Математические методы понимания естественного языка для мониторинга медиа-пространства

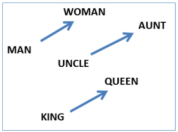
Воронцов Константин Вячеславович
(ФИЦ ИУ РАН, ВМК МГУ, МФТИ)

Заседание Общего собрания ОМН РАН
13 декабря 2021

Содержание

Эволюция подходов машинного обучения к задачам анализа текстов

1. Декомпозиция задач по уровням «пирамиды NLP»
   - морфологический анализ, лемматизация, опечатки
   - синтаксический анализ, выделение терминов, NER
   - семантический анализ, выделение фактов, тем
2. Модели векторных представлений слов (эмбедингов) на основе матричных разложений
   - модели дистрибутивной семантики: word2vec [Mikolov, 2013], FastText [Bojanowski, 2016]
   - тематические модели LDA [Blei, 2003], ARTM [2014]
3. Нейросетевые модели локальных контекстов
   - рекуррентные нейронные сети
   - модели внимания и трансформеры: BERT [2018], GPT-3 [2020] и др.



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

Модели дистрибутивной семантики

Дистрибутивная гипотеза: «слова, появляющиеся в схожих контекстах, имеют схожий смысл» [Харрис, 1954]

Дано: n_{uw} — частота пары слов u, w в одном предложении или окне $\pm h$ слов
Найти: векторные представления (эмбединги) слов x_w и контекстов y_u
Модель: вероятность слова w при условии, что рядом находится слово u:

$$p(w|u) = \underset{w \in W}{\text{SoftMax}} \langle x_w, y_u \rangle = \frac{\exp\langle x_w, y_u \rangle}{\sum_v \exp\langle x_v, y_u \rangle}$$

Критерий: максимизация log-правдоподобия:

$$\sum_{w,u \in W} n_{wu} \ln p(w|u) \;\to\; \max_{\{x_w, y_u\}}$$

Z.Harris. Distributional structure. 1954.
P.Turney, P.Pantel. From frequency to meaning: vector space models of semantics. 2010.
T.Mikolov, K.Chen, G.Corrado, J.Dean. Efficient estimation of word representations in vector space, 2013.

Вероятностное тематическое моделирование

Дано: $p(w|d) = \frac{n_{dw}}{n_d}$ — частотное распределение термов w в документах d

Найти: $p(t|d) = \theta_{td}$ — матрица Θ распределений тем $t \in T$ в документах d

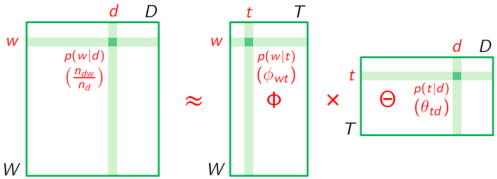
$p(w|t) = \phi_{wt}$ — матрица Φ распределений термов w в темах $t \in T$

Критерий: максимум log-правдоподобия тематической модели

$$\sum_{d \in D} \sum_{w \in d} n_{dw} \ln \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}$$

при ограничениях $\phi_{wt} \geqslant 0$, $\theta_{td} \geqslant 0$, $\sum_w \phi_{wt} = 1$, $\sum_t \theta_{td} = 1$.

Это задача стохастического матричного разложения:



ARTM: тематическая модель с аддитивной регуляризацией и модальностями

Максимизация суммы log-правдоподобий модальностей W_m и регуляризаторов R_i:

$$\sum_{m\in M}\tau_m\sum_{d\in D}\sum_{w\in W_m}n_{dw}\ln\sum_{t\in T}\phi_{wt}\theta_{td}+\sum_i\tau_i R_i(\Phi,\Theta)\ \to\ \max_{\Phi,\Theta}$$

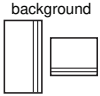
EM-алгоритм: метод простой итерации для решения системы уравнений

$$\text{E-шаг:}\quad p_{tdw}\equiv p(t|d,w)=\operatorname*{norm}_{t\in T}\left(\phi_{wt}\theta_{td}\right)$$

$$\text{M-шаг:}\quad \begin{cases}\phi_{wt}=\operatorname*{norm}_{w\in W_m}\left(\sum_{d\in D}\tau_m n_{dw}p_{tdw}+\sum_i\tau_i\phi_{wt}\frac{\partial R}{\partial\phi_{wt}}\right)\\ \theta_{td}=\operatorname*{norm}_{t\in T}\left(\sum_{w\in d}\tau_m n_{dw}p_{tdw}+\sum_i\tau_i\theta_{td}\frac{\partial R}{\partial\theta_{td}}\right)\end{cases}$$

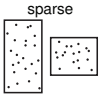
Воронцов К. В. Аддитивная регуляризация тематических моделей коллекций текстовых документов. Доклады РАН. 2014.

