Нейросетевые языковые модели для поиска и анализа научных публикаций

Воронцов Константин Вячеславович
k.vorontsov@iai.msu.ru
д.ф.-м.н., профессор РАН, зав. каф. ММП ВМК МГУ,
зав. лаб. Машинного обучения и семантического анализа
Института искусственного интеллекта МГУ

XIV Международная молодежная
научно-практическая конференция
с элементами научной школы
«Прикладная математика и
фундаментальная информатика»
(ПМиФИ 2024)

ОмГТУ, Омск • 20–25 мая 2024

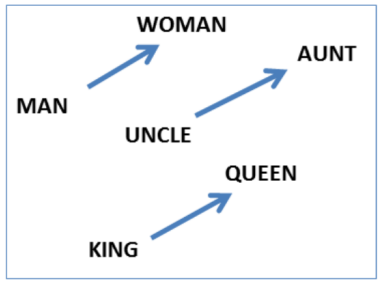
Содержание

Эволюция подходов машинного обучения в анализе текстов

Декомпозиция задач по уровням пирамиды NLP

- морфологический анализ, лемматизация, опечатки
- синтаксический анализ, выделение терминов, NER
- семантический анализ, выделение фактов, тем

Модели векторных представлений (эмбедингов) слов на основе матричных разложений



- модели дистрибутивной семантики:
  word2vec [Mikolov, 2013], FastText [Bojanowski, 2016]
- тематические модели LDA [Blei, 2003], ARTM [2014]

Нейросетевые модели локальных контекстов

- рекуррентные нейронные сети
- модели внимания и трансформеры:
  BERT [2018], GPT-3 [2020], GPT-4 [2023]

$$\text{softmax}\left(\frac{Q \times K^T}{\sqrt{d}}\right) V$$

«Мастерской знаний»: мотивация проекта

«Огромное и все возрастающее богатство знаний разбросано сегодня по всему миру. Этих знаний, вероятно, было бы достаточно для решения всего громадного количества трудностей наших дней, но они рассеяны и неорганизованы. Нам необходима очистка мышления в *своеобразной мастерской*, где можно получать, сортировать, суммировать, усваивать, разъяснять и сравнивать знания и идеи»
— *Герберт Уэллс, 1940*

"An immense and ever-increasing wealth of knowledge is scattered about the world today; knowledge that would probably suffice to solve all the mighty difficulties of our age, but it is dispersed and unorganized. We need a sort of mental clearing house for the mind: a depot where knowledge and ideas are received, sorted, summarized, digested, clarified and compared" — *Herbert Wells, 1940*

Теперь технологии NLP позволяют решать такие задачи

От поиска информации к «Мастерской знаний»

Недостатки обычного поиска:

- как искать новые знания?
- что делать с найденным?

Мастерская знаний — инструментарий для автоматизации последующих этапов работы с текстовыми источниками:

- ищу – чтобы накапливать
- накапливаю – чтобы анализировать
- анализирую – чтобы понимать
- понимаю – чтобы применять и передавать

Это задачи, связанные с *автоматической обработкой текстов* (только применение знаний остаётся за пределами системы)

Концепция сервисов «Мастерской знаний»

Подборка — долгосрочный поисковый интерес пользователя

Поисково-рекомендательные функции:

- поиск тематически близких документов по подборке
- мониторинг новых документов для подборки
- контекстные рекомендации по документу из подборки

Аналитические функции:

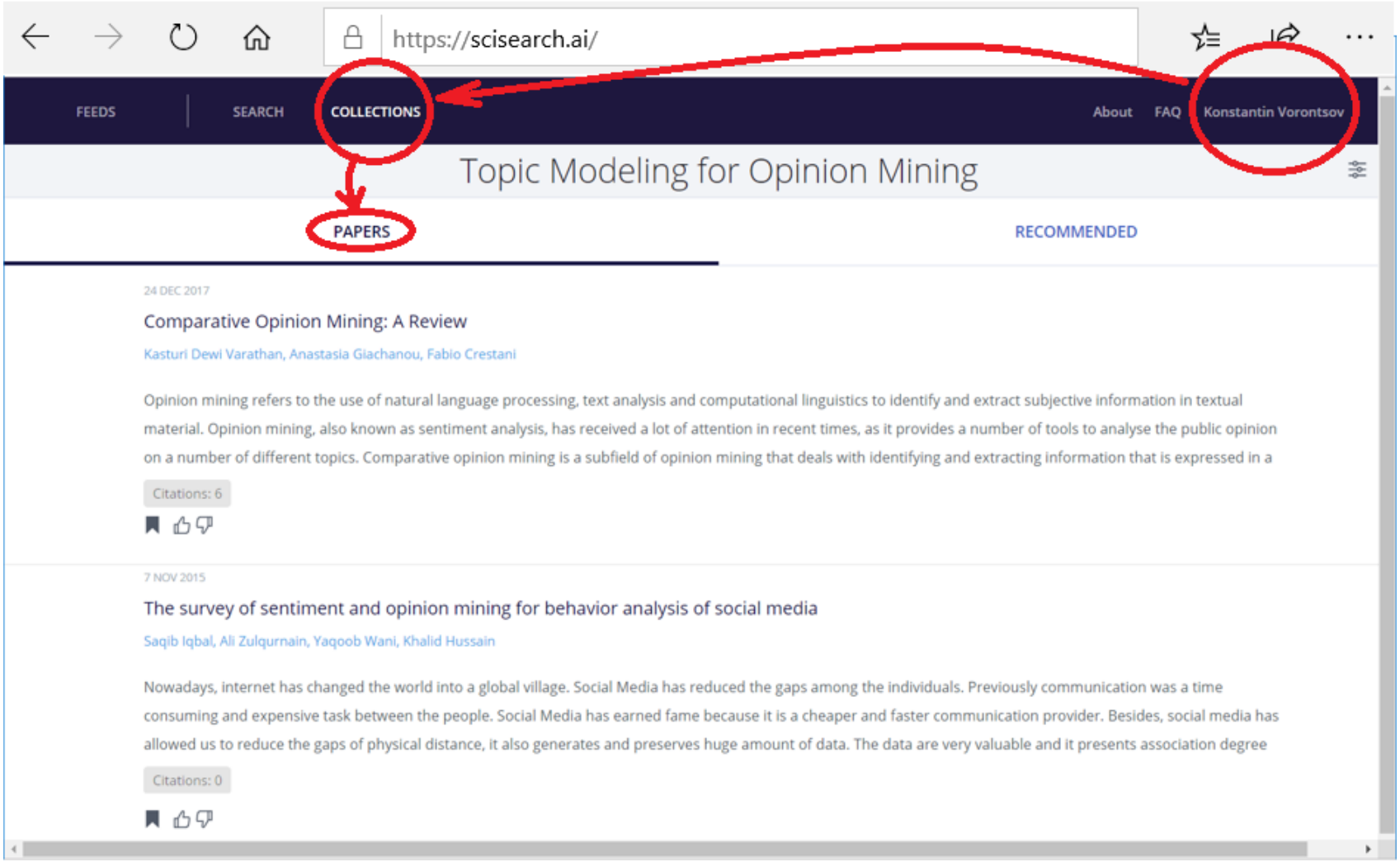
- автоматизация реферирования подборки
- кластеризация трендов, аспектов, отношений в подборке
- рекомендация порядка чтения внутри подборки
- визуализация карт знаний в виде mind-map по подборке

