Автоматическое аннотирование (суммаризация)

Гусев Илья

Московский физико-технический институт

Москва, 2020

Содержание

Задача

- Получение по тексту большего размера текста меньшего размера, каким-то образом отражающего содержание исходного текста.
- Extractive summarization - суммаризация без использования новых слов
- Abstractive summarization - суммаризация с использованием новых слов
- Обычно короткими текстами выступают:
  1. Аннотации текстов
  2. Названия текстов

Корпусы для английского: single-document

1. CNN/Daily Mail
   - Основной датасет, 300k примеров
   - Есть в 2 вариантах: анонимизированный и нет
   - Один текст - много хайлайтов, которые обычно конкатенируют
2. Gigaword
   - Генерация заголовков новостей, 4kk примеров
3. Cornell Newsroom
   - Саммари к новостям, 1.3kk пар
4. X-Sum
   - Саммари к новостям в одну строку, 220k пар
5. Arxiv и PubMed
6. Multi-News
   - Для мультидокументной суммаризации
   - До 10 источников
   - 40к уникальных кластеров
7. DUC 2002, 2004; TAC 2011 - закрытые датасеты
8. New York Times Annotated Corpus - закрытый датасет

Корпусы для русского

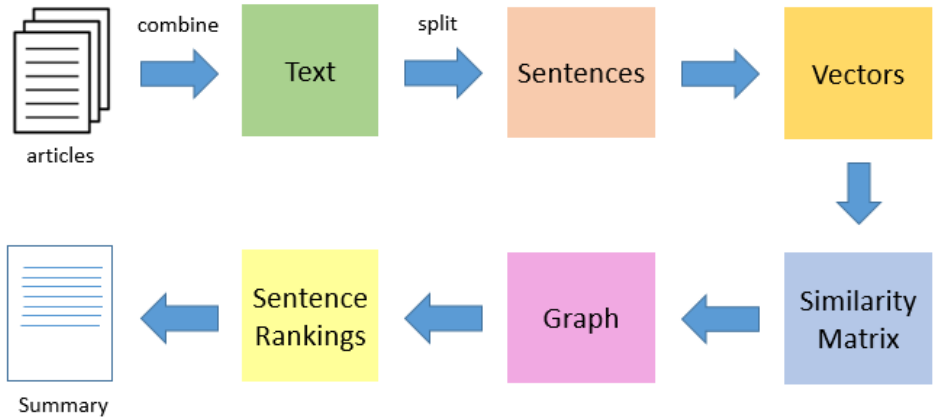
- RIA news dataset
- Lenta
- Новостная коллекция РОМИП

Метрики

Типичная text2text задача:

1. BLEU
2. ROUGE - основная
   - ROGUE-N. Считаем точность, полноту и f-меру вхождений n-грам из сгенерированного ответа в эталонном. Обычно, приводят только f-меру
   - ROUGE-L – длина наибольшей общей подпоследовательности ответа и эталона
3. METEOR

