

Модели векторизации текста для семантического поиска и формализации гуманитарных знаний

Воронцов Константин Вячеславович

voron@mlsa-iai.ru

д.ф.-м.н., профессор РАН, зав. каф. ММП ВМК МГУ,
зав. лаб. Машинного интеллекта и семантического анализа



Международная молодежная научная конференция
«Технологии ИИ в науке и образовании» • 8–9 декабря 2023

- 1 Задачи понимания естественного языка**
 - эволюция подходов в обработке текстов
 - большие предобученные языковые модели
 - чем GPT отличается от всего, что было раньше
- 2 Проект «Мастерская знаний»**
 - поисково-рекомендательный сервис
 - пошаговое полуавтоматическое реферирование
 - тематическое моделирование и визуализация
- 3 Проект «Новостной коллаидер»**
 - задачи детекции постправды
 - технологический конкурс ПРО//ЧТЕНИЕ
 - задачи автоматизации разметки текста

Эволюция подходов машинного обучения в анализе текстов

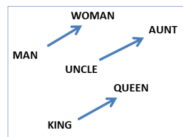
Декомпозиция задач по уровням пирамиды NLP

- 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]

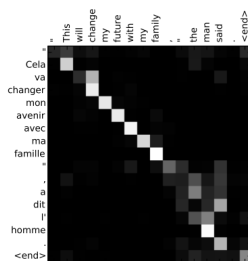
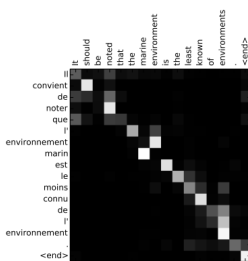
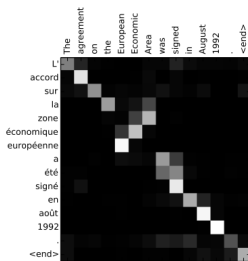


Нейросетевые модели локальных контекстов

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

$$\text{softmax} \left(\frac{\begin{matrix} Q \\ \text{grid} \end{matrix} \times \begin{matrix} KV \\ \text{grid} \end{matrix}}{\sqrt{d}} \right) \begin{matrix} V \\ \text{grid} \end{matrix}$$

Модели внимания для машинного перевода



Вход: $\{x_i\}$ — последовательность слов входного языка

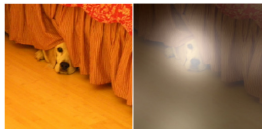
Выход: $\{y_t\}$ — последовательность слов выходного языка

Интерпретация модели внимания: матрица семантического сходства A_{ti} показывает, на какие слова x_i входного текста модель обращает внимание, генерируя слово перевода y_t

Модели внимания для аннотирования изображений



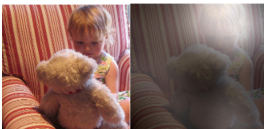
A woman is throwing a frisbee in a park.



A dog is standing on a hardwood floor.



A stop sign is on a road with a mountain in the background.



A little girl sitting on a bed with a teddy bear.



A group of people sitting on a boat in the water.



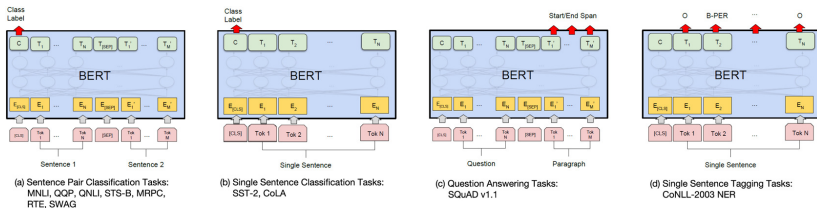
A giraffe standing in a forest with trees in the background.

Интерпретация: на какие области модель обращает внимание, генерируя подчёркнутое слово в описании изображения

Kelvin Xu et al. Show, attend and tell: neural image caption generation with visual attention. 2016

Большие пред-обученные модели языка (трансформеры)

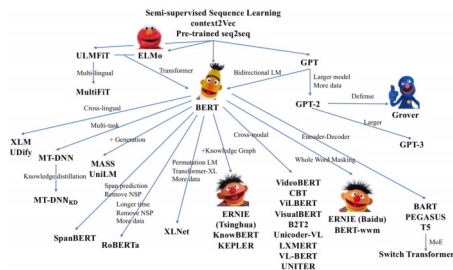
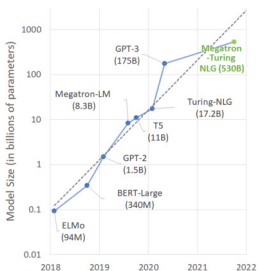
- обучены предсказывать слово по его контексту
- обучены по терабайтам текстов, «они видели в языке всё»
- способны выделять и классифицировать фрагменты текста, генерировать фейковые тексты, не отличимые от реальных
- *мультиязычны*: обучаются на десятках языков
- *мультизадачны*: для каждой новой задачи NLP/NLU достаточно дообучения на малой размеченной выборке



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

Рост больших языковых моделей — быстрее закона Мура

Рост числа параметров в трансформерных моделях языка



- *трансформер-кодировщик* преобразует последовательность слов в числовые векторы, зависящие от контекста
- *трансформер-декодировщик* преобразует векторную последовательность в последовательность слов

Число параметров сети сопоставимо с объёмом исходных данных

ChatGPT-4: проблески общего искусственного интеллекта

Sparks of Artificial General Intelligence: Early experiments with GPT-4

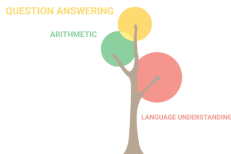
Sébastien Bubeck Varun Chandrasekaran Ronen Eldan Johannes Gehrke
Eric Horvitz Ece Kamar Peter Lee Yin Tat Lee Yuanzhi Li Scott Lundberg
Harsha Nori Hamid Palangi Marco Tulio Ribeiro Yi Zhang

Microsoft Research (27 March 2023)

Новые способности модели, не закладывавшиеся при обучении:

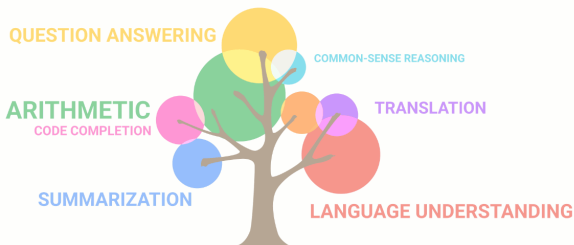
- объяснять свои ответы, перефразировать
- реферировать, генерировать планы, сценарии, шаблоны
- переводить на другие языки, строить аналогии, менять тональность, стиль, глубину изложения
- генерировать программный код на различных языках
- решать некоторые логические и математические задачи
- искать и исправлять собственные ошибки по подсказке

Эмерджентность — появление качественно новых способностей



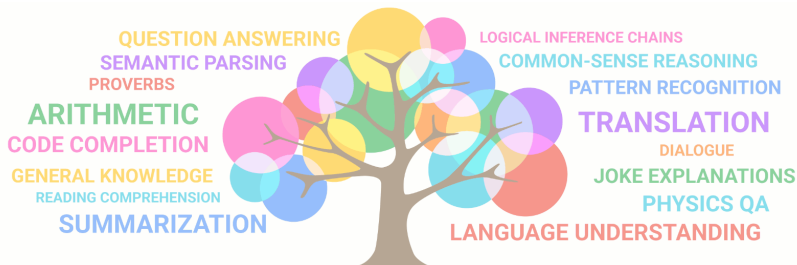
- GPT-2: 14/Feb/2019, контекст 768 слов (1,5 страницы)
- 1,5 млрд. параметров, корпус 10 млрд. токенов (40Gb)
- способность написать эссе, которое конкурсное жюри не смогло отличить от написанного человеком