Коммуникативные функции:

- совместное составление и использование подборок
- интерактивная визуализация и инфографика по подборке

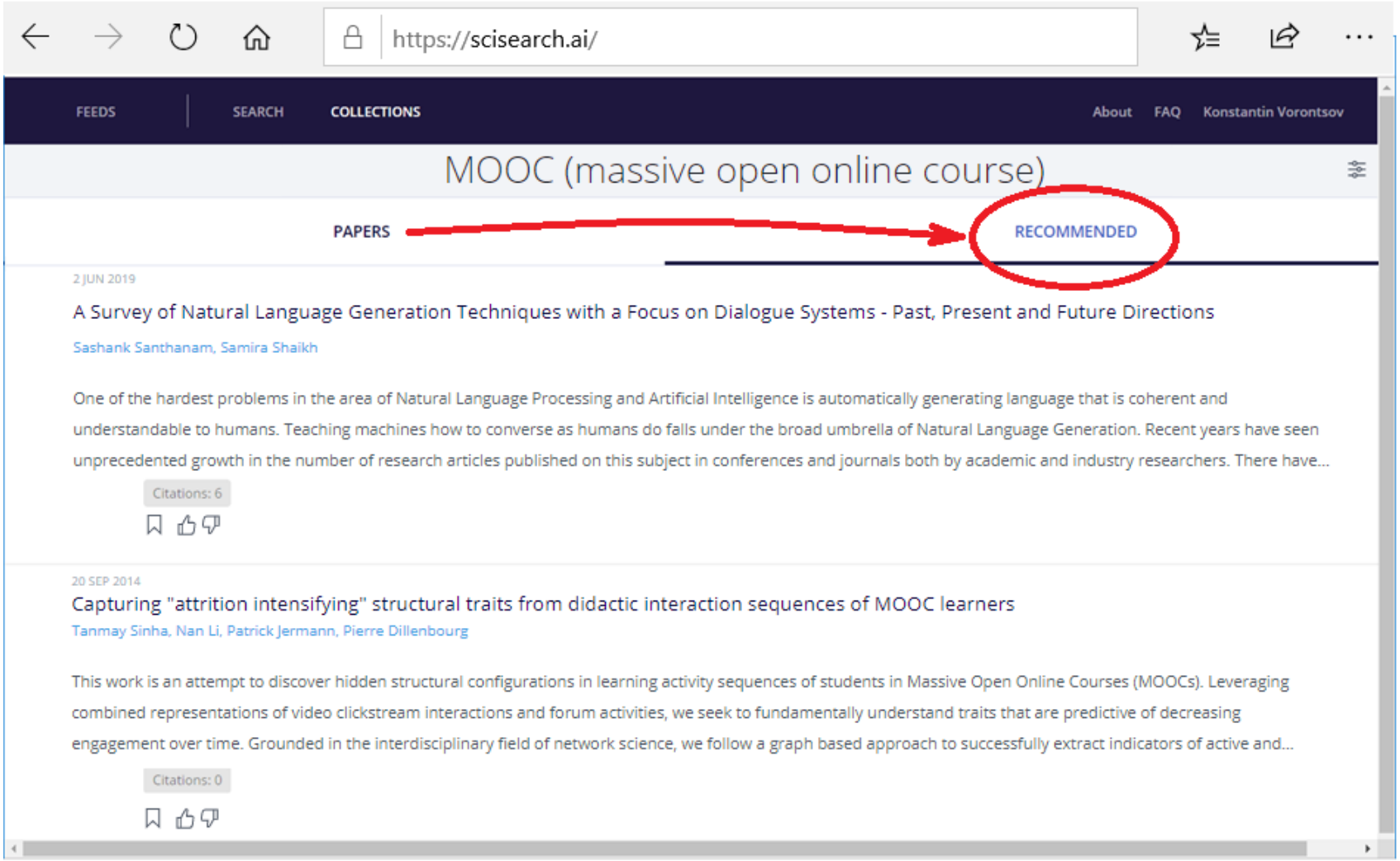
Прототип поисково-рекомендательной системы

Тематическая подборка пользователя:



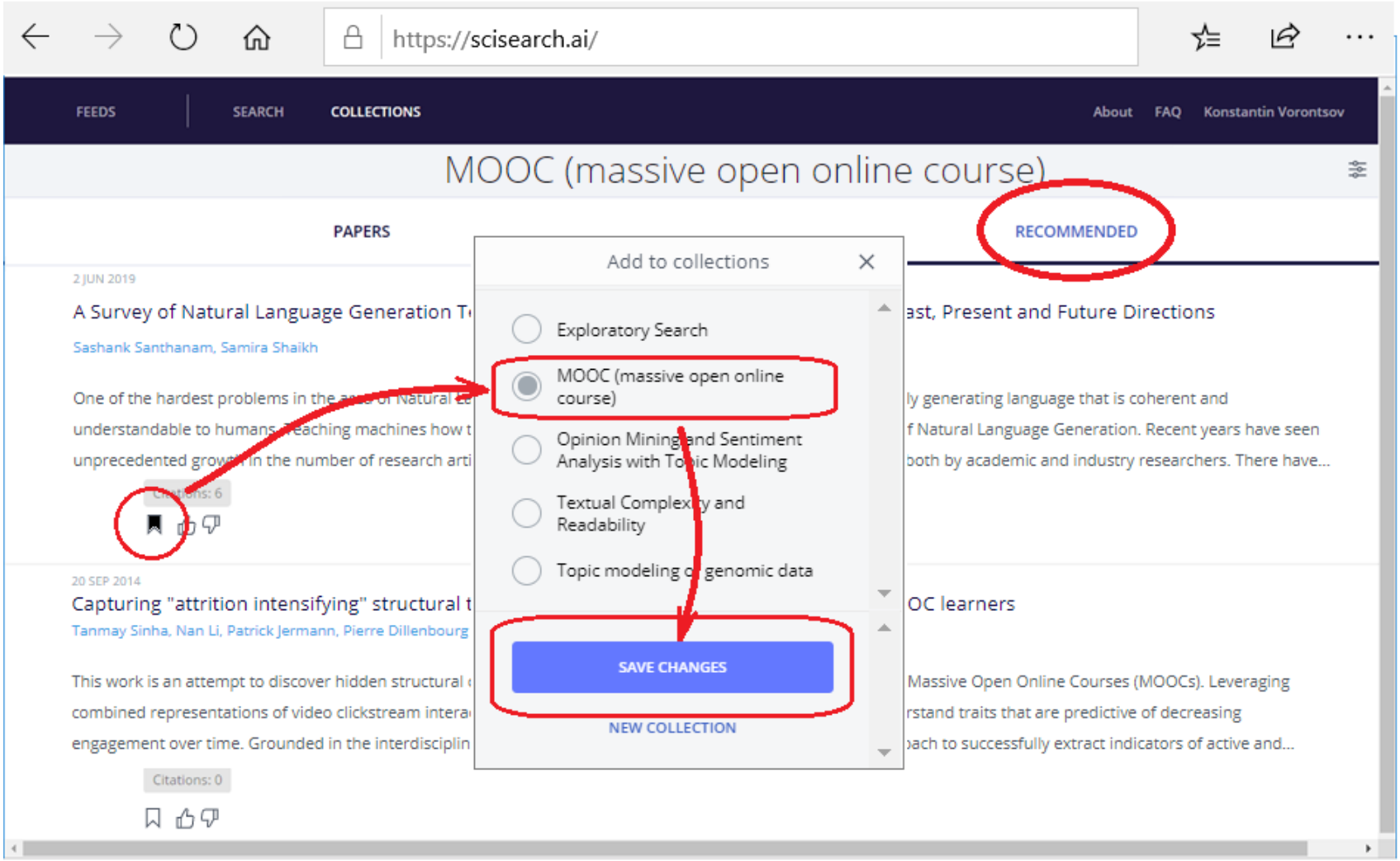
Прототип поисково-рекомендательной системы