Регуляризаторы для улучшения интерпретируемости тем



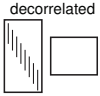
Сглаживание фоновых тем $B \subset T$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in B} \sum_{w} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d} \sum_{t \in B} \alpha_t \ln \theta_{td}$$



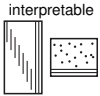
Разреживание предметных тем $S = T \setminus B$:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in S} \sum_{w} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d} \sum_{t \in S} \alpha_t \ln \theta_{td}$$



Декоррелирование для повышения различности тем:

$$R(\Phi) = -\frac{\tau}{2} \sum_{t,s} \sum_{w} \phi_{wt} \phi_{ws}$$



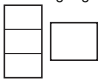
Сглаживание + разреживание + декоррелирование для улучшения интерпретируемости тем

Регуляризаторы для мультимодальных тематических моделей

supervised

Модальность меток классов (категорий) для задач классификации (категоризации) текстов

multilanguage



Модальности языков и регуляризатор вероятностного словаря переводов $\pi_{uwt} = p(u|w,t)$ с языка k на ℓ в мультиязыковой модели:

$$R(\Phi,\Pi) = \tau \sum_{u\in W^k} \sum_{t\in T} n_{ut} \ln \sum_{w\in W^\ell} \pi_{uwt}\phi_{wt}$$

graph

Модальность вершин графа v с подмножествами документов D_v:

$$R(\Phi) = -\frac{\tau}{2} \sum_{(u,v)\in E} S_{uv} \sum_{t\in T} n_t^2 \left(\frac{\phi_{vt}}{|D_v|} - \frac{\phi_{ut}}{|D_u|}\right)^2.$$

geospatial

Модальность геолокаций g с оценками близости $S_{gg'}$:

$$R(\Phi) = -\frac{\tau}{2} \sum_{g,g'\in G} S_{gg'} \sum_{t\in T} n_t^2 \left(\frac{\phi_{gt}}{n_g} - \frac{\phi_{g't}}{n_{g'}}\right)^2$$

Регуляризаторы для учёта дополнительной информации

temporal

Темпоральная модель со сглаживанием по модальности времени i:

$$R(\Phi) = -\tau \sum_{i \in I} \sum_{t \in T} \left| \phi_{it} - \phi_{i-1,t} \right|$$

regression

Линейная модель регрессии $\hat{y}_d = \langle v, \theta_d \rangle$ на документах:

$$R(\Theta, v) = -\tau \sum_{d \in D} \left(y_d - \sum_{t \in T} v_t \theta_{td} \right)^2$$

coherence

Модель дистрибутивной семантики, n_{uv} — частота сочетаемости слов:

$$R(\Phi) = \tau \sum_{u \in W} \sum_{v \in W} n_{uv} \ln \sum_{t \in T} n_t \phi_{ut} \phi_{vt}$$

hierarchy

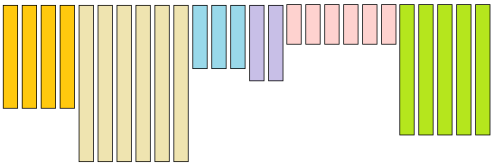
Модель иерархии с родительскими темами t и дочерними подтемами s:

$$R(\Phi, \Psi) = \tau \sum_{t \in T} \sum_{w \in W} n_{wt} \ln \sum_{s \in S} \phi_{ws} \psi_{st}$$

Задача максимизации функции на единичных симплексах

Пусть $\Omega = (\omega_j)_{j \in J}$ — набор нормированных неотрицательных векторов $\omega_j = (\omega_{ij})_{i \in I_j}$ различных размерностей $|I_j|$:

$$\Omega = \left(\quad \right)$$



Задача максимизации функции $f(\Omega)$ на единичных симплексах:

$$\begin{cases} f(\Omega) \to \max\limits_{\Omega}; \\ \sum\limits_{i \in I_j} \omega_{ij} = 1, \quad \omega_{ij} \geqslant 0, \quad i \in I_j,\ j \in J \end{cases}$$

Лемма о максимизации функции на единичных симплексах

Операция нормировки вектора: $p_i = \underset{i \in I}{\operatorname{norm}}(x_i) = \dfrac{\max\{x_i, 0\}}{\sum\limits_{k \in I} \max\{x_k, 0\}}$

Лемма. Пусть $f(\Omega)$ непрерывно дифференцируема по Ω. Тогда векторы ω_j локального экстремума задачи $f(\Omega) \to \max$ удовлетворяют системе уравнений

$$\omega_{ij} = \underset{i \in I_j}{\operatorname{norm}}\left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}}\right), \quad \text{если } \exists i\colon \omega_{ij} \frac{\partial f}{\partial \omega_{ij}} > 0$$

$$\omega_{ij} = \underset{i \in I_j}{\operatorname{norm}}\left(-\omega_{ij} \frac{\partial f}{\partial \omega_{ij}}\right), \quad \text{иначе, если } \exists i\colon \omega_{ij} \frac{\partial f}{\partial \omega_{ij}} < 0$$

$$\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} = 0, \quad \text{иначе}$$