Эмерджентность — появление качественно новых способностей



- GPT-3: 11/Jun/2020, контекст 1536 слов (3 страницы)
- 175 млрд. параметров, корпус 500 млрд. токенов
- способность делать перевод на другие языки,
- решать логические и математические задачи,
- генерировать программный код по описанию

Эмерджентность — появление качественно новых способностей



- GPT-4: 14/Mar/2023, контекст 24 000 слов (48 страниц)
- >1 трл. параметров, корпус >1Tb
- способность описывать и анализировать изображения,
- реагировать на подсказки вроде «Let's think step by step»,
- решать качественные физические задачи по картинке

Концепция проекта «Мастерская знаний»

«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*



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*
- выделение «важных мест» в документе из *подборки*

Коммуникативные функции:

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

Прототип поисково-рекомендательной системы

Тематическая подборка пользователя:

https://arxiv.aitheta.com/collections/Q29sbGVjYjdlGjVjZUFTUEF5aHBH

FEEDS | SEARCH | **COLLECTIONS** | About | FAQ | Konstantin Vorontsov

MOOC (massive open online course)

PAPERS | RECOMMENDED

19 JUL 2014
Towards Feature Engineering at Scale for Data from Massive Open Online Courses
Kalyan Veeramachaneni, Una-May O'Reilly, Colin Taylor

We examine the process of engineering features for developing models that improve our understanding of learners' online behavior in MOOCs. Because feature engineering relies so heavily on human insight, we argue that extra effort should be made to engage the crowd for feature proposals and even their operationalization. We show two approaches where we have started to engage the crowd. We also show how features can be evaluated for their relevance in predictive accuracy. When we...

Citations: 6

2 JUL 2017
Reciprocal Recommender System for Learners in Massive Open Online Courses (MOOCs)
Sankalp Prabhakar, Gerasimos Spanakis, Osmar Zaiane

Massive open online courses (MOOC) describe platforms where users with completely different backgrounds subscribe to various courses on offer. MOOC forums and discussion boards offer learners a medium to communicate with each other and maximize their learning outcomes. However, oftentimes learners are hesitant to approach each other for different reasons (being shy, don't know the right match, etc.). In this paper, we propose a reciprocal recommender system which matches...

Citations: 0

Прототип поисково-рекомендательной системы

Список статей, рекомендуемых для добавления в подборку:

← → ↻ https://arxiv.aitha.com/collections/Q29sbGVjZGljbjJzUjVTVUEFsaHBH

FEEDS | SEARCH | COLLECTIONS | About | FAQ | Konstantin Vorontsov

MOOC (massive open online course)

PAPERS → RECOMMENDED

2 JUN 2019

A Survey of Natural Language Generation Techniques with a Focus on Dialogue Systems - Past, Present and Future Directions

Sashank Santhanam, Samira Shaikh

One of the hardest problems in the area of Natural Language Processing and Artificial Intelligence is automatically generating language that is coherent and understandable to humans. Teaching machines how to converse as humans do falls under the broad umbrella of Natural Language Generation. Recent years have seen unprecedented growth in the number of research articles published on this subject in conferences and journals both by academic and industry researchers. There have...

Citations: 6

🔖 👍 🔄

20 SEP 2014

Capturing "attrition intensifying" structural traits from didactic interaction sequences of MOOC learners

Tanmay Sinha, Nan Li, Patrick Jermann, Pierre Dillenbourg

This work is an attempt to discover hidden structural configurations in learning activity sequences of students in Massive Open Online Courses (MOOCs). Leveraging combined representations of video clickstream interactions and forum activities, we seek to fundamentally understand traits that are predictive of decreasing engagement over time. Grounded in the interdisciplinary field of network science, we follow a graph based approach to successfully extract indicators of active and...

Citations: 0

🔖 👍 🔄

Прототип поисково-рекомендательной системы

Добавление статьи из списка рекомендаций в подборку:

The screenshot shows a web browser window displaying a search results page for 'MOOC (massive open online course)'. The page has a dark blue header with navigation links: FEEDS, SEARCH, COLLECTIONS, About, FAQ, and Konstantin Vorontsov. The main content area is titled 'MOOC (massive open online course)' and lists several papers under the 'PAPERS' section. The first paper is 'A Survey of Natural Language Generation T...' by Sashank Santhanam and Samira Shaikh, dated 2 JUN 2019. A red circle highlights the bookmark icon for this paper. A modal dialog box titled 'Add to collections' is open over the paper, showing a list of collection options: Exploratory Search, MOOC (massive open online course) (selected), Opinion Mining and Sentiment Analysis with Topic Modeling, Textual Complexity and Readability, and Topic modeling of genomic data. A red circle highlights the 'MOOC (massive open online course)' option. Below the list is a blue 'SAVE CHANGES' button, also circled in red. A 'NEW COLLECTION' link is visible at the bottom of the dialog. On the right side of the page, a 'RECOMMENDED' section is visible, with the word 'RECOMMENDED' circled in red. The overall interface is clean and modern, with a focus on user interaction and recommendation management.

Полуавтоматическое реферирование тематических подборок

Рекомендации фраз для реферата с помощью суфлёров:

The screenshot displays a web application interface for paper summarization and recommendation, organized into three main columns: PAPERS, RECOMMENDED, and SUMMARIZATION.

- PAPERS:** A list of papers is shown. The selected paper is "SummaRuNNer: A Recurrent Neural Network based..." by Ramesh Nallapati, Feifei Zhai, and Bowen Zhou (13 NOV 2016).
- RECOMMENDED:** A summary of the selected paper is displayed. It describes a novel method for training neural networks for extractive summarization, called BANDITSUM, and mentions the use of a policy gradient reinforcement learning algorithm.
- SUMMARIZATION:** Recommended phrases are shown, such as "SummaRuNNer, a Recurrent Neural Network (RNN) based sequence model for extractive summarization of documents and show that it achieves performance better than or comparable to state-of-the-art."

At the bottom, a "Promoters" bar allows users to select categories for the summary, including Annotate, Idea, Theory, Method, Citation, Dataset, Experiment, Result, and Conclusion. Red arrows indicate the flow of information from the selected paper to the summary and then to the recommended phrases.

А.Власов. Методы полуавтоматической суммаризации подборок научных статей. Магистерская диссертация, МФТИ, 2020.

С.Крыжановская. Технология полуавтоматической суммаризации тематических подборок научных статей. Магистерская диссертация, ВМК МГУ, 2022.

Концепция MAHS (Machine Aided Human Summarization)

- 1 Система рекомендует *сценарий реферата* — список статей **подборки**, ранжированный в порядке упоминания
- 2 Пользователь может скорректировать сценарий в соответствии со своими целями и творческим замыслом
- 3 В цикле по ранжированному списку статей **подборки**:
 - **пользователь** запрашивает аспекты статьи у суфлёров: «как другие авторы ссылаются на эту статью», «цель», «идея», «подход», «достижение», «недостаток», «результат», «вывод» и т.д.
 - **суфлёр** выдаёт ранжированный список найденных фраз
 - **пользователь** добавляет фразу из поисковой выдачи и корректирует её в соответствии с целями и замыслом

А.Власов. Методы полуавтоматической суммаризации подборок научных статей. Магистерская диссертация, МФТИ, 2020.