Extractive summarization: TextRank

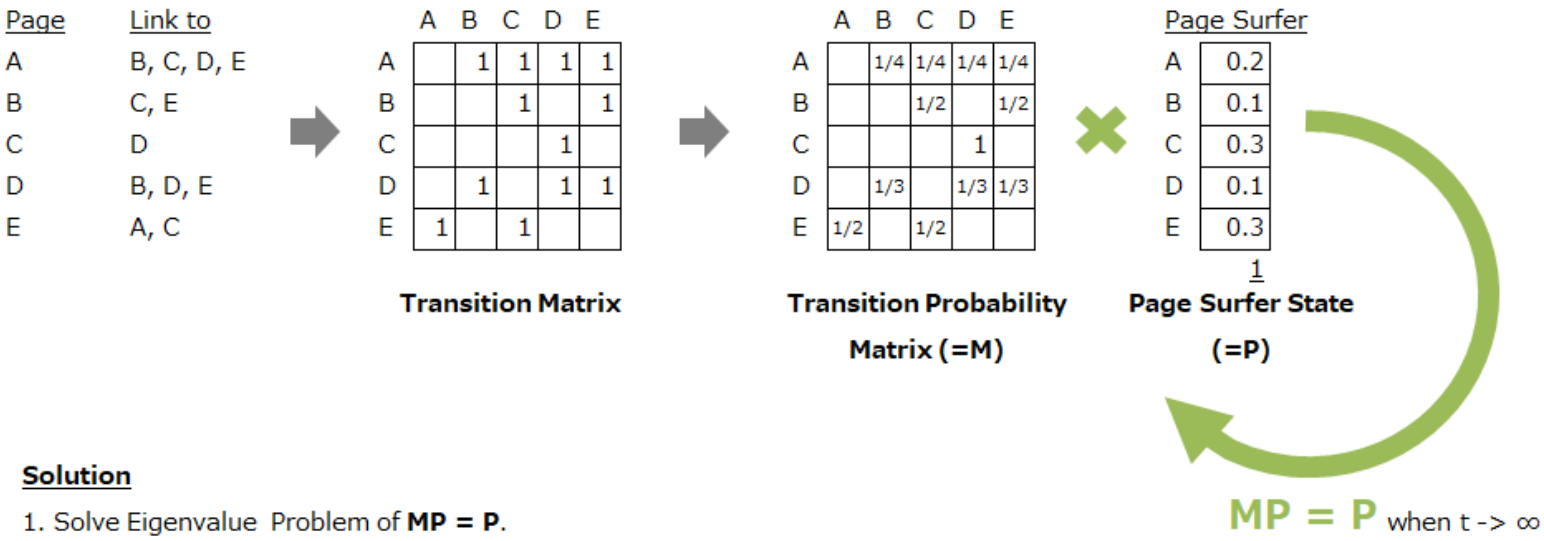


- Решение на уровне предложений.
- Разбиваем на предложения, считаем похожесть каждого с каждым
- Строим матрицу, считаем PageRank, выбираем топ по нему

Источники:

1. [analyticsvidhya.com] An Introduction to Text Summarization using the TextRank AIgorithm
2. [igorshevchenko.ru] Суммаризация с помощью TextRank

TextRank: минутка PageRank'a



Solution

1. Solve Eigenvalue Problem of **MP = P**.
2. Repeat the transition until convergence(**MP - P < threshold**).

$P_i' = (1-d) + d * M_i^T P_i$ *The page surfers randomly click the page with a probability*

$\sum(P_i' - P_i) < threshold$ *of **1-d**. (d = usually 0.85)*

Источники:

1. [github.com] The guide to tackle with the Text Summarization.
2. [ams.org] PageRank explanation

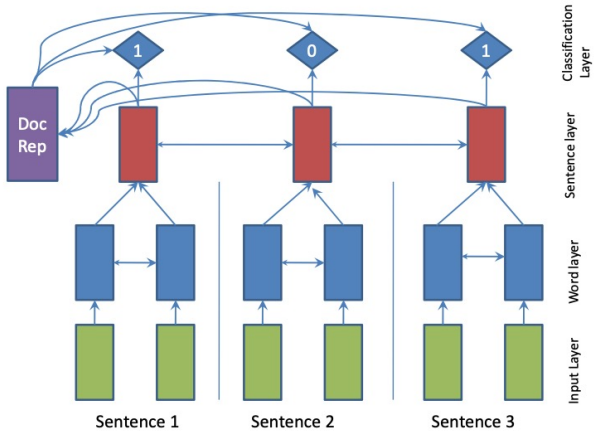
Extractive summarization как задача классификации предложений

- Хотим просто делать бинарную классификацию предложений исходного текста
- Проблема - их нет в саммари
- Идея - нужно выбрать такое покрытие предложений исхожного текста, чтобы масимизировать метрики с настоящими саммари
  1. Прямой перебор - 2^N вычислений метрики, где N - количество предложений в исходном тексте
  2. Жадный алгоритм - выбираем лучшее предложение на каждом шаге, пока метрика улучшается: $N \cdot M$ вычислений метрики, где M - число набранных предложений
  3. Перебераем все одиночные предложения, потом все пары, потом тройки, до тех пор, пока увеличение количества не перестанет давать прирост по метрике: $\binom{N}{1} + \binom{N}{2} + \ldots + \binom{N}{M}$
- В литературе - an oracle summary