Замечания к Лемме о максимизации на единичных симплексах

- Лемма применима для построения широкого класса моделей, параметризованными дискретными вероятностными распределениями
- Численное решение системы — методом простых итераций
- Существование стационарной точки Ω гарантировано
- Первый из трёх случаев является основным:

$$\omega_{ij} := \underset{i \in I_j}{\operatorname{norm}} \left(\omega_{ij} \frac{\partial f}{\partial \omega_{ij}} \right)$$

- В остальных случаях нормирующий знаменатель нулевой; такие векторы можно считать вырожденными и отбрасывать, сокращая размерность модели
- В отличие от градиентной оптимизации, подбор шага η не требуется:

$$\omega_{ij} := \omega_{ij} + \eta \frac{\partial f}{\partial \omega_{ij}}$$

Теорема о сходимости итерационного процесса

$$\omega_{ij}^{t+1} = \underset{i \in I_j}{\text{norm}}\left(\omega_{ij}^{t} \frac{\partial f(\Omega^t)}{\partial \omega_{ij}^{t}}\right), \quad t = 0, 1, 2, \ldots$$

Теорема. Пусть $f(\Omega)$ — ограниченная сверху, непрерывно дифференцируемая функция, и все Ω^t, начиная с некоторой итерации t^0 обладают свойствами:

- $\forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t = 0 \to \omega_{ij}^{t+1} = 0$ (сохранение нулей)
- $\exists \varepsilon > 0 \;\; \forall j \in J \;\; \forall i \in I_j \;\; \omega_{ij}^t \notin (0, \varepsilon)$ (отделимость от нуля)
- $\exists \delta > 0 \;\; \forall j \in J \;\; \exists i \in I_j \;\; \omega_{ij}^t \frac{\partial f(\Omega^t)}{\partial \omega_{ij}} \geqslant \delta$ (невырожденность)

Тогда $f(\Omega^{t+1}) > f(\Omega^t)$ и $\left|\omega_{ij}^{t+1} - \omega_{ij}^t\right| \to 0$ при $t \to \infty$.

Ирхин И. А., Воронцов К. В. Сходимость алгоритма аддитивной регуляризации тематических моделей // Труды Института математики и механики УрО РАН. 2020.

BigARTM: библиотека тематического моделирования

Ключевые возможности:

- Онлайновый параллельный мультимодальный регуляризованный EM-алгоритм
- Пакетная обработка больших данных, коллекция не хранится в памяти
- Встроенная библиотека регуляризаторов и мер качества

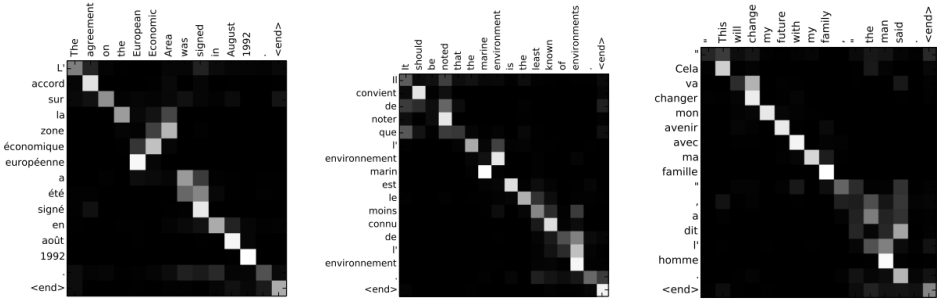
Сообщество:

- 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

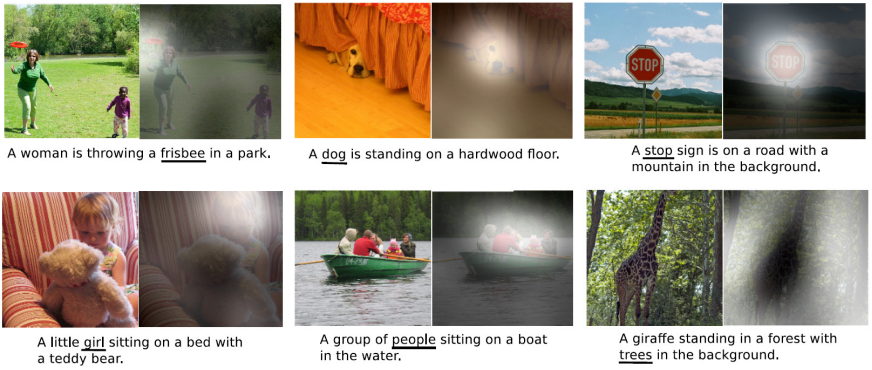
Модели внимания в машинном переводе



Интерпретируемость моделей внимания: матрица семантического сходства $A = (\alpha_{ti})$ показывает, на какие слова x_i из входной последовательности модель обращает внимание, когда генерирует слово перевода y_t