Список статей, рекомендуемых для добавления в подборку:



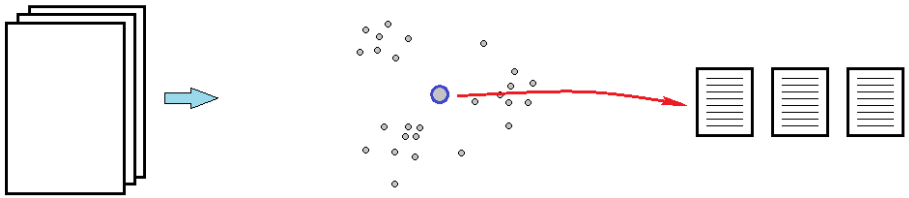
Прототип поисково-рекомендательной системы

Добавление статьи из списка рекомендаций в подборку:

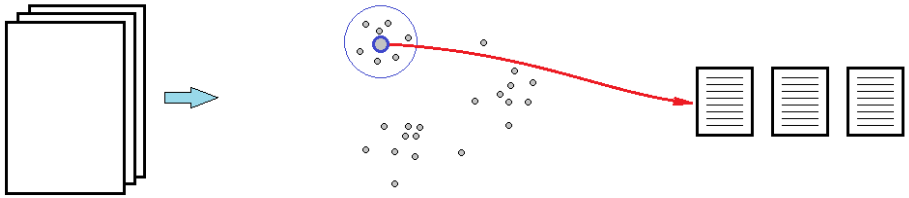


Стратегии векторного поиска документов по документам

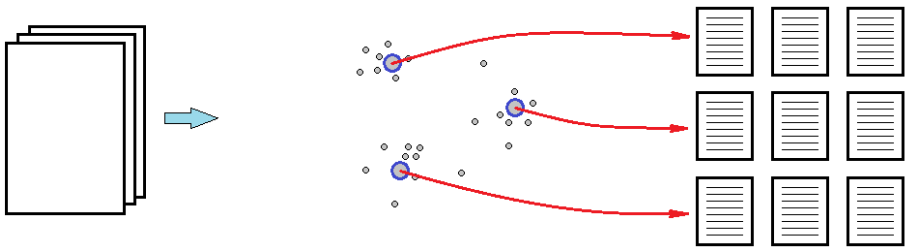
Поиск по среднему вектору подборки (неудачная стратегия):



Поиск по части подборки, близкой к выбранному документу:

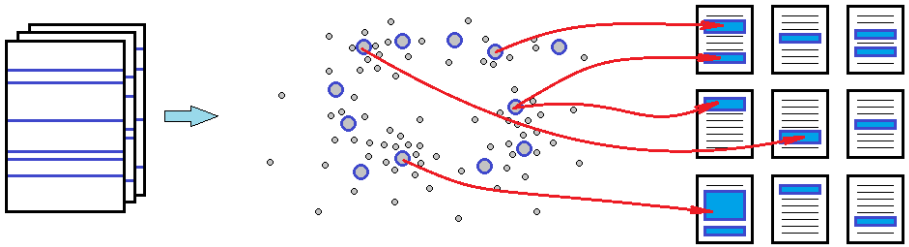


Поиск по тематике кластеров, на которые делится подборка:

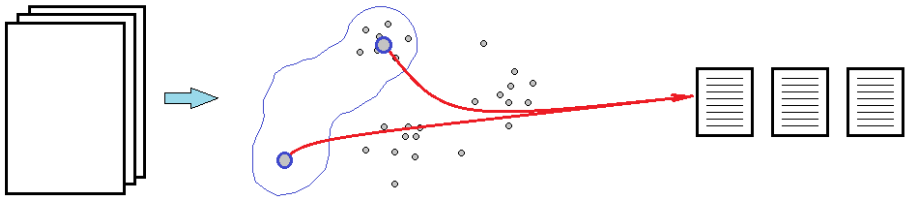


Стратегии векторного поиска документов по документам

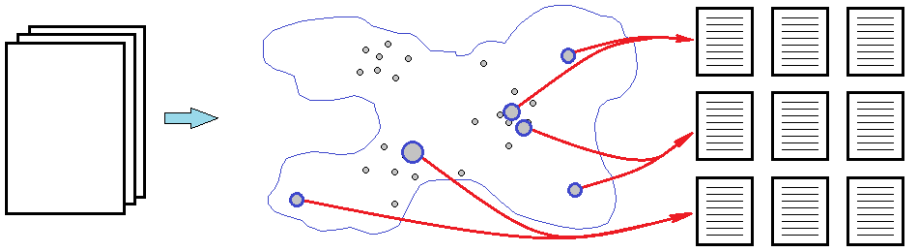
Поиск по тематике сегментов документов подборки:



Поиск по тематике, смежной для части подборки:

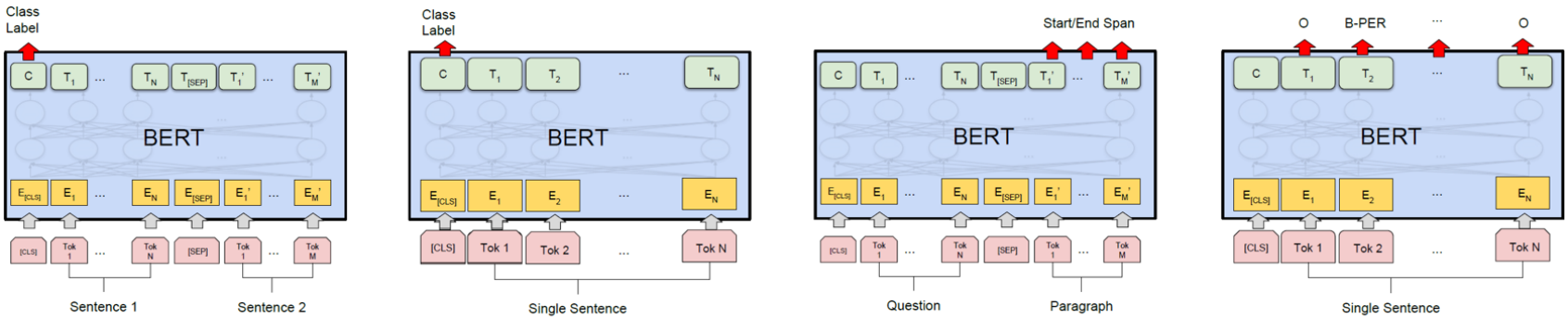


Поиск по тематике, смежной для всей подборки:



Большие пред-обученные модели языка (трансформеры)

- обучены по терабайтам текстов, «они видели в языке всё»
- предсказывая слова по контексту, способны также
  — выделять и классифицировать фрагменты текста,
  — генерировать фейковые тексты, и т. д.
- *мультиязычны*: обучаются на десятках языков
- *мультизадачны*: для каждой новой задачи NLP/NLU достаточно дообучения на малой размеченной выборке



(a) Sentence Pair Classification Tasks: MNLI, QQP, QNLI, STS-B, MRPC, RTE, SWAG

(b) Single Sentence Classification Tasks: SST-2, CoLA

(c) Question Answering Tasks: SQuAD v1.1

(d) Single Sentence Tagging Tasks: CoNLL-2003 NER

J.Devlin et al. BERT: Pre-training of deep bidirectional transformers for language understanding. 2019.