lead-3 baseline

- Первые 3 предложения исходного текста

Extractive summarization: SummaRuNNer

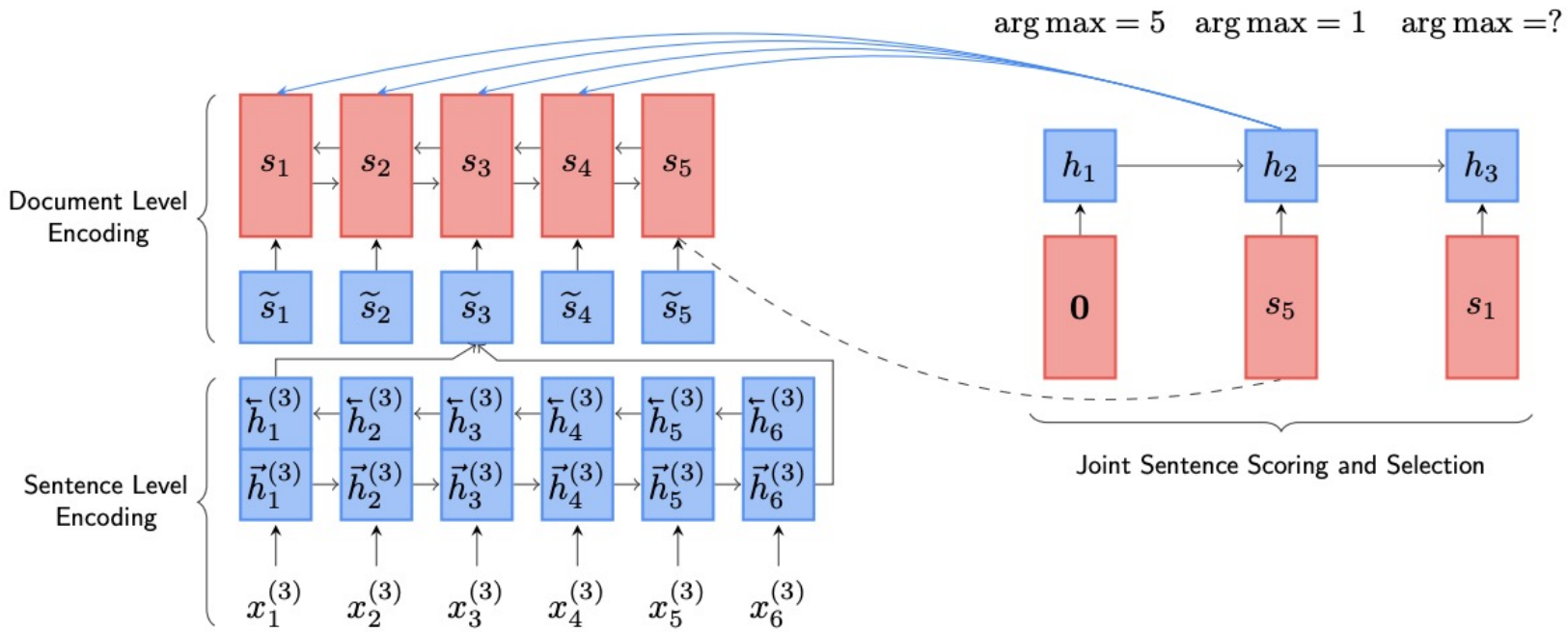


- BiRNN по словам для каждого предложения, усредняем все выходы, получаем S предложения
- BiRNN по S предложений, S документа - линейный слой с активацией над усреднением всех S на выходе
- Предсказываем 1 и 0, используя S на выходе, S документа, S саммари

Источники:

1. [arxiv.org] SummaRuNNer: A Recurrent Neural Network based Sequence Model for Extractive Summarization of Documents