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

Модели внимания на изображениях для генерации описаний



Засветка показывает, на какие области изображения 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

Применения моделей внимания

Преобразование одной последовательности в другую, seq2seq:

- Машинный перевод (machine translation)
- Ответы на вопросы (question answering)
- Ведение диалога (conversational agents)
- Суммаризация текста (text summarization)
- Описание изображений, аудио, видео (multimedia description)
- Распознавание и синтез речи (speech recognition and synthesis)

Обработка последовательности:

- Классификация текстовых документов
- Выделение и классификация фрагментов текста
- Анализ тональности документа / предложений / аспектов

Модель внимания Запрос–Ключ–Значение (Query–Key–Value)

q — *вектор-запрос*, для которого хотим вычислить вектор контекста
$K = (k_1, \ldots, k_n)$ — *векторы-ключи*, которые мы сравниваемые с запросом
$V = (v_i, \ldots, v_n)$ — соответствующие им *векторы-значения*, образующие контекст
$a(k_i, q)$ — *оценка релевантности* (сходства) ключа k_i запросу q

c — искомый вектор контекста, релевантный запросу

Модель внимания — трёх-слойная нейронная сеть, вычисляющая выпуклую комбинацию векторов-значений v_i, наиболее релевантных запросу q по ключам k_i:

$$c = \text{Attn}(q, K, V) = \sum_i v_i \operatorname*{SoftMax}_i a(k_i, q)$$

$c_t = \text{Attn}(W_q y_{t-1}, W_k X, W_v X)$ — в машинном переводе, где $X = (x_1, \ldots, x_n)$ — векторы слов входного предложения, y_{t-1} — предшествующий выходной вектор

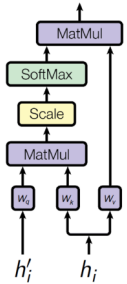
Внутреннее внимание или «самовнимание» (self-attention):
$c_i = \text{Attn}(W_q x_i, W_k X, W_v X)$ — частный случай, когда $x_i \in X$

Разновидности функций сходства векторов

$a(h, h') = h^\mathsf{T} h'$ — скалярное произведение
$a(h, h') = h^\mathsf{T} W h'$ — с матрицей обучаемых параметров W
$a(h, h') = w^\mathsf{T} \operatorname{th}(Uh + Vh')$ — аддитивное внимание с w, U, V

Линейные преобразования векторов query, key, value:

$$a(h_i, h'_{t-1}) = (W_k h_i)^\mathsf{T} (W_q h'_{t-1})/\sqrt{d}$$
$$\alpha_{ti} = \operatorname*{SoftMax}_i a(h_i, h'_{t-1})$$
$$c_t = \textstyle\sum_i \alpha_{ti} W_v h_i$$



$W_{q\,d\times\dim(h')}$, $W_{k\,d\times\dim(h)}$, $W_{v\,d\times\dim(h)}$ — матрицы коэффициентов обучаемых линейных преобразований в пространство размерности d

Возможно упрощение модели: $W_k \equiv W_v$

Dichao Hu. An introductory survey on attention mechanisms in NLP problems. 2018.

Многомерное внимание (multi-head attention)

Идея: J разных моделей внимания совместно обучаются выделять различные аспекты входной информации (например, части речи, синтаксис, фразеологизмы):

$$c^j = \text{Attn}(W_q^j q, W_k^j H, W_v^j H), \quad j = 1, \ldots, J$$

Варианты агрегирования выходного вектора:

$c = \frac{1}{J}\sum_{j=1}^{J} c^j$ — усреднение

$c = \left[c^1 \cdots c^J\right]$ — конкатенация

$c = \left[c^1 \cdots c^J\right] W$ — чтобы вернуться к нужной размерности

Регуляризация: для повышения различности аспектов внимания строки $J \times n$-матрицы A, $\alpha_{ji} = \text{SoftMax}_i a(W_k^j h_i, W_q^j q)$ декоррелируются и разреживаются:

$$\alpha_s^{\mathrm{T}}\alpha_j \to 0, \quad \alpha_j^{\mathrm{T}}\alpha_j \to 1 \qquad \Longrightarrow \qquad \|AA^{\mathrm{T}} - I\|^2 \to \min_{\{W_k^j, W_q^j\}}$$

Zhouhan Lin, Y.Bengio et al. A structured self-attentive sentence embedding. 2017.