С.Крыжановская. Технология полуавтоматической суммаризации тематических подборок научных статей. Магистерская диссертация, ВМК МГУ, 2022.

Полуавтоматическое реферирование тематических подборок

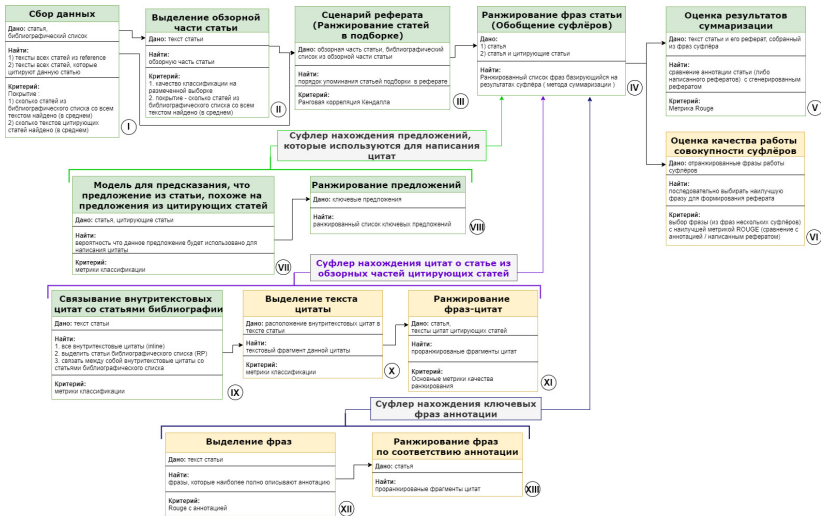
Задачи машинного обучения для МАНС:

- 1 Формирование обучающей выборки: paper \rightarrow (refs, survey)
- 2 Ранжирование статей подборки для сценария реферата
- 3 Выбор релевантных фраз из текста статьи для сфлёрра
- 4 Ранжирование выбранных фраз для каждого сфлёрра
- 5 Выбор начала и конца контекста фразы, в частности, выбор релевантного контекста вокруг ссылки:

Few contextual citation graphs are publicly available. The ACL Anthology Network (AAN) (Radev et al., 2009) is one such contextual citation graph built from the ACL Anthology corpus (Bird et al., 2008), consisting of 24.6K papers manually augmented with citation information. CiteSeer (Giles et al., 1998) provides a large corpus consisting of 1.0M papers with full text and bibliography entries parsed from PDFs. Saier and Farber (2019) introduces a contextual citation graph of approximately 1.0M arXiv papers with full text LaTeX parses where citations are linked to papers in the Microsoft Academic Graph.

M. Yasunaga et al. ScisummNet: A large annotated corpus and content-impact models for scientific paper summarization with citation networks. 2019.

Систематизация задач машинного обучения для MAHS



Тематическое моделирование: «о чём все эти тексты?»

Дано:

- коллекция текстовых документов

Найти:

- T — множество тем, составляющих эту коллекцию
- $p(w|t) = \varphi_{wt}$ — вероятности слов w в каждой теме t
- $p(t|d) = \theta_{td}$ — вероятности тем t в каждом документе d
- $p(w|d) = \sum_{t \in T} \varphi_{wt} \theta_{td}$ — вероятностная тематическая модель

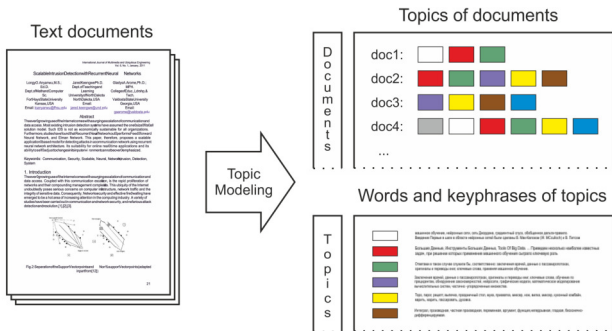
Критерий: правдоподобие предсказания слов w в документах d с дополнительными критериями-регуляризаторами $R_i(\Phi, \Theta)$:

$$\sum_{d \in D} \sum_{w \in d} \ln \sum_{t \in T} \varphi_{wt} \theta_{td} + \sum_i \tau_i R_i(\Phi, \Theta) \rightarrow \max_{\Phi, \Theta}$$

Воронцов К.В. Вероятностное тематическое моделирование: теория регуляризации ARTM и библиотека с открытым кодом BigARTM. URSS, 2023. ISBN 978-5-9519-4345-3.

Мультимодальная тематическая модель

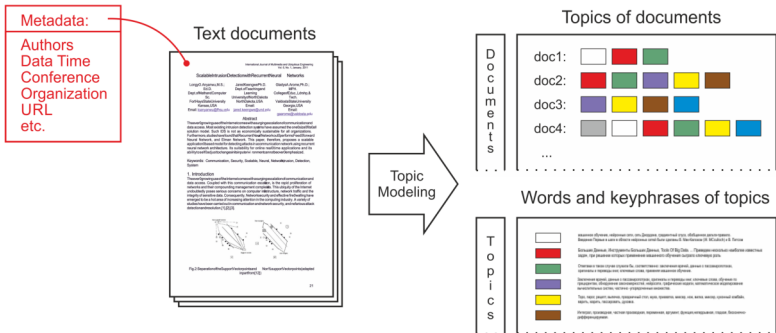
Тема t может содержать термины различных *модальностей*:
 $p(\text{слово} | t)$, $p(n\text{-грамма} | t)$,



Мультимодальная тематическая модель

Тема t может содержать термины различных модальностей:

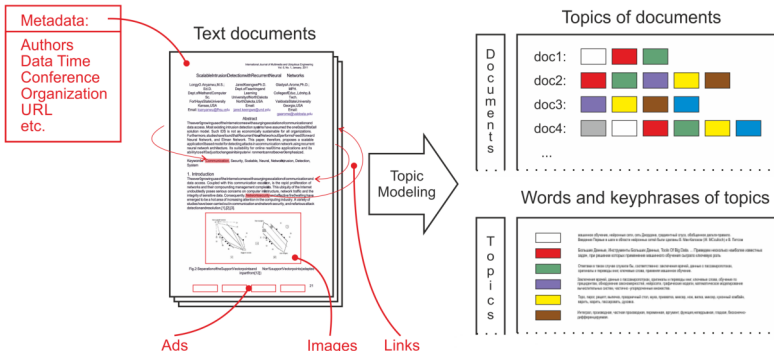
$p(\text{слово} | t)$, $p(n\text{-грамма} | t)$, $p(\text{автор} | t)$, $p(\text{время} | t)$, $p(\text{источник} | t)$,



Мультимодальная тематическая модель

Тема t может содержать термины различных модальностей:

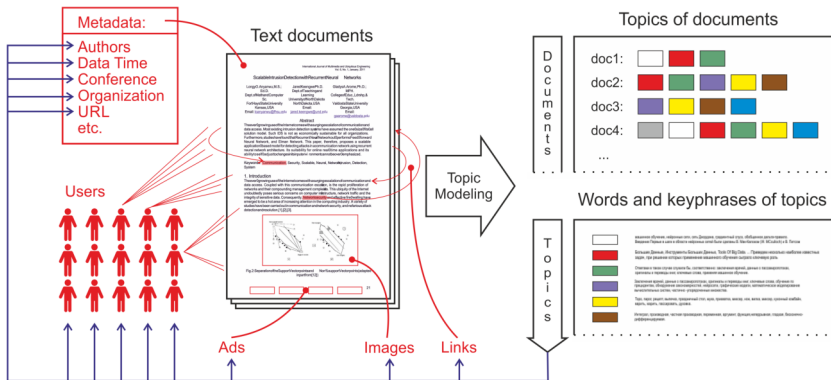
$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
 $p(\text{объект}|t)$, $p(\text{название}|t)$, $p(\text{ссылка}|t)$,