Extractive summarization: NeuSum

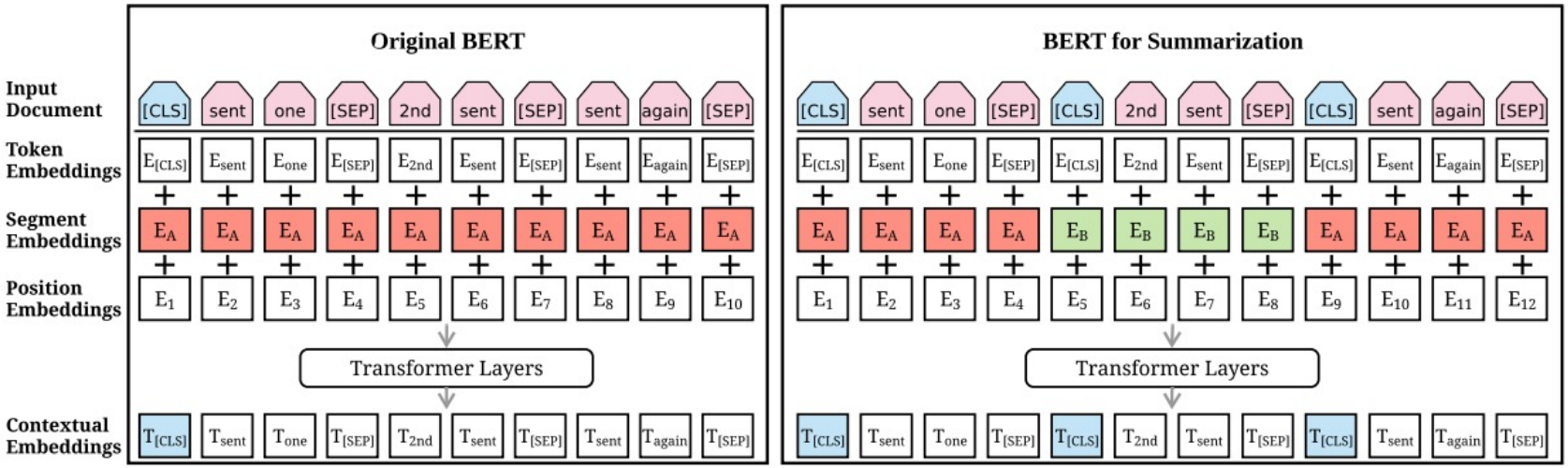


- В обучении KL-дивергенция между предсказанным приростом метрики и реальным, хитро отнормированным

Источники:

1. [aclweb.org] Neural Document Summarization by Jointly Learning to Score and Select Sentences

Extractive summarization: BertSumExt



- CLS токены на входе для каждого предложения
- Чередующиеся E_a и E_b
- Над последовательностью $T_{[CLS]}$ ещё пара слоёв Трансформера для классификаиции

Источники:

1. [arxiv.org] Text Summarization with Pretrained Encoders

Abstractive summarization: Seq2Seq+Attn

Bahdanau et al. в 2014 для машинного перевода.