Трасформер для машинного перевода

Трасформер (transformer) — это нейросетевая архитектура на основе моделей внимания и полносвязных слоёв, без рекуррентности

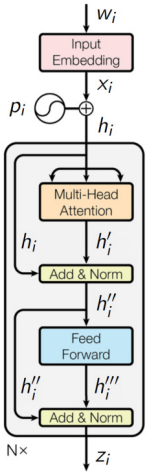
Схема преобразований данных в машинном переводе:

- $S = (w_1, \ldots, w_n)$ — слова предложения на входном языке
  $\downarrow$ обучаемая или пред-обученная векторизация слов
- $X = (x_1, \ldots, x_n)$ — векторные представления слов входного предложения
  $\downarrow$ трансформер-кодировщик
- $Z = (z_1, \ldots, z_n)$ — контекстные векторные представления слов
  $\downarrow$ трансформер-декодировщик, аналогичный кодировщику
- $Y = (y_1, \ldots, y_m)$ — эмбединги 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.

Архитектура трансформера кодировщика

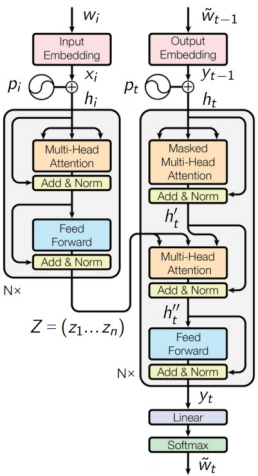
1. К векторам слов x_i добавляются позиционные векторы p_i:
$h_i = x_i + p_i, \quad H = (h_1, \dots, h_n)$
$d = \dim x_i, p_i, h_i = 512$
$\dim H = 512 \times n$
2. Многомерное самовнимание, $j = 1, \dots, J = 8$:
$h_i^j = \mathrm{Attn}(W_q^j h_i, W_k^j H, W_v^j H)$
$\dim h_i^j = 64$
$\dim W_q^j, W_k^j, W_v^j = 64 \times 512$
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 двухслойная сеть:
$h'''_i = W_2\,\mathrm{ReLU}(W_1 h''_i + b_1) + b_2$
$\dim W_1 = 2048 \times 512$
$\dim W_2 = 512 \times 2048$
6. Сквозная связь + нормировка уровня:
$z_i = \mathrm{LN}(h'''_i + h''_i; \mu_2, \sigma_2)$
$\dim z_i, \mu_2, \sigma_2 = 512$



Архитектура трансформера декодировщика

для всех $t = 1, 2, \ldots$ **пока** $\tilde{w}_t \neq \langle \text{EOS} \rangle$:

1. Маскирование «данных из будущего»:
   $h_t = y_{t-1} + p_t; \quad H_t = (h_1, \ldots, h_t)$
2. Многомерное самовнимание:
   $h'_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(W_q^j h_t, W_k^j H_t, W_v^j H_t)$
3. Многомерное внимание на кодировку Z:
   $h''_t = \text{LN} \circ \text{MH}_j \circ \text{Attn}(\tilde{W}_q^j h'_t, \tilde{W}_k^j Z, \tilde{W}_v^j Z)$
4. Двухслойная полносвязная сеть:
   $y_t = \text{LN} \circ \text{FFN}(h''_t)$
5. Линейный предсказывающий слой:
   $p(\tilde{w}|t) = \underset{\tilde{w}}{\text{SoftMax}}(W_y y_t + b_y)$
   $\tilde{w}_t = \arg\max_{\tilde{w}} p(\tilde{w}|t)$ — генерация слова перевода



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.

Критерии обучения трансформеров

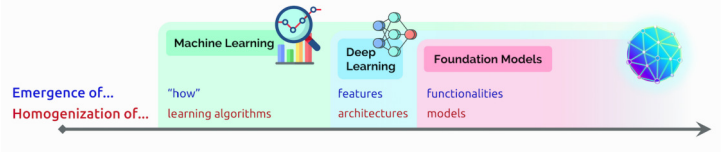
- **Машинный перевод:** максимизация правдоподобия слов перевода $\tilde{w}_t$ по обучающей выборке пар «предложение S, его перевод $\tilde{S}$»:

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

- **BERT MLM (masked language modeling):** максимизация правдоподобия при предсказании маскированных слов по их локальному контексту.
- **BERT NSP (next sentence prediction):** максимизация правдоподобия при предсказании, следуют ли два предложения друг за другом.
- **Fine-tuning:** для дообучения трансформера $Z(S, W)$ на задаче задаётся модель $f(Z(S, W), W_f)$, выборка $\{S\}$ и критерий $\mathscr{L}(S, f) \to \max$
- **Multi-task learning:** для дообучения на наборе задач $\{t\}$ задаются модели $f_t(Z(S, W), W_t)$, выборки $\{S\}_t$ и сумма критериев $\sum_t \lambda_t \sum_S \mathscr{L}_t(S, f_t) \to \max$

Концепция фундаментальных моделей (Foundation Models)

Обучаемая векторизация данных — глобальный тренд в области AI/ML



R.Bommasani et al. (Center for Research on Foundation Models, Stanford University)
On the opportunities and risks of foundation models // CoRR, 20 August 2021.

