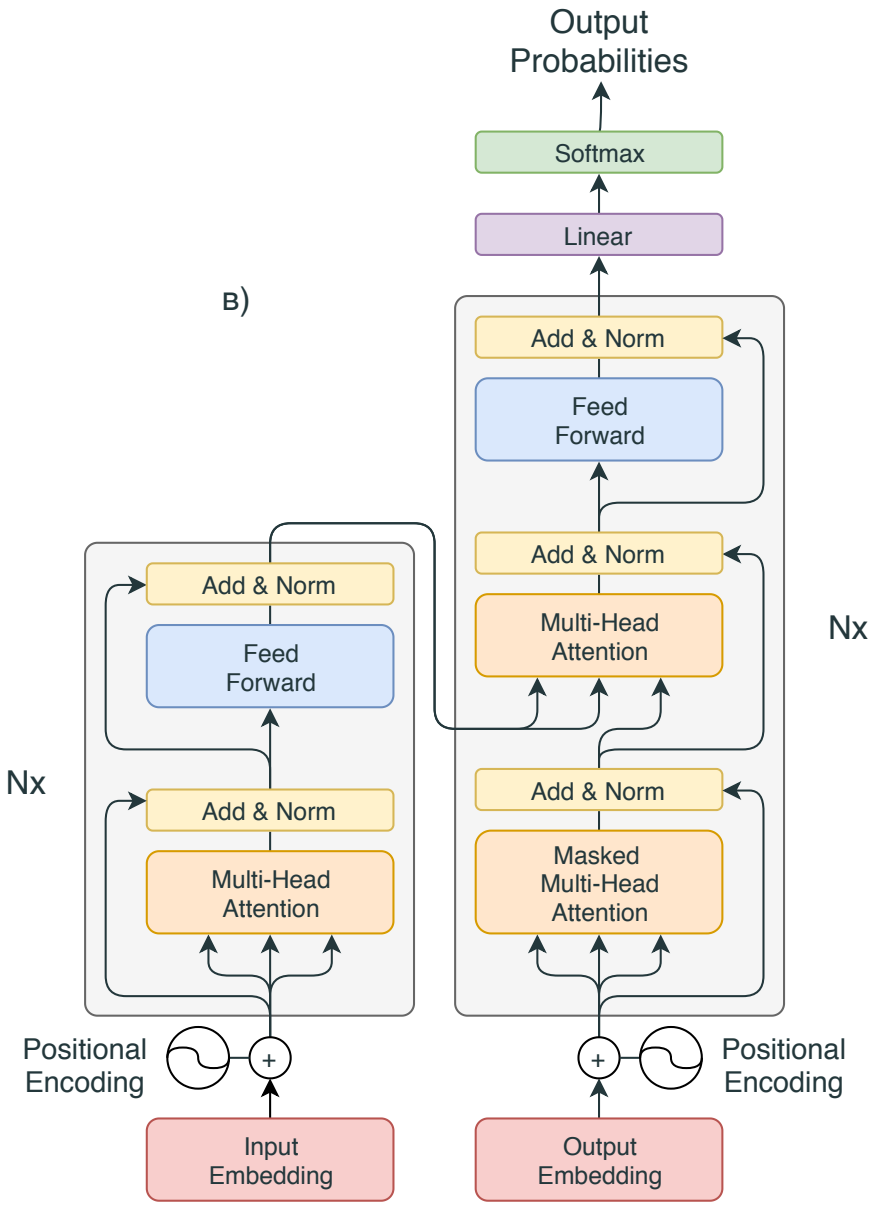
а)