Трасформер для машинного перевода

Трасформер (transformer) — это нейросетевая архитектура для трансформации векторов слов в контекстно-зависимые

Схема преобразований данных в машинном переводе:

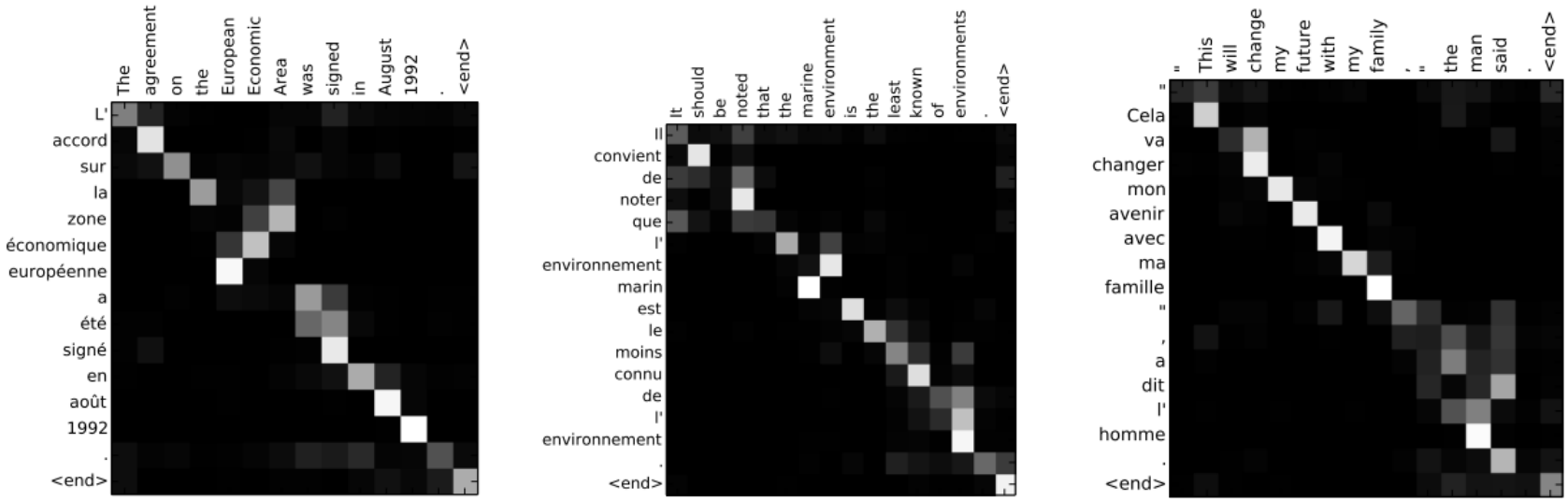
- $S = (w_1, \ldots, w_n)$ — слова предложения на входном языке
  $\downarrow$ обучаемая или пред-обученная векторизация слов
- $X = (x_1, \ldots, x_n)$ — векторы слов входного предложения
  $\downarrow$ трансформер-кодировщик
- $Z = (z_1, \ldots, z_n)$ — контекстно-зависимые векторы слов
  $\downarrow$ трансформер-декодировщик, похож на кодировщика
- $Y = (y_1, \ldots, y_m)$ — векторы слов выходного предложения
  $\downarrow$ генерация слов из построенной языковой модели
- $\tilde{S} = (\tilde{w}_1, \ldots, \tilde{w}_m)$ — слова предложения на выходном языке

Vaswani et al. (Google) Attention is all you need. 2017.

Модели внимания для машинного перевода

$X = (x_1, \ldots, x_n)$ — векторы слов входного предложения
$Y = (y_1, \ldots, y_m)$ — векторы слов выходного предложения

Модель внимания оценивает матрицу семантического сходства $A_{ti} = a(x_i, y_t)$ — насколько входное слово x_i важно (требует внимания) для обработки выходного слова y_t



Bahdanau et al. Neural machine translation by jointly learning to align and translate. 2015.

Модель внимания Query–Key–Value

q — *вектор-запрос* для трансформации в вектор-контекст z
$K = (k_1, \ldots, k_n)$ — *векторы-ключи*, сравниваемые с запросом
$X = (x_1, \ldots, x_n)$ — *векторы-значения*, образующие контекст

Модель внимания — трёхслойная сеть, вычисляющая z как выпуклую комбинацию векторов x_i, релевантных запросу q:

$$z = \text{Attn}(q, K, X) = \sum_i x_i \,\text{SoftMax}_i\, a(k_i, q),$$

где $a(k, q)$ — *оценка релевантности* ключа k запросу q,
например $a(k, q) = k^\intercal q$ или $k^\intercal W q$ с матрицей параметров W

Модель внутреннего внимания (самовнимания, self-attention):

$$z_i = \text{Attn}(W_q x_i, W_k X, W_v X)$$

трансформирует входную последовательность $X = (x_1, \ldots, x_n)$
в выходную последовательность векторов контекста $(z_1, \ldots, z_n)$

Архитектура трансформера-кодировщика

1. Добавляются позиционные векторы p_i:
$h_i = x_i + p_i, \quad H = (h_1, \ldots, h_n)$
$d = \dim x_i, p_i, h_i = 512$
$\dim H = 512 \times n$

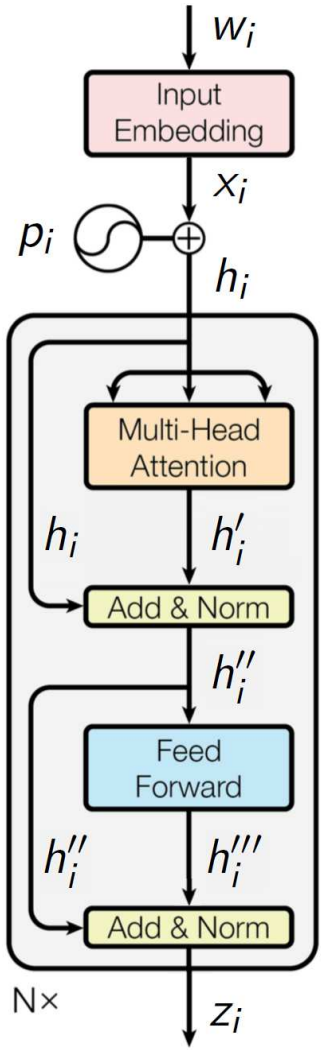
2. Многомерное самовнимание:
$h_i^j = \mathrm{Attn}(W_q^j h_i, W_k^j H, W_v^j H)$
$j = 1, \ldots, J = 8$
$\dim h_i^j = 64$
$\dim W_q^j, W_k^j, W_v^j = 64 \times 512$