Выводы. Два вызова в области понимания естественного языка

1. Решение трудных прикладных задач

- выявление потенциально опасного дискурса в СМИ, соцсетях, интернете: фейков, обмана, пропаганды, речевых манипуляций

2. Создание доверенных (интерпретируемых) языковых моделей

- Глубокие нейросети способны аппроксимировать очень сложные функции, но их параметры и векторные представления не интерпретируемы
- Тематические модели интерпретируемы благодаря словарям $p(w|t)$, но имеют простую структуру и слабые аппроксимационные свойства

Возможно ли «объединить лучшее от двух миров»? То есть создать

- глубокие нейросетевые архитектуры,
- параметризованные неотрицательными нормированными векторами $p(t|x)$,
- с мультипликативными шагами в градиентной оптимизации,
- и возможностью интерпретации координат $p(t|x)$ через словари $p(w|t)$

Предпосылки явления и политики постправды (post-truth)

Психологические

- для людей факты менее значимы, чем эмоции и личные убеждения
- люди охотнее распространяют ложь и негатив, чем правду и позитив
- люди подвержены даже таким грубым приёмам пропаганды, как повтор

Политические

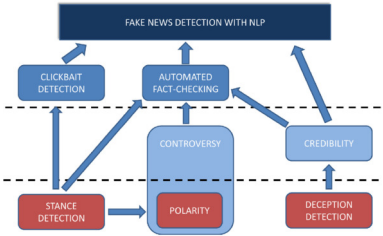
- концепция пропаганды Уолтера Липпмана вытеснила концепцию Джона Дьюи
- постправда — удобный инструмент «мягкой силы» в гибридных войнах

Технологические

- интернет увеличил скорость распространения информации и охват аудитории
- появились технологии генерации фейковых новостей, изображений, видео
- соцсети и рекомендательные системы породили «информационные пузыри»
- СМИ лишились рекламных бюджетов и части самостоятельности

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**
   классификация позиций по многим темам
6. **Clickbait Detection**
   выявление противоречий заголовка и текста
7. **Credibility Scores**
   оценка достоверности источника или новости



E.Saquete et al. Fighting post-truth using natural language processing: a review and open challenges // Expert Systems With Applications, Elsevier, 2020.

Deception Detection (выявление обмана)

История: 50 лет исследований в психологии и криминологии

Задача классификации текста на 2 класса: обман / не обман

Обучающие выборки:

- Контролируемый эксперимент: люди врут / не врут на заданную тему
- Материалы судебных заседаний (датасет DECOUR)
- Отзывы на товары и услуги, проверяемые с помощью краудсорсинга

Признаки: лингвистические маркеры (Linguistic-Based Cues, LBC)

Критерии: accuracy или F-мера 70–92% в зависимости от задачи

На небольших датасетах классический ML лучше и проще DL

Есть проблема переноса моделей на другие датасеты

Типы лингвистических маркеров

Манипулятивные и суггестивные приёмы

- многословие: плеоназмы, лишние слова, тавтологии, расщепления сказуемого
- избыточные повторы слов и фраз
- повышенная когнитивная сложность текста, перегруженные синтаксические конструкции
- повышенная экспрессивность, преобладание негативной тональности
- категоричность, психологическое давление

Уход от личной ответственности

- безличные глаголы, глаголы абстрактной семантики, модальные глаголы, объективация
- неконкретность, уклончивость, безличность, неопределённость высказываний

Подача информации

- оторванность от контекста: пониженная детализация места, времени, событий
- упрощение, пониженное лексическое разнообразие, лексическая недостаточность
- замалчивание фактов, сообщение ложных сведений (fact-checking, см. далее)

Automated Fact-Checking (проверка фактов)

История: ручной fact-checking давно используется в журналистике

Задача классификации текста целиком, по порядковой шкале:
True, Mostly True, Half True, Mostly False, False

Обучающие выборки:

- Платформы для проверки фактов: Politifact, FullFact, FactCheck и др.
- Соревнования: CLEF-2018,19,20,21, FEVER, SemEval (Rumour-Eval)
- Датасеты: NELA-GT-2018,19, FakeNewsNet, Snopes и др.