Мультимодальная тематическая модель

Тема t может содержать термины различных модальностей:

$p(\text{слово}|t)$, $p(n\text{-грамма}|t)$, $p(\text{автор}|t)$, $p(\text{время}|t)$, $p(\text{источник}|t)$,
 $p(\text{объект}|t)$, $p(\text{название}|t)$, $p(\text{ссылка}|t)$, $p(\text{пользователь}|t)$



Пример 1. Мультиязычная тематическая модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их частоты $p(w|t)$ в %:

Тема №68				Тема №79			
research	4.56	институт	6.03	goals	4.48	матч	6.02
technology	3.14	университет	3.35	league	3.99	игрок	5.56
engineering	2.63	программа	3.17	club	3.76	сборная	4.51
institute	2.37	учебный	2.75	season	3.49	фк	3.25
science	1.97	технический	2.70	scored	2.72	против	3.20
program	1.60	технология	2.30	cup	2.57	клуб	3.14
education	1.44	научный	1.76	goal	2.48	футболист	2.67
campus	1.43	исследование	1.67	apps	1.74	гол	2.65
management	1.38	наука	1.64	debut	1.69	забивать	2.53
programs	1.36	образование	1.47	match	1.67	команда	2.14

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

K. Vorontsov, O. Frej, M. Apishev, P. Romov, M. Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 1. Мультиязычная тематическая модель Википедии

216 175 русско-английских пар статей. Языки — модальности.
Первые 10 слов и их частоты $p(w|t)$ в %:

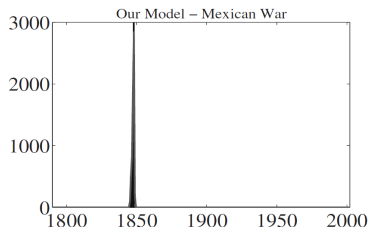
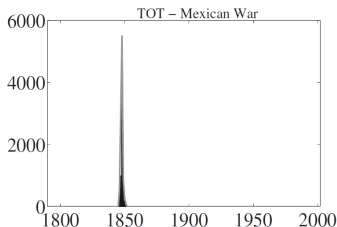
Тема №88				Тема №251			
opera	7.36	опера	7.82	windows	8.00	windows	6.05
conductor	1.69	оперный	3.13	microsoft	4.03	microsoft	3.76
orchestra	1.14	дирижер	2.82	server	2.93	версия	1.86
wagner	0.97	певец	1.65	software	1.38	приложение	1.86
soprano	0.78	певица	1.51	user	1.03	сервер	1.63
performance	0.78	театр	1.14	security	0.92	server	1.54
mozart	0.74	партия	1.05	mitchell	0.82	программный	1.08
sang	0.70	сопрано	0.97	oracle	0.82	пользователь	1.04
singing	0.69	вагнер	0.90	enterprise	0.78	обеспечение	1.02
operas	0.68	оркестр	0.82	users	0.78	система	0.96

Ассессор оценил 396 тем из 400 как хорошо интерпретируемые.

K. Vorontsov, O. Frej, M. Apishev, P. Romov, M. Suvorova. BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections. AIST-2015.

Пример 2. Совмещение темпоральной и n -граммной модели

По коллекции выступлений президентов США



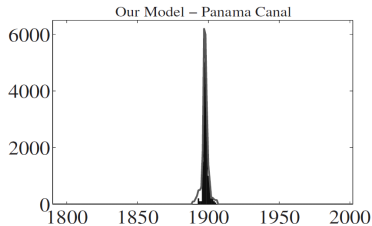
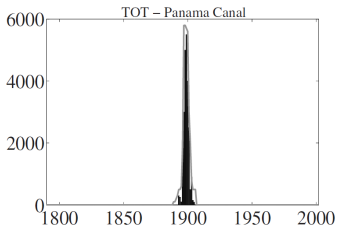
1. mexico	8. territory
2. texas	9. army
3. war	10. peace
4. mexican	11. act
5. united	12. policy
6. country	13. foreign
7. government	14. citizens

1. east bank	8. military
2. american coins	9. general herrera
3. mexican flag	10. foreign coin
4. separate independent	11. military usurper
5. american commonwealth	12. mexican treasury
6. mexican population	13. invaded texas
7. texan troops	14. veteran troops

Shoaib Jameel, Wai Lam. An N -Gram Topic Model for Time-Stamped Documents. ECIR 2013.

Пример 2. Совмещение темпоральной и n -граммной модели

По коллекции выступлений президентов США



1. government	8. spanish
2. cuba	9. island
3. islands	10. act
4. international	11. commission
5. powers	12. officers
6. gold	13. spain
7. action	14. rico

1. panama canal	8. united states senate
2. isthmian canal	9. french canal company
3. isthmus panama	10. caribbean sea
4. republic panama	11. panama canal bonds
5. united states government	12. panama
6. united states	13. american control
7. state panama	14. canal

Shoaib Jameel, Wai Lam. An N-Gram Topic Model for Time-Stamped Documents. ECIR 2013.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Большие данные: коллекция не хранится в памяти
- Онлайн-овый параллельный мультимодальный ARTM
- Встроенная библиотека регуляризаторов и метрик качества

Сообщество:

- Открытый код <https://github.com/bigartm>
(discussion group, issue tracker, pull requests)
- Документация <http://bigartm.org>



Лицензия и среда разработки:

- Свободная коммерческая лицензия (BSD 3-Clause)
- Кросс-платформенность: Windows, Linux, MacOS (32/64 bit)
- Интерфейсы API: command-line, C++, and Python

Качество и скорость: BigARTM vs Gensim и Vowpal Wabbit

3.7М статей Википедии, 100К слов:

время min (перплексия)

проц.	T	Gensim	Vowpal Wabbit	BigARTM	BigARTM асинхрон
1	50	142m (4945)	50m (5413)	42m (5117)	25m (5131)
1	100	287m (3969)	91m (4592)	52m (4093)	32m (4133)
1	200	637m (3241)	154m (3960)	83m (3347)	53m (3362)
2	50	89m (5056)		22m (5092)	13m (5160)
2	100	143m (4012)		29m (4107)	19m (4144)
2	200	325m (3297)		47m (3347)	28m (3380)
4	50	88m (5311)		12m (5216)	7m (5353)
4	100	104m (4338)		16m (4233)	10m (4357)
4	200	315m (3583)		26m (3520)	16m (3634)
8	50	88m (6344)		8m (5648)	5m (6220)
8	100	107m (5380)		10m (4660)	6m (5119)
8	200	288m (4263)		15m (3929)	10m (4309)

D.Kochedykov, M.Apishev, L.Golitsyn, K.Vorontsov.

Fast and Modular Regularized Topic Modelling. FRUCT ISMW, 2017.

Разведочный поиск в технологических блогах

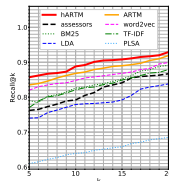
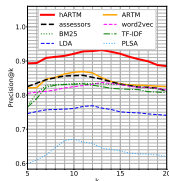
Цель: поиск документов
по длинным текстовым запросам
— Habr.ru (175К документов),
— TechCrunch.com (760К док.).

