Федеральное государственное автономное образовательное учреждение
высшего образования
«Московский физико-технический институт
(национальный исследовательский университет)»
Физтех-школа прикладной математики и информатики
Кафедра интеллектуальных систем

Карпов Матвей Леонидович

# ПАРАМЕТРИЗАЦИЯ ТЕМАТИЧЕСКИХ МОДЕЛЕЙ ЛОКАЛЬНОГО КОНТЕКСТА

03.03.01 — Прикладные математика и физика

Выпускная квалификационная работа бакалавра

**Научный руководитель:**
Воронцов Константин Вячеславович,
докт. физ.-мат. наук

Москва — 2026

## Аннотация

В работе рассматривается вероятностная тематическая модель локального контекста, в которой тематическое представление токена строится по соседним словам внутри того же документа. В отличие от моделей мешка слов, такая постановка сохраняет информацию о порядке токенов и локальных совместных встречаемостях. В отличие от локально-контекстной Attentive ARTM (AARTM; статья об этой модели находится на рецензировании, авторы — Воронцов Константин и Илья Дьяков) с заранее заданными позиционными коэффициентами, в данной работе предлагается обучать веса контекста по данным.

Модель задается через словесные тематические эмбеддинги $p(t \mid w)$. Контекстное распределение темы строится как взвешенная комбинация эмбеддингов слов локального окна, причем веса параметризуются относительным смещением и словом. Для модели сформулирована регуляризованная целевая функция, получен EM-алгоритм с обновлением коэффициентов контекста, изучена сходимость итераций обучения весов контекста.

Экспериментальная проверка проводится на корпусах 20 Newsgroups и AG News. Результаты показывают, что обучаемые коэффициенты контекста отражают структуру локального окружения слова. При включении центрального токена в контекст модель быстро приходит к вырожденному самопредсказанию: основная масса веса концентрируется на позиции $k = 0$. Исключение центрального элемента устраняет это самовлияние и приводит к содержательному распределению весов между соседними позициями. В среднем вклад контекстных слов убывает с расстоянием от центрального токена, что согласуется с лингвистической интуицией о большей информативности ближайших слов. При этом значения когерентности остаются ниже, чем у AARTM, поэтому улучшение локально-контекстного вероятностного описания не следует отождествлять с ростом интерпретируемости тем. Выигрышь в перплексии удается получить на длинных текстах.

## Содержание

**1. Введение** **4**

**2. Определения и обозначения** **7**

2.1. Тематическая модель мешка слов . . . . . 8

2.2. Тематическая модель локального контекста . . . . . 10

**3. Предлагаемый метод** **12**

3.1. Параметризация коэффициентов контекста . . . . . 12

3.2. Связь с фиксированной AARTM . . . . . 13

3.3. Обновление коэффициентов контекста . . . . . 14

3.4. Итерационный алгоритм . . . . . 18

**4. Данные** **20**

4.1. Датасеты . . . . . 20

4.1.1. 20 Newsgroup . . . . . 20

4.1.2. Ag news . . . . . 21

4.2. Предобработка . . . . . 21

**5. Вычислительный эксперимент** **23**

5.1. Метрики . . . . . 23

5.2. Переобучение и самовнимание . . . . . 24

5.3. Коэффициенты внимания . . . . . 25

**Заключение** **27**

**Список литературы** **29**

## 1. Введение

Вероятностные тематические модели остаются удобным инструментом разведочного анализа текстовых коллекций. Они представляют документ распределением по скрытым темам, а тему — распределением по словам; такие векторы можно интерпретировать вручную, использовать как признаки документов и применять для навигации по корпусу. Классические PLSA [1, 2] и LDA [3] задают эту идею через вероятностную факторизацию:

$$p(w \mid d) = \sum_{t \in T} p(w \mid t) p(t \mid d).$$

Эта постановка сделала тематическое моделирование простым и масштабируемым, но одновременно закрепила два сильных предположения: документ рассматривается как мешок слов, а все вхождения слов внутри документа объясняются одной и той же документной смесью тем.