Вспомогательная задача: стоит ли отправлять текст на проверку?
Три класса: *Non-Factual Sentence, Unimportant, Check-Worthy*
(пример: ClaimBuster, https://idir.uta.edu/claimbuster, 2015)

Credibility Scores (Оценивание надёжности)

История: старая задача в социологии, психологии, маркетинге

Задача: оценить уровень доверия (credibility, trustworthiness)
для источника (СМИ, блогера, пользователя) или отдельной новости

Признаки:

- распространение ненадёжного контента (spam, deception, fake и др.)
- вероятность быть ботом (по диспропорции рассылок и качеству контента)
- стиль контента, геолокация и образовательный уровень читателей

Обучающие выборки:

- много несопоставимых датасетов, отсутствует «золотой стандарт»

Критерии: AUC до 89%; accuracy до 81%; MSE до 0.33

- много критериев, не хватает методологического единства

Stance Detection (выявление позиции)

История: задача текстового следования (textual entailment) — классификация пар текстов «текст $t \rightarrow$ гипотеза h» на три класса: «h следует из t», «h противоречит t», «h не относится к t»

Задача: классификация текста h относительно запроса (claim) t:
agree, disagree, discusses (позиция не высказана), unrelated

Обучающие выборки:

- SNLI: 570K пар предложений: entail, contradict, independent
- Датасеты: Emergent, SemEval-2016 6A(stance), FakeNewsChellenge FNC-1

Критерии: F1-мера до 97% на новостях; Accuracy до 68% на Twitter

Clickbait Detection (обнаружение кликбейта)

История: задача появилась в 2016 году. Обнаружение заголовков или ссылок-приманок, не соответствующих сути контента

Задача: классификация пары «заголовок, текст» на 2 класса «противоречит/нет» Задача аналогична Textual Entailment и Stance Detection

Признаки: гиперболизация, противоречия, web-трафик

Обучающие выборки:

- Датасеты: Webis-Clickbait 2017 (32K заголовков) и др.
- Соревнование: Clickbait challenge 2017

Критерии: F1-мера до 68%; Accuracy до 86%

Controversy / Polarization Detection

Две специальные разновидности задачи Stance Detection

Controversy Detection (выявление полемики, разногласий):

- кластеризация мнений без учителя
- выделение сообществ сторонников каждого мнения в социальной сети
- количественное оценивание объёма и динамики сообществ

Polarization Detection (выявление поляризованности общества):

- выявление разногласий по совокупности запросов или тем

Обучающие выборки:

- Датасеты социальных сетей, обычно Twitter
- Википедия

Критерии: Accuracy 73–83% (на Википедии, методом kNN)

Пример. Поляризация мнений о событии

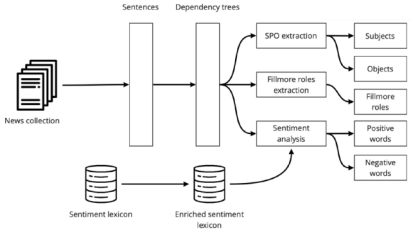
... Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР "национализировали" украинские предприятия ... При этом Кремль защитил конфискацию предприятий в ЛДНР ... Украина потребует расширить 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)***

Subject | Object | Agent | Locative | Negative lexicon | Dependent word

- Слова «Порошенко», «Россия», «Украина» встречаются одинаково часто
- «Порошенко» — субъект в первом тексте и объект во втором
- «Россия» — агенс в первом тексте и локация во втором
- Негативная тональность: «Россия», «Кремль» в 1-м, «Киев», «Украина» во 2-м

Пример. Поляризация мнений о событии



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**

LPR Business

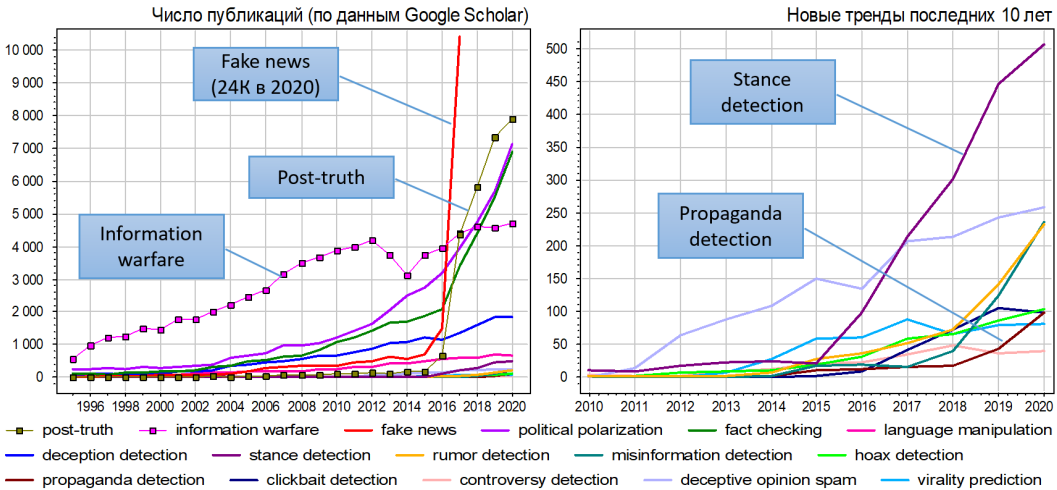
Modalities	*Pr*	*Rec*	*F1*
TF-IDF	0.57	0.97	0.72
SPO	0.56	0.99	0.72
FR	0.67	0.97	0.79
Sent	0.56	0.55	0.55
SPO+FR	0.72	0.99	0.83
SPO+Sent	0.57	0.99	0.72
FR+Sent	0.73	0.97	0.83
All	**0.77**	**0.94**	**0.85**

Paris Trump

- Мнение формализуется как устойчивое сочетание слов, терминов, объектов и субъектов, их семантических ролей по Филлмору и их тональных окрасок
- Все они используются в тематической модели как отдельные модальности

Feldman D. G., *Sadekova T. R.*, *Vorontsov K. V.* Combining facts, semantic roles and sentiment lexicon in a generative model for opinion mining // Computational Linguistics and Intellectual Technologies. Dialogue 2020.