Регуляризаторы:

$$\mathcal{L} \left(\begin{array}{|c|} \hline \text{PLSA} \\ \hline \Phi \quad \Theta \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{hierarchy} \\ \hline \text{graph} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{interpretable} \\ \hline \text{matrix} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{multimodal} \\ \hline \text{img} \quad \text{text} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{n-gram} \\ \hline \text{grid} \\ \hline \end{array} \right) \rightarrow \max$$

Результаты:

- Точность и полнота **93%**, превосходит ассессоров и другие методы (tf-idf, BM25, word2vec, PLSA, ARTM).
- Увеличилась оптимальная размерность векторов:
200 → 1400 (Habr.ru), 475 → 2800 (TechCrunch.com).



A. Ianina, K. Vorontsov. Regularized multimodal hierarchical topic model for document-by-document exploratory search. FRUCT-ISMW, 2019.

Поиск и классификация этно-релевантных тем в соцсетях

Цель: выявление как можно большего числа тем о национальностях и межнациональных отношениях (затравка — словарь 300 этнонимов).

Регуляризаторы:

$$\mathcal{L} \left(\begin{array}{|c|} \hline \text{PLSA} \\ \hline \Phi \quad \Theta \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{seed words} \\ \hline \text{[Bar chart]} \quad \square \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{interpretable} \\ \hline \text{[Bar chart]} \quad \text{[Scatter plot]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{multimodal} \\ \hline \text{[Image]} \quad \square \\ \hline \end{array} \right) \\ + R \left(\begin{array}{|c|} \hline \text{temporal} \\ \hline \text{[Line graph]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{geospatial} \\ \hline \text{[Map]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{sentiment} \\ \hline \text{[Sentiment scale]} \\ \hline \end{array} \right) \rightarrow \max$$

Результаты: число релевантных тем: 45 (LDA) \rightarrow 83 (ARTM).

M. Apishev, S. Koltcov, O. Koltsova, S. Nikolenko, K. Vorontsov. Additive regularization for topic modeling in sociological studies of user-generated text content. MICAI, 2016.

Mining ethnic content online with additively regularized topic models. 2016.

[японцы] японский, япония, япон, корей, китайский, жилища, азия, фукусима, цунами, сакура, бики, слиши, озон, рабон, нана, гласко, диньши,
[иностранцы] дитя, ребенок, родился, детский, семья, воспитаный, певца, возраст, отец, воспитаный, нидерландский, родителский, родител, мальчик, взрослый, отец, сын,
[американцы] айба, колора, инвестор, член, президент, ит, индур, бонжон, фидель, гласко, катанский, вивальдийский, лидер, болгарская, нидерландский, зальме, лидер,
[китайцы] китайский, россия, производство, китай, продукция, страна, предприятие, компания, технология, военный, регион, производство, промышленный, экономика, россия, российский, экономика, мир,
[азербайджанцы] русский, азербайджан, азербайджанец, росси, азербайджанский, тикси, диспоза, анал, жарод, москва, страна, землянк, слово, рынок,
[германы] германский, спецназ, военный, август, батальон, российский, спецназ, министр, операция, рунч, братва, микрофинансовый, абскал, группа, война, русский, цинвале,
[осетины] конституция, осетия, азиат, русский, осетинский, цинвал, северный, россия, жаб, республика, мирот, алака, республика, «осетины», компания, (азиат) переклет, алака, шатка, ларши, место, страна, деньги, время, работа, жизнь, дух, дин, цинвалский, нарицательное,

Аналогичные исследования по выделению узкой тематики

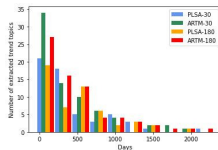
Задачи «поиска и классификации иголок в стоге сена»

- поиск и кластеризация новостей [1]
- поиск в социальных медиа информации, связанной с болезнями, симптомами и методами лечения [2]
- поиск чатов, связанных с преступностью и экстремизмом [3, 4]
- поиск выступлений о правах человека в ООН [5]

-
1. *J.Jagarlamudi, H.Daumé III, R.Udupa*. Incorporating lexical priors into topic models. 2012.
 2. *M.Paul, M.Dredze*. Discovering health topics in social media using topic models. 2014.
 3. *M.A.Basher, A.Rahman, B.C.M.Fung*. Analyzing topics and authors in chat logs for crime investigation. 2014.
 4. *A.Sharma, M.Pawar*. Survey paper on topic modeling techniques to gain useful forecasting information on violant extremist activities over cyber space. 2015.
 5. *Kohei Watanabe, Yuan Zhou*. Theory-driven analysis of large corpora: semisupervised topic classification of the UN speeches. 2022.

Выявление трендов в коллекции научных публикаций

Цель: раннее обнаружение трендовых тем с начальным экспоненциальным ростом в области AI/ML 2009–2021 гг.



Регуляризаторы:

$$\mathcal{L} \left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{[Bar Chart Icon]} \quad \text{[Scatter Plot Icon]} \end{array} \right) + R \left(\begin{array}{c} \text{dynamic} \\ \text{[Line Graph Icon]} \end{array} \right) + R \left(\begin{array}{c} \text{multimodal} \\ \text{[Stacked Boxes Icon]} \quad \text{[Square Icon]} \end{array} \right) + R \left(\begin{array}{c} \text{n-gram} \\ \text{[Grid Icon]} \end{array} \right) \rightarrow \max$$

Результаты:

- выделение 90 из 91 тренда в области машинного обучения
- 63% тем выделяется за год, 79% за два года

Н.Герасименко, А.Чернявский, М.Никифорова, М.Никитин, К.Воронцов.

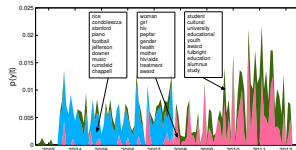
Инкрементальное обучение тематических моделей для поиска трендовых тем в научных публикациях. Доклады РАН, 2022.

Выявление динамики тем в новостных потоках

Цель: выделение тем в коллекции пресс-релизов МИДов 4х стран, с привязкой ко времени.

Регуляризаторы:

$$\mathcal{L} \left(\begin{array}{|c|} \hline \text{PLSA} \\ \hline \Phi \quad \Theta \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{interpretable} \\ \hline \text{[bar chart]} \quad \text{[scatter plot]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{temporal} \\ \hline \text{[line graph]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{multimodal} \\ \hline \text{[stacked bar]} \quad \text{[box]} \\ \hline \end{array} \right) \\ + R \left(\begin{array}{|c|} \hline \text{n-gram} \\ \hline \text{[grid]} \\ \hline \end{array} \right) + R \left(\begin{array}{|c|} \hline \text{multilanguage} \\ \hline \text{[stacked bar]} \quad \text{[box]} \\ \hline \end{array} \right) \rightarrow \max$$



Результаты:

- разделение тем на событийные и перманентные
- когерентность тем: 5.5 \rightarrow 6.5

Н. Дойков. Адаптивная регуляризация вероятностных тематических моделей.
ВКР бакалавра, ВМК МГУ, 2015.

Выделение поляризованных мнений в политических новостях

Цель: найти признаки, по которым
событийная тема разделяется
на кластеры-мнения

Modalities	Pr	Rec	F1
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	0.77	0.97	0.86

Регуляризаторы:

$$\mathcal{L} \left(\begin{array}{c} \text{PLSA} \\ \Phi \quad \Theta \end{array} \right) + R \left(\begin{array}{c} \text{interpretable} \\ \text{matrix} \end{array} \right) + R \left(\begin{array}{c} \text{multimodal} \\ \text{matrix} \end{array} \right) + R \left(\begin{array}{c} \text{n-gram} \\ \text{matrix} \end{array} \right) + R \left(\begin{array}{c} \text{syntax} \\ \text{tree} \end{array} \right) \rightarrow \max$$

Результаты:

- выделение мнений внутри тем: F1-мера = 0.86%
- совместное использование трёх модальностей:
 - факты как триплеты «субъект–предикат–объект»
 - семантические роли слов по Филлмору
 - тональности именованных сущностей

D.Feldman, T.Sadekova, K.Vorontsov. Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining. Dialogue 2020.