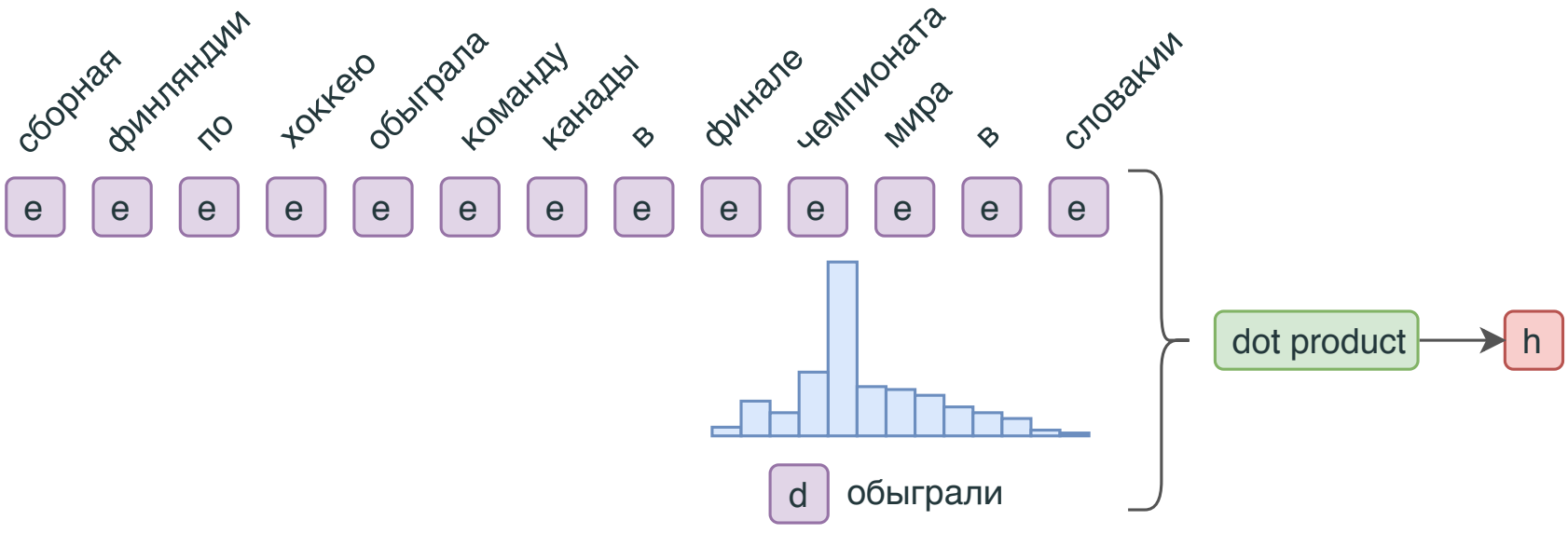
б)



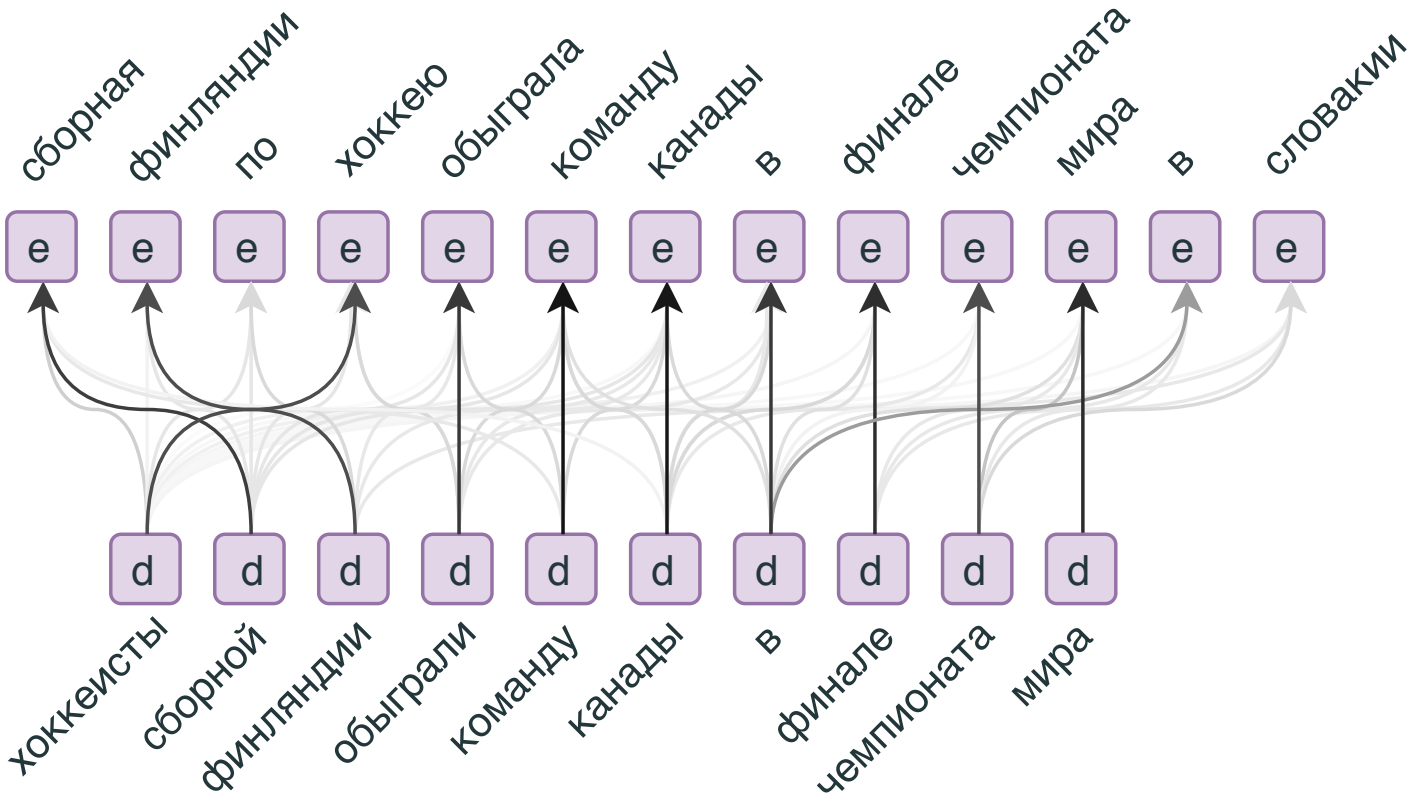
в)