3. Конкатенация (multi-head attention):
$h'_i = \mathrm{MH}_j(h_i^j) \equiv \left[h_i^1 \cdots h_i^J\right]$
$\dim h'_i = 512$

4. Сквозная связь + нормировка уровня:
$h''_i = \mathrm{LN}(h'_i + h_i; \mu_1, \sigma_1)$
$\dim h''_i, \mu_1, \sigma_1 = 512$

5. Полносвязная 2х-слойная сеть FFN:
$h'''_i = W_2\,\mathrm{ReLU}(W_1 h''_i + b_1) + b_2$
$\dim W_1 = 2048 \times 512$
$\dim W_2 = 512 \times 2048$

6. Сквозная связь + нормировка уровня:
$z_i = \mathrm{LN}(h'''_i + h''_i; \mu_2, \sigma_2)$
$\dim z_i, \mu_2, \sigma_2 = 512$



Несколько дополнений и замечаний

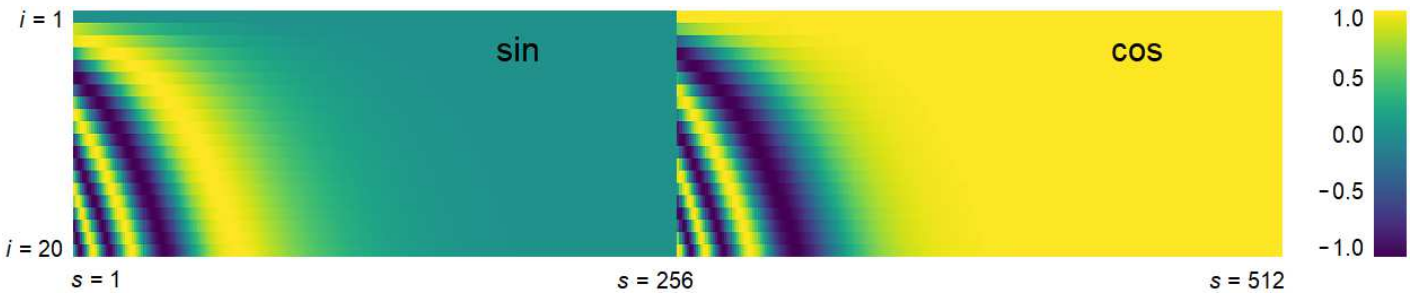
- $N = 6$ блоков $h_i \to \square \to z_i$ соединяются последовательно
- эмбединги слов $x_i \in \mathbb{R}^d$ — обучаемые или пред-обученные
- нормировка уровня (Layer Normalization), $x, \mu, \sigma \in \mathbb{R}^d$:

$$\mathrm{LN}_s(x; \mu, \sigma) = \sigma_s \frac{x_s - \bar{x}}{\sigma_x} + \mu_s, \quad s = 1, \ldots, d,$$

$\bar{x} = \frac{1}{d}\sum_s x_s$ и $\sigma_x^2 = \frac{1}{d}\sum_s (x_s - \bar{x})^2$ — среднее и дисперсия x

- Позиции слов i кодируются векторами p_i, $i = 1, \ldots, n$;
чем больше $|i - j|$, тем больше $\|p_i - p_j\|$, n не ограничено:

$$p_{is} = \sin(i\, 10^{-8\frac{s}{d}}), \quad p_{i,s+\frac{d}{2}} = \cos(i\, 10^{-8\frac{s}{d}}), \quad s = 1, \ldots, \tfrac{d}{2}$$

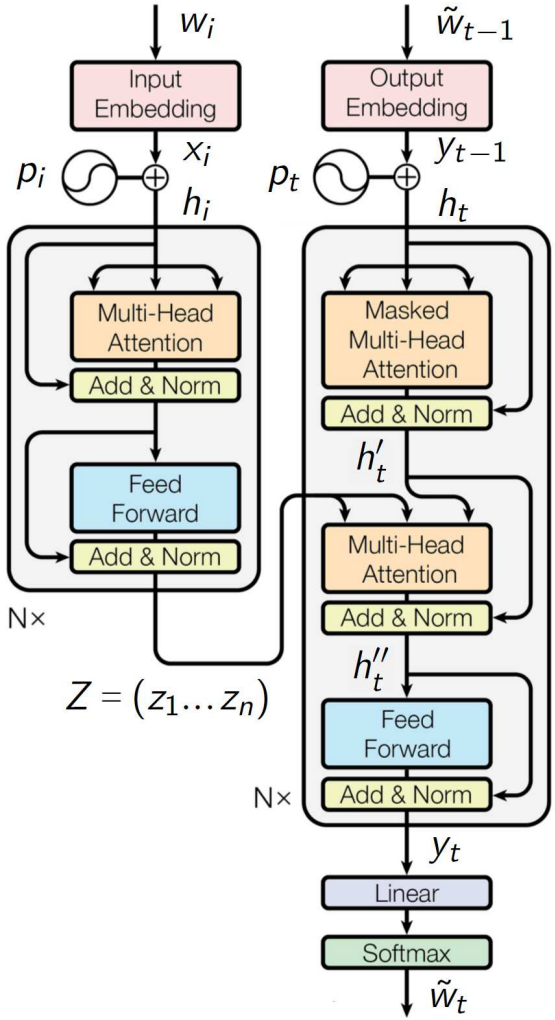


Архитектура трансформера декодировщика

Авторегрессионный синтез последовательности:
$y_0 = \langle \text{BOS} \rangle$ — вектор символа начала;
для всех $t = 1, 2, \ldots$:

1. Маскирование «данных из будущего»:
   $h_t = y_{t-1} + p_t; \quad H_t = (h_1, \ldots, h_t)$
2. Многомерное самовнимание:
   $h'_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(W^j_q h_t, W^j_k H_t, W^j_v H_t)$
3. Многомерное внимание на кодировку Z:
   $h''_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(\tilde{W}^j_q h'_t, \tilde{W}^j_k Z, \tilde{W}^j_v Z)$
4. Двухслойная полносвязная сеть:
   $y_t = \text{LN} \circ \text{FFN}(h''_t)$
5. Линейный предсказывающий слой:
   $p(\tilde{w}|t) = \text{SoftMax}_{\tilde{w}}(W_y y_t + b_y)$

генерация $\tilde{w}_t = \arg\max_{\tilde{w}} p(\tilde{w}|t)$ **пока** $\tilde{w}_t \neq \langle \text{EOS} \rangle$



Vaswani et al. (Google) Attention is all you need. 2017.

Критерии обучения и валидации для машинного перевода

Критерий для обучения параметров нейронной сети W по обучающей выборке предложений S с переводом $\tilde{S}$:

$$\sum_{(S,\tilde{S})} \sum_{\tilde{w}_t \in \tilde{S}} \ln p(\tilde{w}_t | t, S, W) \to \max_W$$

Критерии оценивания моделей (недифференцируемые) по выборке пар предложений «перевод S, эталон S_0»:

BiLingual Evaluation Understudy:

$$\text{BLEU} = \min\left(1, \frac{\Sigma \text{len}(S)}{\Sigma \text{len}(S_0)}\right) \underset{(S_0,S)}{\text{mean}} \left(\prod_{n=1}^{4} \frac{\#n\text{-грамм из } S, \text{ входящих в } S_0}{\#n\text{-грамм в } S} \right)^{\frac{1}{4}}$$

Word Error Rate:

$$\text{WER} = \underset{(S_0,S)}{\text{mean}} \left(\frac{\#\text{вставок} + \#\text{удалений} + \#\text{замен}}{\text{len}(S)} \right)$$

Vaswani et al. (Google) Attention is all you need. 2017.