Выделение поляризованных мнений в политических новостях

... Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР "национализировали" украинские предприятия ... При этом Кремль защитил конфискацию предприятий в ЛДНР ... Украина потребует расширить санкции ... За все эти действия обязательно наступит наказание. Украина потребует расширения санкций на тех, кто украл украинские предприятия ... (*Kiev opinion*)

... По словам Захарченко, Киев встретит свой "ужасный конец" ... Киев возьмется за ум, и в целях спасения собственной промышленности снимет блокаду ... Обстановка, которую искусственно создала Украина с блокадой Донбасса, вынудила ... кошмарит свой народ ... если в Киеве были приняты какое-либо постановление ... положительные результаты, как в республиках, так и в России ... Если им удастся сместить Порошенко и при этом не развалить Украину то все вернется на свои места ... (*Moscow opinion*)

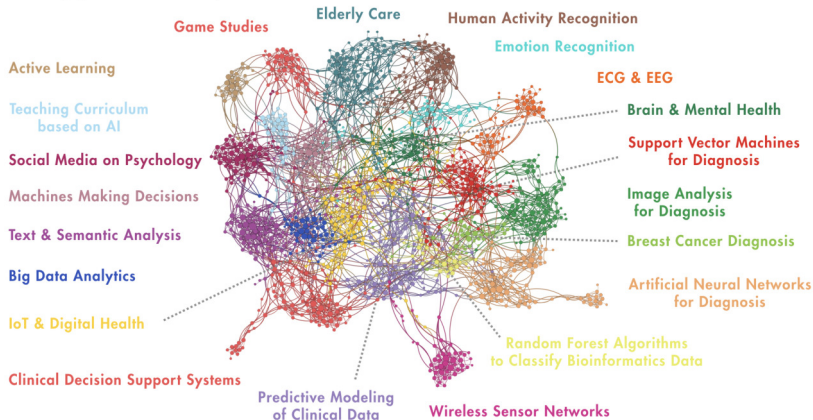


Слова «Порошенко», «Россия», «Украина» встречаются в тексте-1 и тексте-2 одинаково часто, однако:

- «Порошенко» — субъект в тексте-1 и объект в тексте-2;
- «Россия» — агенс в тексте-1 и локация в тексте-2;
- негативная тональность: «Россия», «Кремль» в тексте-1, «Киев», «Украина» в тексте-2.

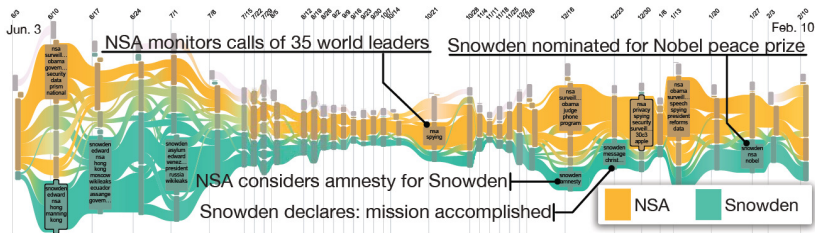
Пример тематической карты «ИИ в биомедицине»

Academic papers on AI in Healthcare published in 2016



C.Folgar, J.McCuan. The 3 most-cited studies in healthcare and AI. Quid, 2017.

Динамика тем: эволюция предметной области



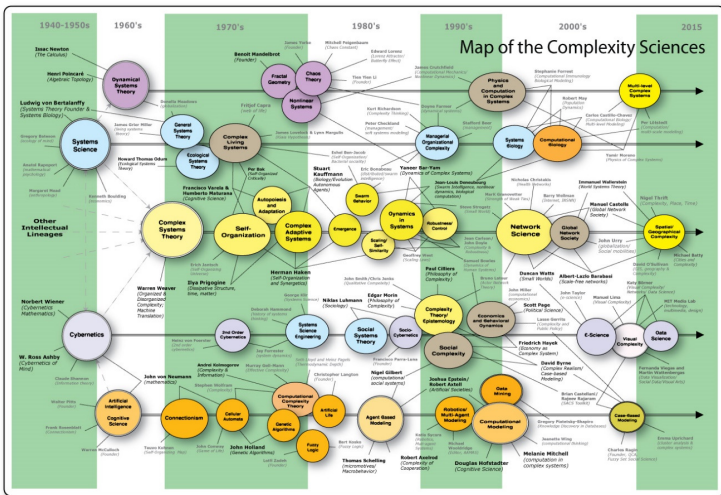
Эволюция выбранных тем иерархии. Данные Prism (2013/06/03–2014/02/09)

Стратегия визуализации в системах TextFlow и RoseRiver:

- эксперт задаёт сечение иерархии (дерева) тем,
- интерактивно выбирает подмножество тем и событий,
- получает сгенерированный отчёт с инфографикой.

Weiwei Cui, Shixia Liu, Zhuofeng Wu, Hao Wei. How hierarchical topics evolve in large text corpora. 2014.

Пример карты предметной области (построено вручную)



<http://www.theoryculturesociety.org/brian-castellani-on-the-complexity-sciences>

Источники вдохновения: <http://textvis.lnu.se>

Интерактивный обзор 440 средств визуализации текстов



Shixia Liu, Weiwei Cui, Yingcai Wu, Mengchen Liu. A survey on information visualization: recent advances and challenges. 2014.

Айсина Р. М. Обзор средств визуализации тематических моделей коллекций текстовых документов // JMLDA, 2015.

Концепция «новостного коллаидера»

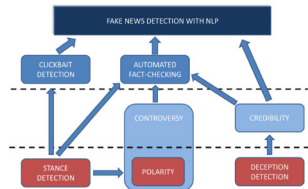
Цель создания
адронного коллаидера —
сталкивая потоки частиц,
узнать больше о строении
материи



Цель создания
новостного коллаидера —
сталкивая потоки новостей,
защитить общество от угроз
эпохи постправды
и информационных войн

Область исследований «Fake News Detection»

- 1 **Deception Detection**
выявление обмана в тексте
- 2 **Automated Fact-Checking**
автоматическая проверка фактов
- 3 **Stance Detection**
выявление позиции за или против
- 4 **Controversy Detection**
выявление и кластеризация разногласий
- 5 **Polarization Detection**
выявление полярных позиций
- 6 **Clickbait Detection**
противоречия заголовка и текста
- 7 **Credibility Scores**
оценка достоверности источников



*E.Saquete, D.Tomas, P.Moreda,
P.Martinez-Barco, M.Palomar.*

**Fighting post-truth using
natural language processing:
a review and open challenges.**

Expert Systems With
Applications, Elsevier, 2020.