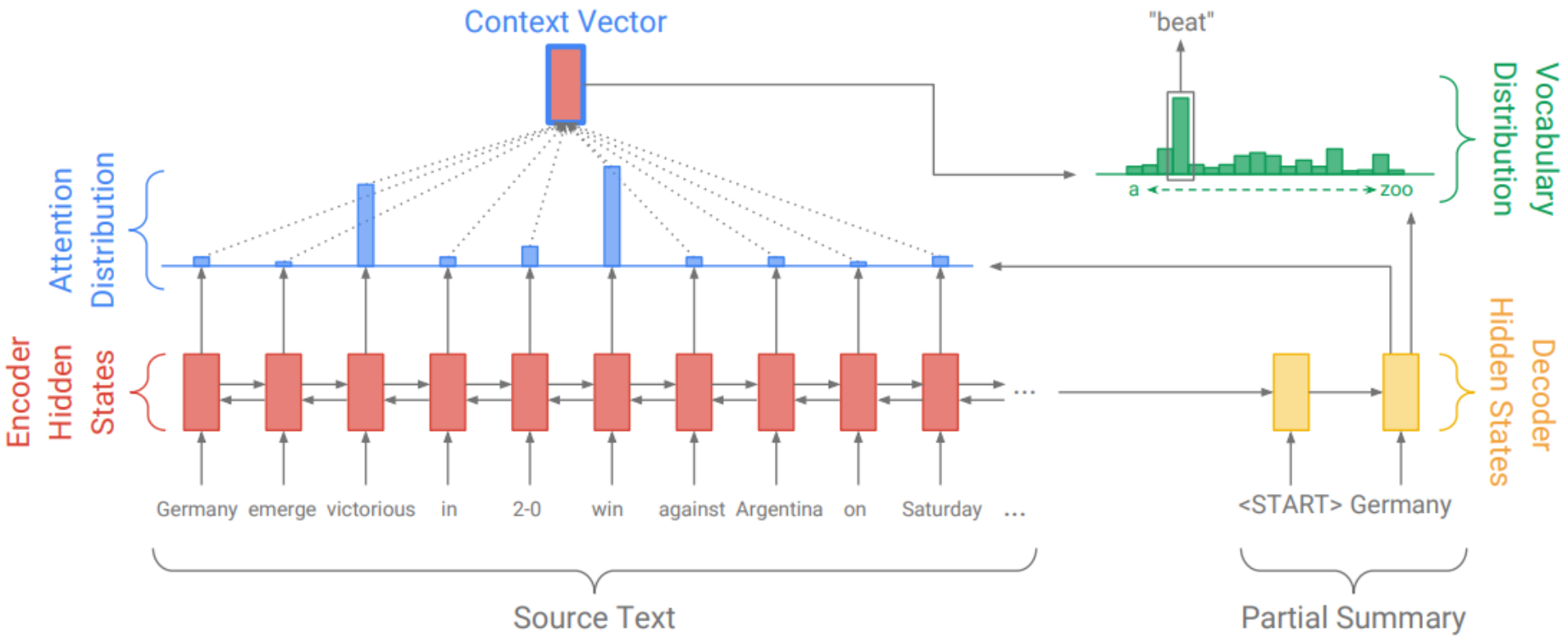
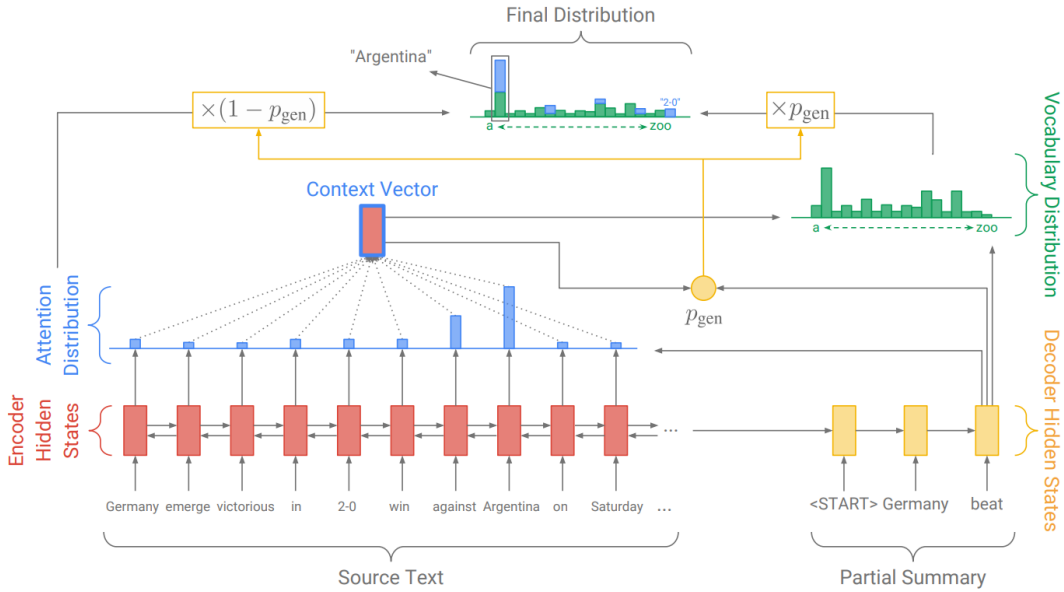