BERT (Bidirectional Encoder Representations from Transformers)

Трансформер BERT — это кодировщик без декодировщика, предобучаемый на большой текстовой коллекции для решения широкого класса задач NLP

Схема преобразования данных в задачах NLP:

- $S = (w_1, \ldots, w_n)$ — токены предложения входного текста
  $\downarrow$ обучение эмбедингов вместе с трансформером
- $X = (x_1, \ldots, x_n)$ — эмбединги токенов входного предложения
  $\downarrow$ трансформер кодировщика
- $Z = (z_1, \ldots, z_n)$ — трансформированные эмбединги
  $\downarrow$ дообучение на конкретную задачу
- Y — выходной текст / разметка / классификация и т.п.

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language)
BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Критерий MLM (masked language modeling) для обучения BERT

Критерий маскированного языкового моделирования MLM, строится автоматически по текстам (self-supervised learning):

$$\sum_S \sum_{i \in M(S)} \ln p(w_i | i, S, W) \to \max_W,$$

где $M(S)$ — подмножество маскированных токенов из S,

$$p(w|i, S, W) = \underset{w \in V}{\text{SoftMax}}(W_z z_i(S, W_T) + b_z)$$

— языковая модель, предсказывающая i-й токен предложения S;

$z_i(S, W_T)$ — контекстный эмбединг i-го токена предложения S на выходе трансформера-кодировщика с параметрами W_T;

$W = (W_T, W_z, b_z)$ — все параметры языковой модели

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language) BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Критерий NSP (next sentence prediction) для обучения BERT

Критерий предсказания связи между предложениями NSP, строится автоматически по текстам (self-supervised learning):

$$\sum_{(S,S')} \ln p\big(y_{SS'}|S,S',W\big) \to \max_{W},$$

где $y_{SS'} = [\text{за } S \text{ следует } S']$ — классификация пары предложений,

$$p(y|S,S',W) = \underset{y\in\{0,1\}}{\text{SoftMax}}\big(W_y\,\text{th}(W_s z_0(S,S',W_T) + b_s) + b_y\big)$$

— вероятностная модель бинарной классификации пар (S,S'), $z_0(S,S',W_T)$ — контекстный эмбединг токена $\langle\texttt{CLS}\rangle$ для пары предложений, записанной в виде $\langle\texttt{CLS}\rangle\ S\ \langle\texttt{SEP}\rangle\ S'\ \langle\texttt{SEP}\rangle$

Jacob Devlin, Ming-Wei Chang, Kenton Lee, Kristina Toutanova (Google AI Language) BERT: pre-training of deep bidirectional transformers for language understanding. 2019.

Ещё несколько замечаний про трансформеры

- **Fine-tuning:** для дообучения на задаче задаётся модель $f(Z(S, W_T), W_f)$, выборка $\{S\}$ и критерий $\mathscr{L}(S, f) \to \max$
- **Multi-task learning:** для дообучения на наборе задач $\{t\}$ задаются модели $f_t(Z(S, W_T), W_t)$, выборки $\{S\}_t$ и сумма критериев $\sum_t \lambda_t \sum_S \mathscr{L}_t(S, f_t) \to \max$
- *GLUE*, *SuperGLUE*, *Russian SuperGLUE* — наборы тестовых задач на понимание естественного языка
- Трансформеры обычно строятся не на словах, а на токенах, получаемых BPE (Byte-Pair Encoding) или WordPiece
- Первый трансформер: $N = 6$, $d = 512$, $J = 8$, весов 65M
- $\text{BERT}_{\text{BASE}}$, GPT1: $N = 12$, $d = 768$, $J = 12$, весов 110M
- $\text{BERT}_{\text{LARGE}}$: $N = 24$, $d = 1024$, $J = 16$, весов 340M

Большие языковые модели научных текстов

- SciBERT (Beltagy et al., 2019)
- SPECTER (Cohan et al., 2020)
- MPNet (Song et al., 2020)
- LaBSE (Feng et al., 2020)
- SPECTER-2 (Singh et al., 2022)
- SciNCL (Ostendorff et al., 2022)
- mE5 (Wang et al., 2024)

I.Beltagy, K.Lo, A.Cohan. SciBERT: A pretrained language model for scientific text. 2019.
A.Cohan, S.Feldman, I.Beltagy, D.Downey, D.S.Weld. SPECTER: Document-level representation learning using citation-informed transformers. 2020.
K.Song et al. MPNet: Masked and permuted pre-training for language understanding. 2020.
F.Feng et al. Language agnostic BERT sentence embedding. 2020.
A.Singh, M.D'Arcy, A.Cohan, D.Downey, S.Feldman. SciRepEval: A multi-format benchmark for scientific document representations. 2022.
M.Ostendorff, N.Rethmeier, I.Augenstein, B.Gipp, G.Rehm. Neighborhood contrastive learning for scientific document representations with citation embeddings. 2022.
L.Wang et al. Multilingual E5 text embeddings: A technical report. 2024.

Мотивация нашего исследования

Модель должна быть применима в русскоязычных сервисах поиска, рекомендации, классификации, анализа научных публикаций: таких, как eLibrary.ru, «Мастерская знаний» и др.

Требования к модели:

- минимизация размера модели (23M параметров)
- при качестве, сопоставимом с лучшими (SOTA) моделями
- возможность вычисления эмбедингов без GPU
- мультиязычность: английский, русский,...
- возможность дообучения по данным о цитировании
- оценивание качества — по стандартным методикам

Данные для обучения и дообучения модели научных текстов

Данные для обучения:

- **S2ORC — Semantic Scholar Open Research Corpus**:
  205M публикаций, 121M авторов
  30M (12B токенов) отобрано для обучения модели,
  title+abstract, 85% на английском, 2% на русском
- **eLibrary**, заголовки и аннотации (title+abstract):
  8.6M (2B токенов) на русском
  8.8M (1.2B токенов) на английском

Данные для дообучения:

- **S2AG — Semantic Scholar Academic Graph**:
  источники: Crossref, PubMed, Unpaywall и др.
  2.5B связей цитирования

K.Lo, L.L.Wang, M.Neumann, R.Kinney, D.S.Weld. S2ORC: The semantic scholar open research corpus. 2019.