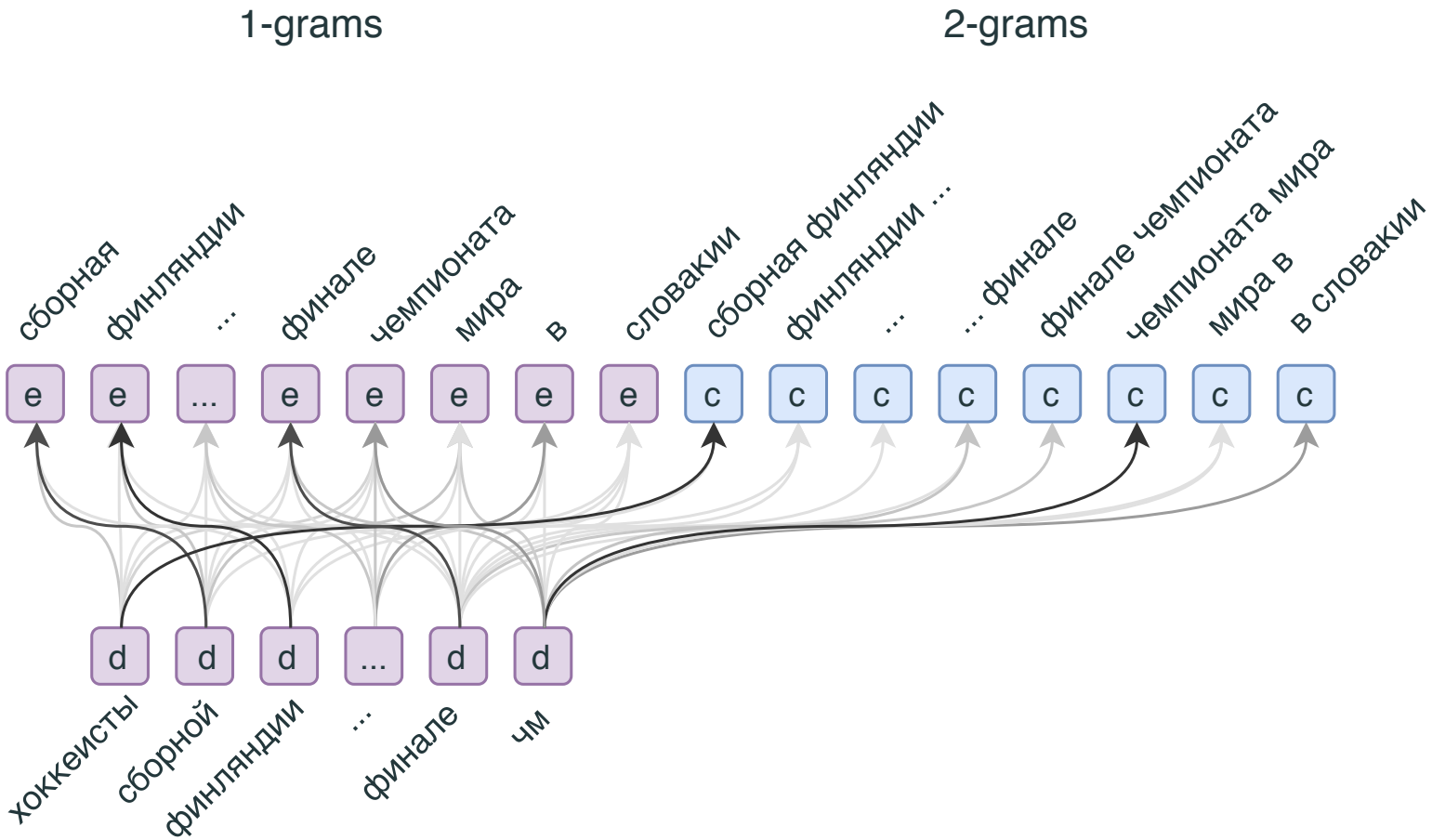
Attention


сборная
финляндии
по
хоккею
обыграла
команду
канады
в
финале
чемпионата
мира
в
словакии
e
e
e
e
e
e
e
e
e
e
e
e
e
d
обыграли
dot product
h

Attention


сборная
финляндии
по
хоккею
обыграла
команду
канады
в
финале
чемпионата
мира
в
словакии
e
e
e
e
e
e
e
e
e
e
e
e
e
d
d
d
d
d
d
d
d
d
d
хоккеисты
сборной
финляндии
обыграли
команду
канады
в
финале
чемпионата
мира

Phrase-based attention



Tokenization

Headline:
ракета-носитель falcon 9 вывела на орбиту 60 спутников

Tokenization

Headline:
ракета-носитель falcon 9 вывела на орбиту 60 спутников

Word-level tokenization

```
['<unk>', '<unk>', '9', 'вывела', 'на', 'орбиту',
'6','0', 'спутников']
```

Tokenization

Headline:
ракета-носитель falcon 9 вывела на орбиту 60 спутников

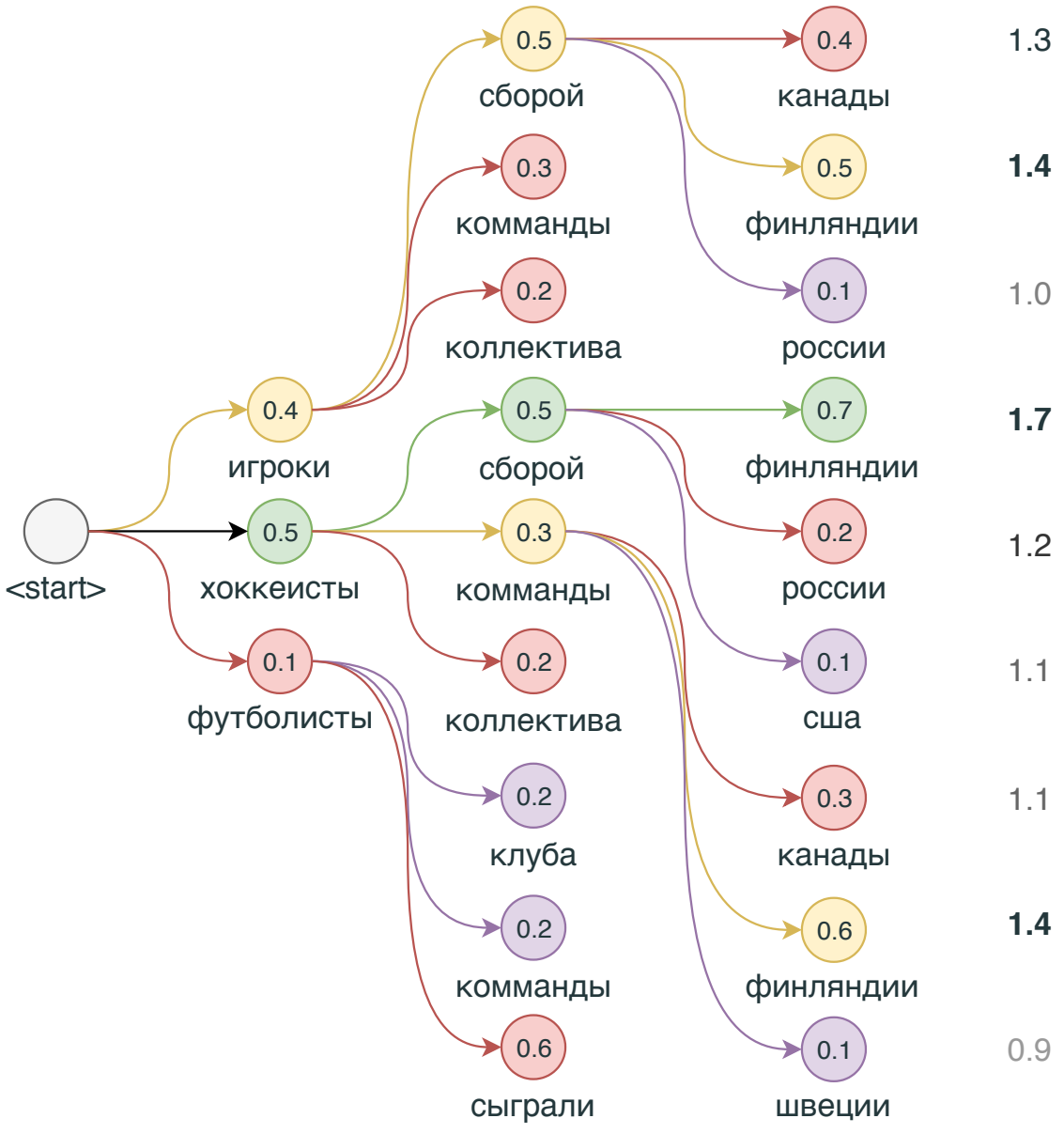
Word-level tokenization

```
['<unk>', '<unk>', '9', 'вывела', 'на', 'орбиту',
'6','0', 'спутников']
```