Рис.: Seq2seq model with attention. Illustration from See et al., 2017

Abstractive summarization: Pointer-Generator Networks



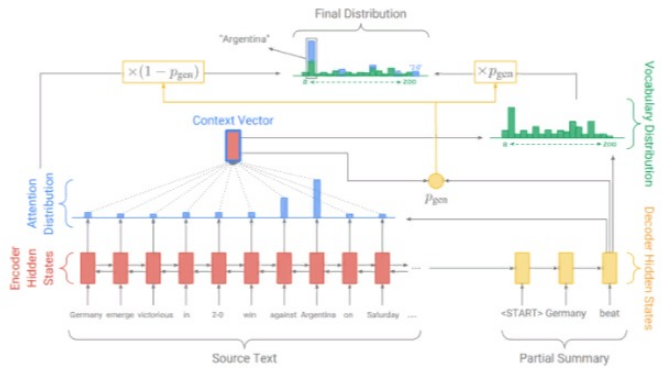
- С вычислимой вероятностью берём слова из оригинального текста на основе распределения внимания
- Coverage - храним сумму всех весов внимания на всех шагах, подаём её на вход самому вниманию, делаем специальный лосс

Источники:

1. [arxiv.org] Get To The Point: Summarization with Pointer-Generator Networks

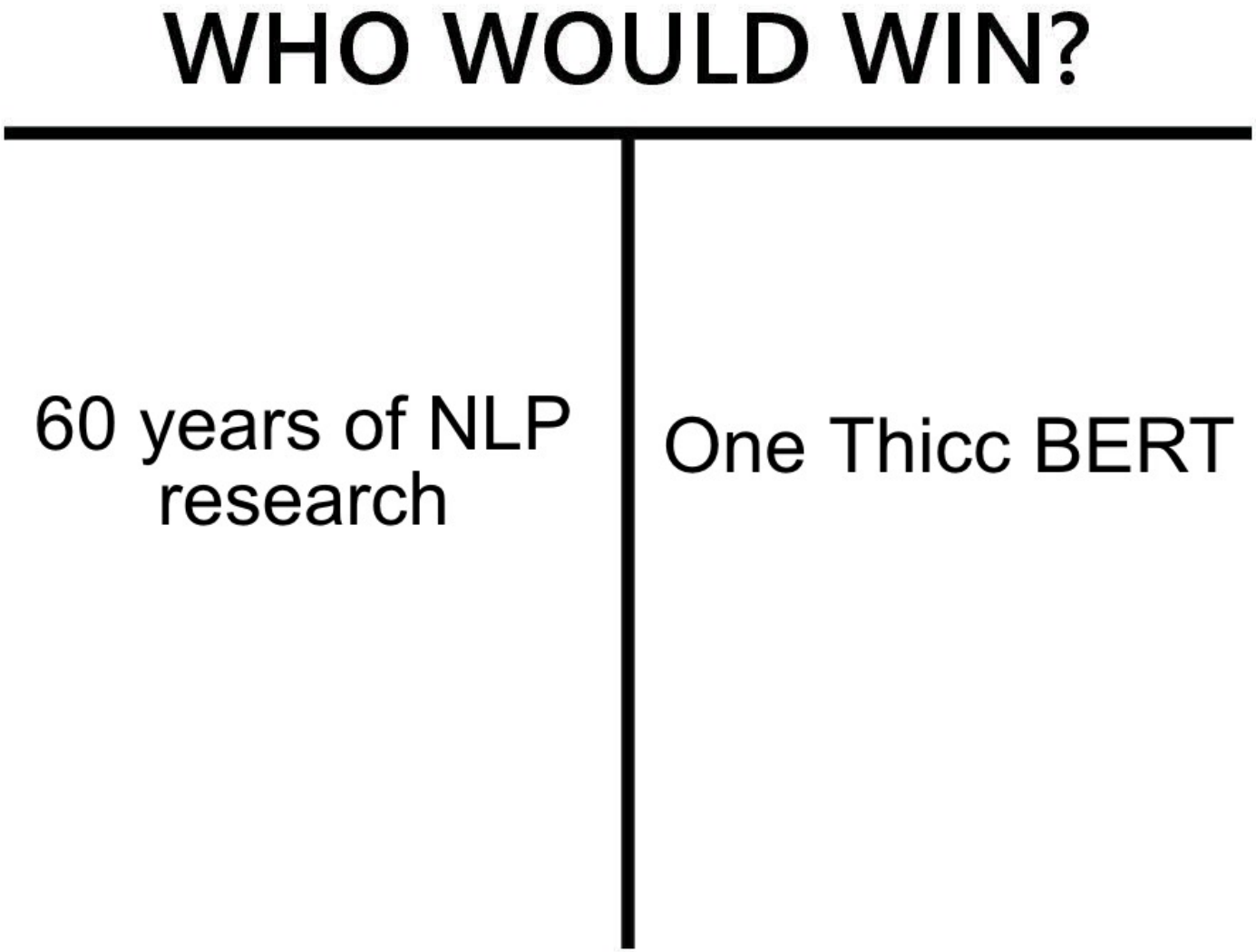
Abstractive summarization: Pointer-Generator Networks