Fake News и близкие тренды исследований (по данным Google Scholar)



Типология потенциально опасного дискурса

воздействия → фейки → пропаганда → инф.война

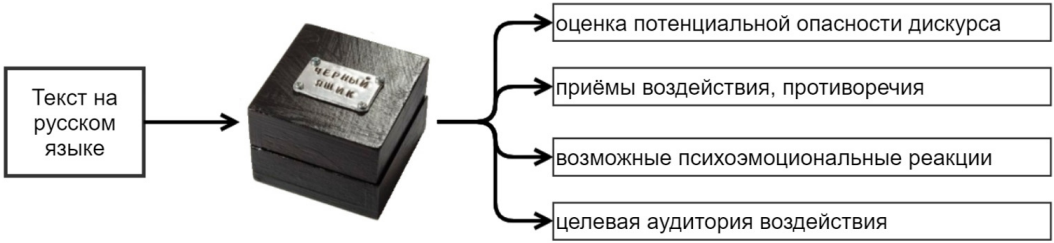
1. детекция приёмов манипулирования
2. детекция замалчивания
3. детекция обмана (deception detection), слухов (rumors d.), мистификаций (hoaxes d.)
4. детекция кликбэйта (clickbait detection)
5. автоматическая проверка фактов (auto fact-checking)
6. детекция позиции (stance d.), противоречий (controversy d.), поляризации (polarization d.)
7. выявление конструктов 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)

Четыре основных типа подзадач ML/NLP 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 конструктов 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, их семантических ролей и тональностей
   - выявление «картин мира» – устойчивых сочетаний суждений и идеологем

Модели выявления потенциально опасного дискурса

Модель принимает на входе текст, на выходе отдаёт оценки текста в целом, разметку текста на фрагменты, оценки и теги отдельных фрагментов.

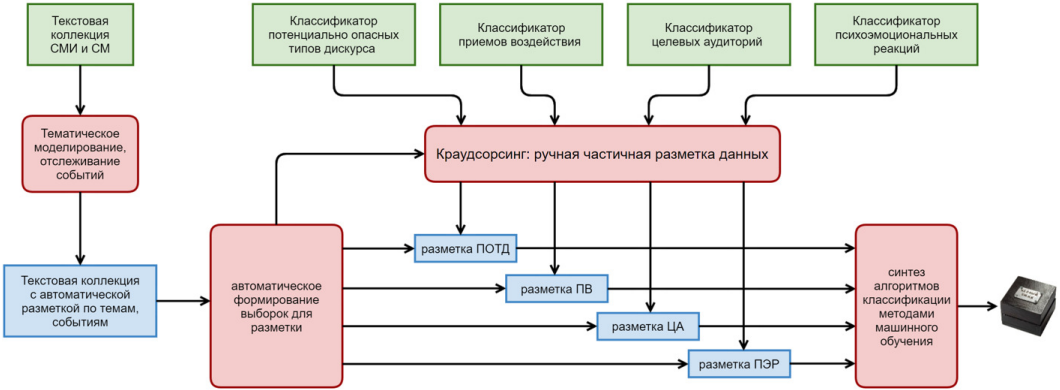


Модель универсальна!

Что именно модель находит в текстах, зависит от обучающей выборки.
Например, можно находить позитивные новости или успешные практики.

Обучение моделей выявления потенциально опасного дискурса

Формирование размеченных выборок — магистральный способ формализации гуманитарных знаний для автоматизации мониторинга медиа-пространства



Резюме

- Противостояние угрозам политики постправды — социально значимая задача, миссия и вызов для научно-технологического сообщества ML/NLP
- Спектр задач детекции фейков расширяется до выявления всех видов информационных угроз (манипуляций, пропаганды, информационной войны)
- Теми же методами могут решаться задачи поиска позитивных новостей или активностей, в частности, с целью их распространения и поддержки
- Одна технология, два целеполагания:
  — детекция негатива с целью противодействия
  — детекция позитива с целью поддержки
- Задачи детекции вполне решаемы современными средствами ML/NLP
- Решение требует междисциплинарного подхода, объединения усилий лингвистов, психологов, политологов, журналистов, AI-инженеров