Задачи Propaganda/Manipulation/Persuasion Detection

Базовая разметка: «фрагмент, метка класса»



Gallia est omnis divisa in partes tres, quarum unam incolunt Belgae, aliam Aquitani, tertiam qui ipsorum lingua Celtae, nostra Galli appellantur. Hi omnes lingua, institutis, legibus inter se differunt. Gallos ab Aquitania Garumna flumen, a Belgis Matrona et Sequana dividit. Horum **omnium fortissimi** sunt Belgae, propterea quod a cultu atque humanitate provinciae longissime absunt, minimeque ad eos mercatores saepe commeant atque ea quae ad effeminandos **animos pertinent important**, proximique sunt Germanis, qui trans Rhenum incolunt, quibuscum continenter bellum gerunt. Qua de causa **Helvetii quoque reliquos Gallos virtute praecedunt, quod fere cotidianis proelis cum Germanis contendunt**, cum aut suis finibus eos prohibent aut ipsi in eorum finibus bellum gerunt. Eorum una pars, quam Gallos obtinere dictum est, initium capit a flumine Rhodano, continetur Garumna flumine, Oceano, finibus Belgarum, attingit etiam ab Sequanis et Helvetiis flumen Rhenum, vergit ad septentriones. Belgae ab extremis Galliae finibus oriuntur, pertinent

Manipulative Wording: Loaded Language

Attack on Reputation: Smears

Manipulative Wording: Exaggeration

Justification: Appeal to Values



Commissio
PopulusQue
Europaea

Упрощённая разметка: «предложение, метка класса»

Продвинутая разметка: «фрагмент, мишень, метка класса»













SemEval-2023 task 3. Detecting the genre, the framing, and the persuasion techniques in online news in a multi-lingual setup.

<https://propaganda.math.unipd.it/semEval2023task3>

G.Martino, P.Nakov et al. A survey on computational propaganda detection. 2020.

Типология угроз и задачи их автоматической детекции

воздействия → **фейки** → **пропаганда** → **инф.война**

1.  детекция приёмов манипулирования
2.  детекция замалчивания
3.  детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)
4.  детекция кликбэйта (clickbait detection)
5.  автоматическая проверка фактов (auto fact-checking)
6.  детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)
7.  выявление конструкторов картины мира: идеологем, мифологем
8.  оценивание возможных психо-эмоциональных реакций
9.  выявление целевых аудиторий воздействия
10.  оценивание и предсказание скорости распространения (virality prediction)
11.  оценивание достоверности источников (credibility scores)
12.  детекция прямой агрессии (угрозы, призывы, провокации, вербовка, экстремизм)

E.Saquete, D.Tomas, P.Moreda, P.Martinez-Barco, M.Palomar. Fighting post-truth using natural language processing: A review and open challenges // Expert Systems With Applications, Elsevier, 2020.

Типы задач ML/NLU для мониторинга медиа-пространства

- 1. Классификация текста (сообщения/предложения) целиком**
 - deception detection, fact-checking, text credibility
- 2. Классификация пары текстов**
 - stance, controversy, polarization, clickbait detection
 - выявление противоречий, разногласий, замалчивания
- 3. Разметка текста (выделение и классификация фрагментов)**
 - поиск лингвистических маркеров (linguistic-based cues) в тексте
 - детекция приёмов манипулирования
 - выявление конструкторов картины мира: мифологем, идеологем
 - выявление психо-эмоциональных реакций и целевых аудиторий
- 4. Кластеризация или тематическое моделирование**
 - кластеризация мнений по заданной теме (controversy detection)
 - выявление поляризованных мнений (polarization detection)
 - выявление мнений как сочетаний слов, семантических ролей и тональностей
 - выявление «картин мира» – устойчивых сочетаний суждений и идеологем

ПРО//ЧТЕНИЕ — технологический конкурс Up Great

Задача: поиск смысловых ошибок в сочинениях ЕГЭ по русскому, литературе, истории, обществознанию, английскому

Период: декабрь 2019 — декабрь 2022

Призовой фонд:

— 100М руб. русский язык

— 100М руб. английский язык

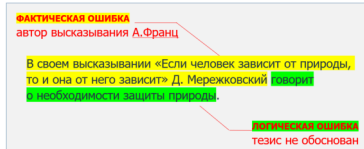
Типов ошибок: 152

(р:70 л:16 о:23 и:20 а:23)

Подтипов ошибок: 236

(р:112 л:19 о:29 и:26 а:50)

Алгоритм должен выделять ошибки и давать их объяснения.



Официальный сайт конкурса: <http://ai.upgreat.one>

Сравнение двух разметок (алгоритма и эксперта)

Алгоритмическая разметка

Нередко люди совершают плохие поступки, забывая о том, что, даже скрыв свой поступок от других, человек не скроется от своей совести. Что же такое безразличный поступок? Безразличный поступок - это поступок, не соответствующий моральным нормам.

Можно ли оправдать безразличный поступок? Именно эту проблему В. Ф. Тендряков поднимает в своем тексте. Докажем сказанное примерами из представленного отрывка.

В тексте В. Ф. Тендряков говорит о том, что человек во благо себе может легко совершить низкий поступок, не испытывая при этом чувства стыда. Человек сможет оправдать свой поступок перед самим собой, объяснив причину. В пример автор приводит поведение героя, который часто в жизни совершал безразличные поступки. Он врал, дрался и крал. Мы видим, что до войны герой привык совершать плохие поступки. Он всегда оправдывался, потому что не хотел нести ответственность за свои действия, а значит не испытывал мучения совести. Мы знаем, что муки совести — это первое и самое сильное наказание, которое получает человек, совершивший плохой поступок. Но наш герой не получал никакого наказания и поэтому продолжал совершать безразличные поступки. Проанализировав поведение главного героя, я убедилась в том, что человек обязан нести ответственность за свои поступки всегда, и поэтому я утверждаю, что нельзя оправдывать даже мелкие безразличные поступки.

человек поступок
человек человек поступок
человек человек поступок
человек
человек поступок поступок поступок
человек поступок поступок поступок
человек человек поступок поступок
человек человек поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок

Экспертная разметка 2

Нередко люди совершают плохие поступки, забывая о том, что, даже скрыв свой поступок от других, человек не скроется от своей совести. Что же такое безразличный поступок? Безразличный поступок - это поступок, не соответствующий моральным нормам.

Можно ли оправдать безразличный поступок? Именно эту проблему В. Ф. Тендряков поднимает в своем тексте. Докажем сказанное примерами из представленного отрывка.

В тексте В. Ф. Тендряков говорит о том, что человек во благо себе может легко совершить низкий поступок, не испытывая при этом чувства стыда. Человек сможет оправдать свой поступок перед самим собой, объяснив причину. В пример автор приводит поведение героя, который часто в жизни совершал безразличные поступки. Он врал, дрался и крал. Мы видим, что до войны герой привык совершать плохие поступки. Он всегда оправдывался, потому что не хотел нести ответственность за свои действия, а значит не испытывал мучения совести. Мы знаем, что муки совести — это первое и самое сильное наказание, которое получает человек, совершивший плохой поступок. Но наш герой не получал никакого наказания и поэтому продолжал совершать безразличные поступки. Проанализировав поведение главного героя, я убедилась в том, что человек обязан нести ответственность за свои поступки всегда, и поэтому я утверждаю, что нельзя оправдывать даже мелкие безразличные поступки.

человек поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек
человек поступок поступок поступок
человек человек поступок поступок
человек человек поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок
человек поступок поступок поступок

- 1 насколько точно предсказана оценка за сочинение
- 2 насколько точно предсказаны фрагменты ошибок и блоков
- 3 насколько точно совпадают границы фрагментов
- 4 совпадают ли типы и подтипы ошибок
- 5 насколько содержательны сгенерированные пояснения

Разметка как способ формализации гуманитарных знаний

Цель — автоматизация обработки текстовых источников (контент-анализа и др.) в социогуманитарных исследованиях.