WHO WOULD WIN?

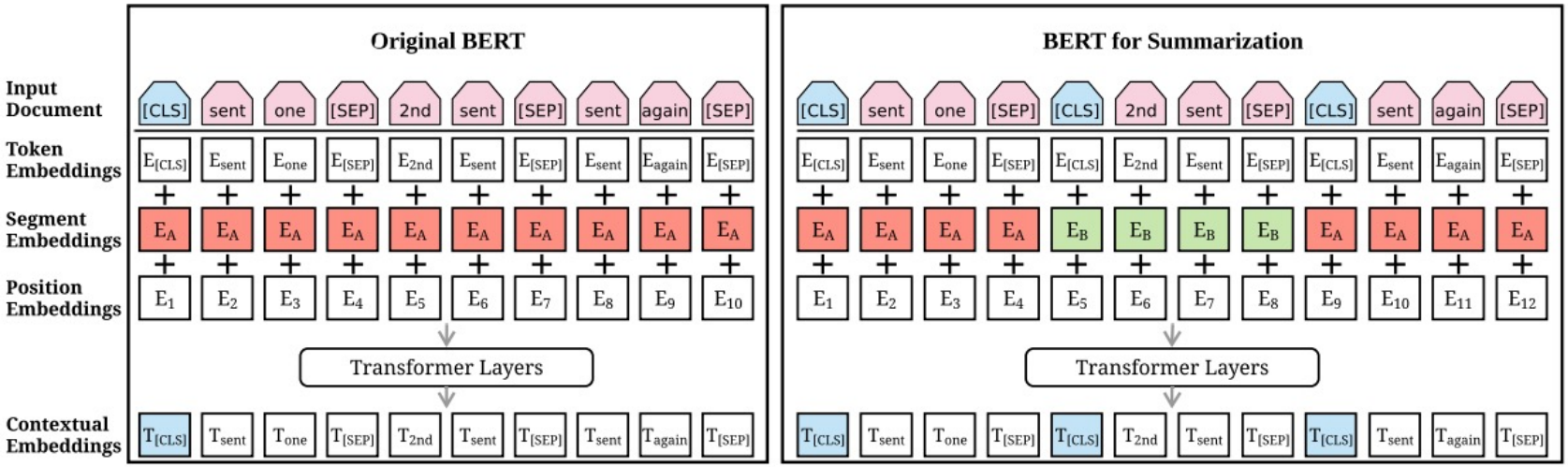


One Thicc lead-3

Abstractive summarization: BertSumAbs



Extractive summarization: BertSumAbs



- Декодер: 6-слойный Трансформер
- Хитрые схемы обучения

Источники:

1. [arxiv.org] Text Summarization with Pretrained Encoders

Стоит упомянуть

1. [arxiv.org] Incorporating Copying Mechanism in Sequence-to-Sequence Learning
2. [arxiv.org] Deep Reinforcement Learning for Sequence-to-Sequence Models
3. [aclweb.org] Deep Communicating Agents for Abstractive Summarization
4. [aclweb.org] Encode, Tag, Realize: High-Precision Text Editing

Полезные ссылки I

- Data Scientist's Guide to Summarization: A text summarization tutorial for beginners
  shorturl.at/dFHSX
- awesome-text-summarization
  https://github.com/icoxfog417/awesome-text-summarization