Оба предположения часто оказываются слишком грубыми. В коротких текстах мало слов для надежной оценки $p(t \mid d)$, а в длинных документах отдельные фрагменты могут относиться к разным темам. В таких случаях тема конкретного вхождения слова определяется скорее его ближайшим окружением, чем всем документом целиком. Поэтому в тематическом моделировании возникло несколько линий работ, которые ослабляют предположение мешка слов: bigram topic model учитывает соседние пары слов [4], biterm topic model и word network topic model используют локальные совместные встречаемости [5, 6], а Local LDA вводит перекрывающиеся контекстные окна [7]. Эти модели показывают важность локальной информации, но обычно требуют отдельной вероятностной постановки и отдельного вывода алгоритма обучения.

С точки зрения оптимизационного подхода PLSA является задачей приближенной стохастической матричной факторизации, которая нуждается в регуляризации. Аддитивная регуляризация тематических моделей (ARTM), предложенная К. В. Воронцовым [8, 9], возвращает байесовские и специализированные тематические модели в общую оптимизационную схему. В ARTM к логарифму правдоподобия добавляется взвешенная сумма регуляризаторов, отвечающих за разреживание, сглаживание, декорреляцию, использование внешней информации или другие желаемые свойства решения. Это позволяет переносить регуляризаторы между моделями, комбинировать несколько ограничений

и сохранять общий EM-подобный алгоритм вместо вывода новой процедуры инференса для каждой модификации.

Параллельно активно развиваются нейросетевые тематические модели. BERTopic [10] строит темы как кластеры контекстных эмбеддингов и затем описывает их ключевыми словами с помощью class-based TF-IDF. Такой подход хорошо использует предобученные языковые представления и служит сильным эмпирическим ориентиром, особенно для коротких текстов и лексически разнообразных корпусов. Однако темы в нем не являются скрытыми компонентами единого правдоподобия, а величины $p(w \mid t)$ и $p(t \mid d)$ не выступают центральными параметрами модели. Поэтому BERTopic не заменяет вероятностную постановку, в которой можно аналитически выводить обновления и контролировать вклад отдельных компонент.

В данной работе рассматривается компромиссный путь: сохранить вероятностную интерпретацию тем и модульность ARTM, но заменить глобальный документный контекст локальным. Локально-контекстная Attentive ARTM (AARTM; статья об этой модели находится на рецензировании, авторы — Воронцов Константин и Илья Дьяков) [11] задает для каждой позиции $i$ окно соседних токенов $C_i$ и строит контекстное тематическое распределение

$$p(t \mid C_i) = \sum_{c \in C_i} \alpha_{ci} p(t \mid w_c).$$

Вероятность центрального токена затем выражается через это локальное распределение вместо $p(t \mid d)$. Такая модель выходит за пределы мешка слов и сохраняет интерпретируемые распределения $p(t \mid w)$ и $p(w \mid t)$, но в исходной AARTM коэффициенты $\alpha_{ci}$ задаются заранее: фиксируются размер окна, ориентация контекста, затухание по расстоянию.

В работе предлагается Adaptive Attentive ARTM (Adaptive AARTM, AAARTM) — локально-контекстная тематическая модель с обучаемыми коэффициентами контекста. Вместо фиксированного правила агрегации соседних слов модель параметризует веса контекста относительным смещением и токеном, а затем обучает их совместно со словесными тематическими эмбеддингами $p(t \mid w)$. Благодаря этому разные слова могут по-разному использовать левый, правый, центральный и дальний контекст, а само правило агрегации становится частью модели, а не только внешним гиперпараметром.