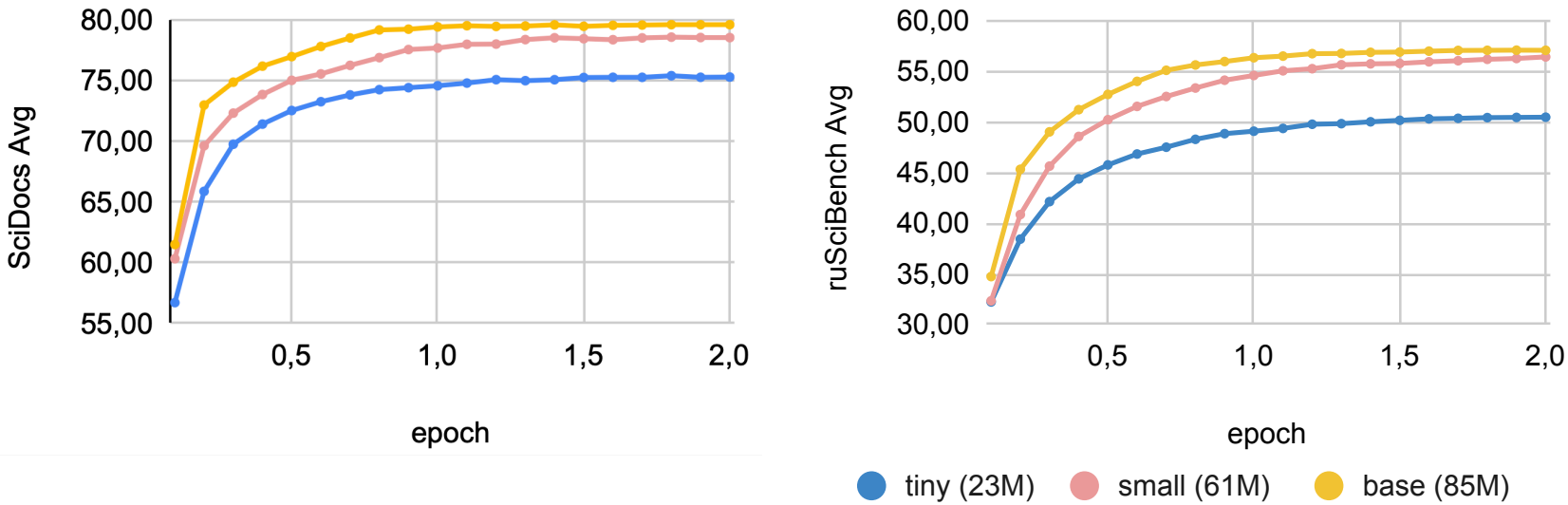
Методика оценивания и сравнения моделей (benchmark)

- **SciDocs: 6 задач**
  классификация статей по MeSH / по тематике,
  предсказание цитирования / ко-цитирования,
  пользовательской активности, рекомендации статей
- **SciRepEval: 24 задач**, вкл. SciDocs (кроме рекомендаций)
  классификация, регрессия, сходство, поиск,
  подбор рецензента для статьи,
  разрешение неоднозначности авторов
- **RuSciBench: 8 задач**
  классификация OECD/ГРНТИ по аннотации ru / en / ru+en
  поиск аннотации по её переводу ru→en / en→ru

A.Cohan, S.Feldman, I.Beltagy, D.Downey, D.S.Weld. SPECTER: Document-level representation learning using citation-informed transformers. 2020.
A.Singh, M.D'Arcy, A.Cohan, D.Downey, S.Feldman. SciRepEval: A multi-format benchmark for scientific document representations. 2022.
N.Gerasimenko, A.Vatolin. RuSciBench benchmark. 2023.
https://github.com/mlsa-iai-msu-lab/ru_sci_bench/tree/main

Этап 1: предобучение модели SciRus-tiny (MSU)

- Архитектура RoBERTa, случайная инициализация:
  tiny (23M, d:312), small (61M, d:768), base (85M, d:1024)
- критерий маскированного языкового моделирования MLM
- две эпохи обучения
- Avg — F1-мера, усреднённая по всем задачам бенчмарка



Y.Liu et al. RoBERTa: A robustly optimized BERT pretraining approach. 2019.

Этап 2: дообучение с контрастирующей функцией потерь

$(q, d) \in Q$ — заданное множество пар текстов,
которые должны иметь схожие векторные представления

$S(q, d; W) = \langle z(q, W), z(d, W) \rangle$ — модель сходства запроса q и документа d, где W — вектор параметров трансформера

Обучение с контрастированием (contrastive learning) — это максимизация правдоподобия для вероятностной модели бинарной классификации пар текстов на схожие и несхожие:

$$\sum_{(q,d)\in Q} \log \frac{e^{S(q,d;W)}}{e^{S(q,d;W)} + \sum_{\{\bar{d}\}} e^{S(q,\bar{d};W)}} \to \max_W$$

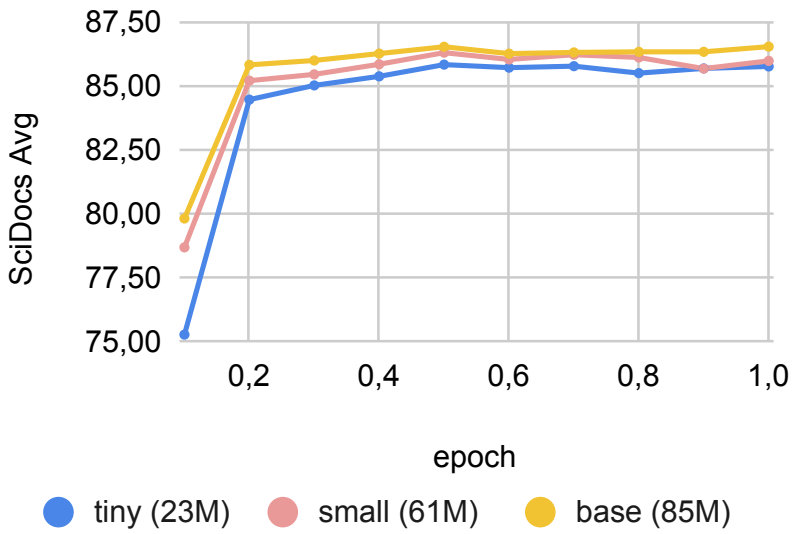
где $\{\bar{d}\}$ — множество «негативных примеров», несхожих с d, обычно это тексты, сэмплируемые случайно (negative sampling)

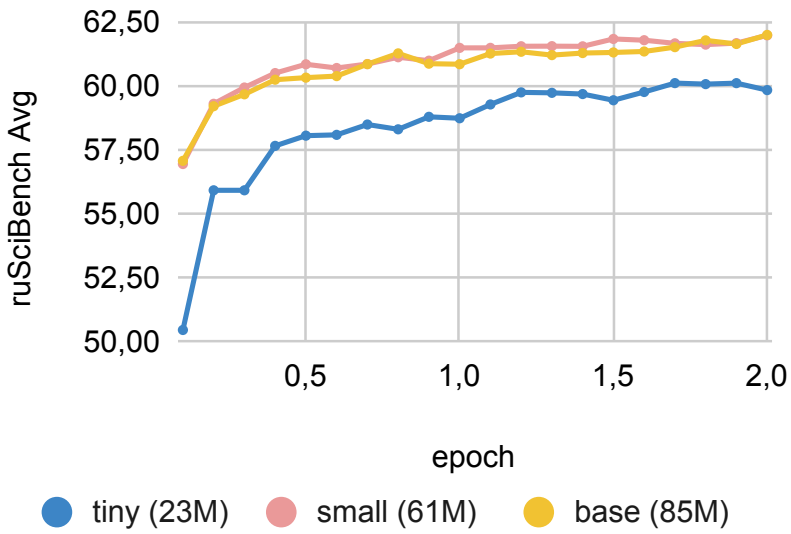
L.Wang et al. Text embeddings by weakly-supervised contrastive pretraining. 2022.