Byte pair encoding (BPE)

```
['_ракета', '-', 'носитель', '_f', 'al', 'con', '_9',
 '_вывел', 'а', '_на', '_орбиту', '_60', '_спутников']
```

Inference


<start>
0.4
игроки
0.5
хоккеисты
0.1
футболисты
0.5
сборой
0.3
комманды
0.2
коллектива
0.5
сборой
0.3
комманды
0.2
коллектива
0.2
клуба
0.2
комманды
0.6
сыграли
0.4
канады
0.5
финляндии
0.1
россии
0.7
финляндии
0.2
россии
0.1
сша
0.3
канады
0.6
финляндии
0.1
швеции
1.3
1.4
1.0
1.7
1.2
1.1
1.1
1.4
0.9

Sample 1

Text:

Сборная Финляндии по хоккею обыграла команду Канады в финале чемпионата мира в Словакии. Встреча в Братиславе завершилась со счетом 3:1 (0:1, 1:0, 2:0) в пользу финнов, в составе которых отличились Марко Анттила (23, 43) и Харри Песонен (56). ...

Headline:

Финляндия празднует победу в чемпионате мира по хоккею

Generated headline:

хоккеисты сборной финляндии обыграли команду канады в финале чм

Sample 2

Text:

Заключительная серия киносаги в жанре фэнтези "Игра престолов"установила новый рекорд просмотров в США, серию посмотрели более 19 миллионов зрителей, сообщает издание Hollywood Reporter. Шестая и заключительная серия восьмого сезона собрала 19,3 миллиона ...

Headline:

Финал "Игры престолов"установил новый рекорд просмотров в США

Generated headline:

"игра престолов"в сша побила рекорд просмотров

Sample 3

Text:

Искусственный интеллект как технология сам по себе не представляет угрозу, опасность заключается в том, как человек будет использовать создаваемые технологии, считает профессор Междисциплинарного центра биоэтики Йельского университета США Венделл Валлах. ...

Headline:

Американский ученый назвал главную опасность искусственного интеллекта

Generated headline:

искусственный интеллект не представляет угрозу, считает валлах

Results

Model	ROUGE-1-f	ROUGE-2-f	ROUGE-l-f
First Sentence	24.08	10.57	16.70
RNN	37.98	20.51	35.36
Universal Transformer	39.75	22.15	36.81
Vanilla Transformer	42.42	25.06	39.50
PBA Transformer	**42.96**	**25.43**	**40.02**

Таблица 1: Evaluation results on RIA dataset

Conclusion

- The problem is nontrivial
- Style sensitivity
- Domain sensitivity
- Text size limit
- Evaluation