Гипотеза: достаточно четырёх базовых операций разметки:

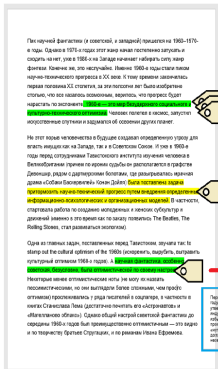
- 1 выделить фрагмент
- 2 классифицировать (тегировать) фрагмент по рубрикатору
- 3 связать несколько фрагментов
- 4 дать комментарий (затекст) к фрагменту или связи

Задачи универсализации обучаемой модели разметки:

- 1 унификация правил разметки и инструментария разметки
- 2 унификация нейросетевой архитектуры модели разметки
- 3 унификация методики оценивания моделей разметки

Унификация правил разметки и инструментария разметки

Обобщение классических задач компьютерной лингвистики (NER, SentAn, SemRL, SyntPars), задач выявления манипуляций, поляризации, смысловых ошибок в академических эссе и др.



Разметка состоит из элементов

Элемент разметки может содержать любое число фрагментов, затекстов и тегов

Теги (классы) выбираются из словаря тегов

Фрагмент задаётся началом и концом, может иметь один или несколько тегов:

BEGIN была поставлена задача

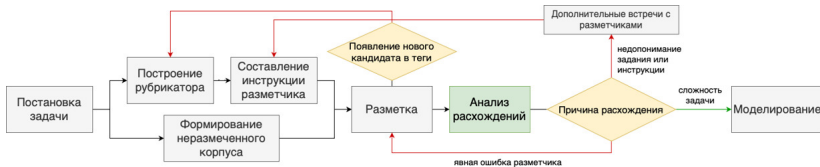
притормозить научно-технический прогресс путем внедрения определенных информационно-психологических и организационных моделей.

END TAG

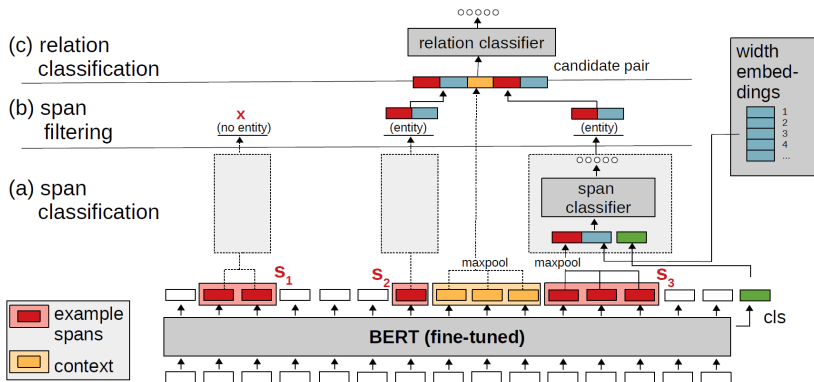
Затекст может выбираться из словаря фраз или свободно генерироваться по контексту, может иметь один или несколько тегов

Технический регламент конкурса ПРО//ЧТЕНИЕ (<http://ai.upgreat.one>)

Процесс разметки и инструмент разметки



Унификация нейросетевых архитектур моделей разметки

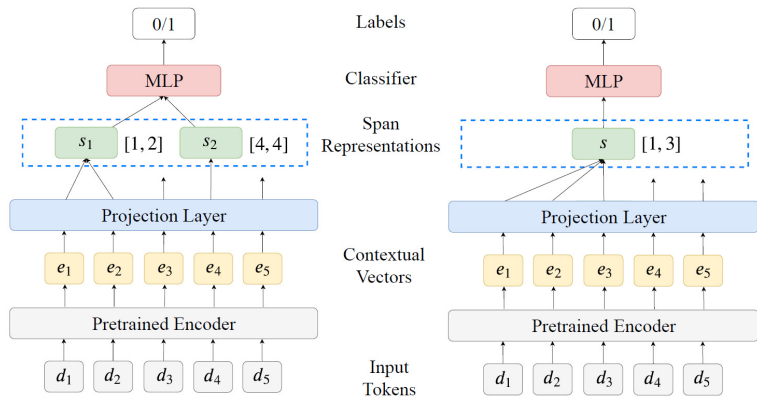


M.Eberts, A.Ulges. Span-based joint entity and relation extraction with transformer pre-training. 2020.

L.Anisiutin, T.Batura, N.Shvarts. Information extraction from news texts using a joint deep learning model. 2021.

Wayne Xin Zhao et al. A Survey of Large Language Models. ArXiv, 29 Jun 2023.

Сравнение методов формирования эмбедингов фрагментов

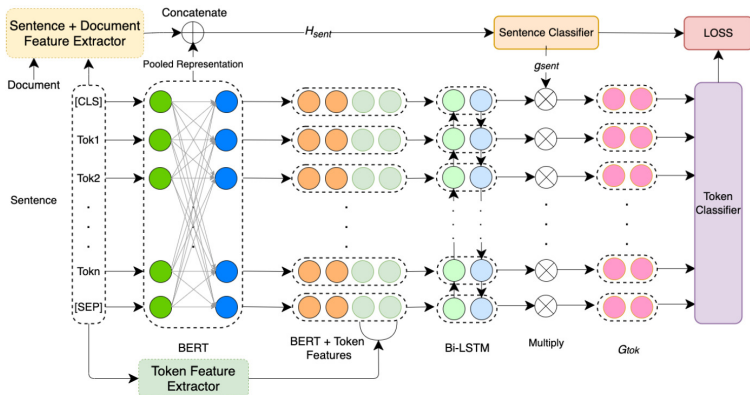


Полнота (recall) до 90% на задачах NER, SRL, Mention Detection

Xiaoya Li et al. A Unified MRC Framework for Named Entity Recognition. 2022.
S. Toshniwal et al. A Cross-Task Analysis of Text Span Representations. 2020.

Извлечение признаков предложений и документов

Задача детекции фрагментов с приёмами пропаганды



Sopan Khosla et al. LTIatCMU at SemEval-2020 Task 11: Incorporating Multi-Level Features for Multi-Granular Propaganda Span Identification. 2020.

Унификация методики оценивания моделей разметки

- В основе методики — сравнение пар разметок текста: «алгоритм – эксперт», «эксперт-1 – эксперт-2», путём оптимального сопоставления их элементов
- Вводятся меры согласованности пары разметок $\text{Con}_k(A, B)$
- Вводится их средневзвешенная согласованность $\text{Con}(A, B)$
- СТАР (Средняя Точность Алгоритмической Разметки) — средняя по размеченной выборке согласованность $\text{Con}(A, E)$ разметки модели A и разметки эксперта E
- СТЭР (Средняя Точность Экспертной Разметки) — средняя по размеченной выборке согласованность $\text{Con}(E_1, E_2)$ разметок двух экспертов, E_1 и E_2
- ОТАР = СТАР / СТЭР (Относительная Точность Алгоритмической Разметки) — если выше 100%, то это означает, что алгоритм работает не хуже экспертов

Выводы

- Нынешний бум искусственного интеллекта обязан развитию методов обучаемой (по большим данным) векторизации сложно структурированных объектов.
- В анализе текстов это большие языковые модели, размер которых сопоставим с размером обучающих данных.
- Эти модели позволяют сегодня решать те задачи, которые ещё 5 лет назад считались непреодолимо трудными.
- В том числе задачи понимания текста для автоматизации и масштабирования социогуманитарных исследований.
- Причём «гибридный интеллект» не заменяет специалиста, а уменьшает объём рутинной работы и ускоряет её.

Воронцов Константин Вячеславович • voron@mlsa-iai.ru