Этап 2: дообучение на парах title-abstract

(q, d) — пары текстов «название–аннотация»:

- 30.6M пар из S2AG
- 17.8M пар из eLibrary





Контрастирующее дообучение улучшает среднее качество

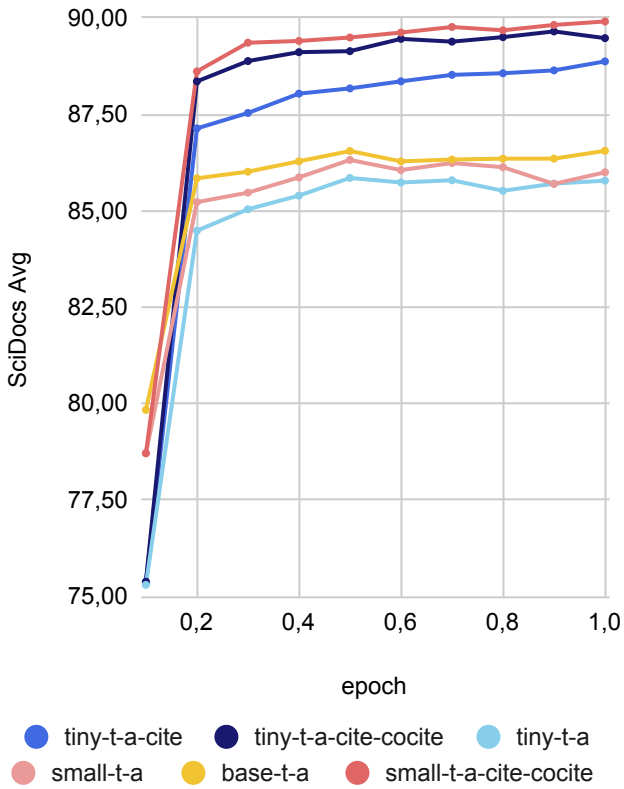
L.Wang et al. Text embeddings by weakly-supervised contrastive pretraining. 2022.

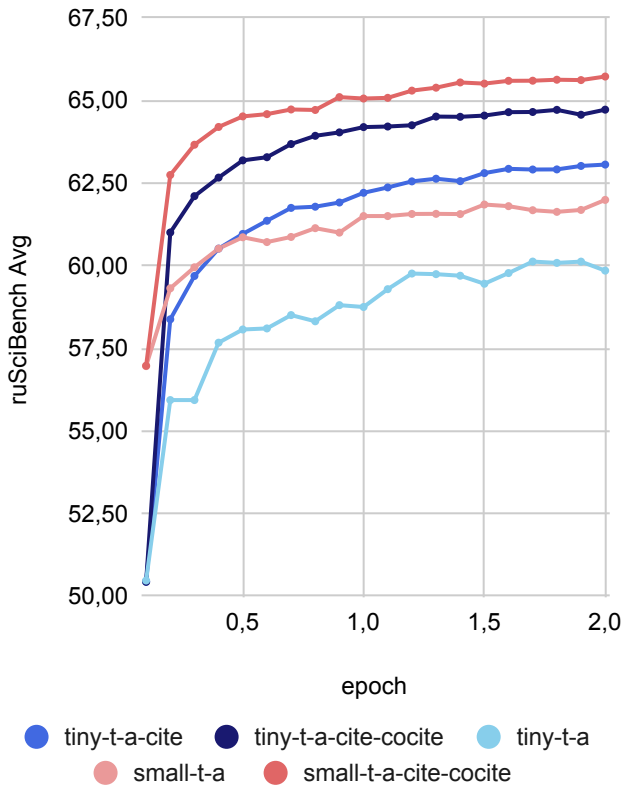
Этап 3: дообучение на парах cite-cocite

cite (q, d) — статья q цитирует статью d

co-cite (q, d) — третья статья q' цитирует статьи q и d

- 13.3M пар cite и 62M пар co-cite из S2AG
- 40M пар cite и 33.7M пар co-cite из eLibrary





Сравнение моделей по метрикам ruSciBench

oecd — качество (macro-F1) классификации en+ru рубрикатору
ru→en, en→ru — качество поиска (recall@1) перевода

модель	размер	oecd	ru→en	en→ru
e5-mistral-7b-instruct	7.11B	67.28	3.65	18.11
multilingual-e5-large	560M	63.70	99.19	99.37
scirus-tiny v3	**23M**	**61.13**	**94.83**	**95.81**
scirus-tiny v2	**23M**	**60.86**	**96.70**	**95.11**
multilingual-e5-base	278M	62.00	97.00	98.00
LaBSE	471M	60.21	98.31	97.20
LaBSE-en-ru	128M	60.05	98.26	96.93
paraphrase-multilingual-mpnet-base-v2		60.03	66.33	78.18
FRED-T5-large	360M	59.80	22.25	0.79
distiluse-base-multilingual-cased-v1		58.69	92.04	90.83
paraphrase-multilingual-MiniLM-L12-v2		56.48	72.87	77.49
mfaq		54.84	86.75	90.11
scirus-tiny v1	**23M**	**54.83**	**88.00**	**88.00**

Сравнение моделей по метрикам SciDocs

Avg — среднее качество по всем задачам бенчмарка

модель	размер	Avg
all-mpnet-base-v2	110M	91.03
scincl	110M	90.84
scirus-tiny3	**23M**	**90.10**
e5-large-v2	335M	88.70
e5-base	109M	88.58
multilingual-e5-large	560M	87.53
e5-small-v2	33.4M	86.99
multilingual-e5-base	278M	86.91
e5-mistral-7b-instruct 4byte	7.11B	86.03
scirus-tiny2	**23M**	**84.21**
sentence-transformers/LaBSE	471M	80.78
e5-pretrain-longer-240000-similarity-step-5581	23M	80.51
cointegrated/rubert-tiny2	29.4M	71.60
allenai/scibert-scivocab-uncased	110M	69.04
scirus-tiny	**23M**	**67.92**

Выводы по результатам сравнения моделей

- Размер и качество модели в сравнении с SciNCL:
  — меньше параметров: 23M против 110M
  — меньше размерность: 312 против 768
  — больше контекст: 1024 против 512
  — сопоставимое качество (SciDocs Avg): 90.10 против 90.84
- Контрастивное дообучение на парах title-abstract
  — существенно улучшает метрики качества,
  — особенно качество кросс-языкового поиска
  — компенсирует недостаточность данных о цитировании
- Контрастивное дообучение на парах cite / cocite
  — компенсирует недостаточность кросс-язаковых данных

Н.Герасименко, А.Ватолин, А.Янина, К.Воронцов. «Маленькая большая языковая модель для обработки научных текстов». 2024. (на рецензировании)

Внедрение в поисковый сервис eLibrary

«Разработанная в рамках данного проекта модель уже широко используется в Научной электронной библиотеке для решения целого ряда задач, связанных с оценкой тематической близости научных документов. Уже протестирован специалистами полезный сервис для ученых, позволяющий для заданной статьи или подборки статей найти тематически похожие документы, как среди всего массива eLIBRARY.RU (более 55 млн. научных публикаций), так и только среди новых поступлений. Важной для нас особенностью данной модели является ее мультиязычность, поскольку Научная электронная библиотека содержит документы на различных языках.»

— Геннадий Еременко, генеральный директор НЭБ

Научная электронная библиотека, портал eLIBRARY.RU.
Пресс-релиз 24-04-2024: Открыт поиск близких по тематике публикаций с применением нейросети МГУ для анализа научных текстов.
https://elibrary.ru/projects/news/search_similar_publ.asp