Основные результаты работы заключаются в следующем:

1. сформулирована локально-контекстная тематическая модель через словесные тематические эмбеддинги и обучаемые коэффициенты контекста;
2. введена параметризация обучаемых коэффициентов контекста $\alpha_{wh}$, позволяющая задавать вес позиции в окне в зависимости от центрального слова и смещения;
3. выведен EM-алгоритм для локально-контексной модели с обучаемыми коэффициентами внимания;
4. исследована сходимость алгоритма;
5. проведены эксперименты на корпусах 20 Newsgroups и AG News, включающие сравнение с фиксированной AARTM, PLSA и BERTopic по когерентности тем, разнообразию тем и времени обучения;
6. экспериментально показано, что обучаемые коэффициенты контекста восстанавливают осмысленную структуру локального окружения: при исключении центрального токена наибольший вес получают ближайшие позиции, а вклад более удаленных слов постепенно уменьшается.

Эксперименты показывают, что обучаемый локальный контекст позволяет модели лучше адаптироваться к структуре данных, чем фиксированная схема весов окна. В частности, анализ коэффициентов $\alpha_{wh}$ демонстрирует, что при включении центрального токена возникает вырожденное самопредсказание, тогда как его исключение приводит к содержательному распределению веса между соседними позициями. При этом значения когерентности остаются ниже, чем у AARTM, что указывает на ограничение предложенного подхода: улучшение локально-контекстного вероятностного описания не гарантирует автоматического роста интерпретируемости тем. На длинных текстах значение перплексити ниже, чем у AARTM. Поэтому предложенная модель рассматривается прежде всего как вероятностный способ учета порядка слов и локального контекста.

## 2. Определения и обозначения

Коллекцию документов обозначим $D$. Пусть после предобработки коллекция представлена последовательностью токенов

$$w_1, \ldots, w_N, \qquad w_i \in W,$$

где $W$ – словарь, $|W| = V$, а $N$ – суммарное число токенов в коллекции. Границы документов задаются индексами

$$0 = b_0 < b_1 < \cdots < b_{|D|} = N,$$

так что токены документа $d$ имеют номера $b_{d-1} + 1, \ldots, b_d$. Множество тем обозначается через $T$. Эмпирическая вероятность слова $w$ в коллекции равна $p(w) = n_w/N$, где

$$n_w = \sum_{i=1}^{N} \mathbf{1}[w_i = w].$$

В дальнейшем основным параметром является матрица словесных тематических эмбеддингов

$$\Phi = (\varphi_{tw})_{w \in W, t \in T}, \qquad \varphi_{tw} = p(t \mid w), \qquad \sum_{t \in T} \varphi_{tw} = 1.$$

Априорная вероятность темы на уровне коллекции определяется как

$$p(t) = \sum_{w \in W} p(w) \varphi_{tw}.$$

При $p(t) > 0$ соответствующее распределение слов в теме восстанавливается по формуле Байеса:

$$\varphi_{wt} = p(w \mid t) = \frac{p(w)\varphi_{tw}}{p(t)}.$$

Такая запись удобна для локально-контекстной модели, поскольку контекстные представления строятся как выпуклые комбинации распределений $p(t \mid w)$.

### 2.1. Тематическая модель мешка слов

**Определение 1** (Тематическая модель мешка слов)**.** Задана коллекция документов $D$. Каждый документ $d \in D$ является последовательностью токенов из словаря $W$, но в модели мешка слов порядок токенов отбрасывается, и документ представляется мультимножеством слов с частотами $n_{dw}$. Требуется оценить распределения слов в темах $\varphi_{wt} = p(w \mid t)$ и тематические смеси документов $\theta_{td} = p(t \mid d)$.

Для классической модели мешка слов документ $d$ имеет тематическую смесь $\theta_{td} = p(t \mid d)$, а вероятность слова записывается как

$$p(w \mid d) = \sum_{t \in T} \varphi_{wt} \theta_{td}.$$

По формуле Байеса та же вероятность принимает вид

$$p(w \mid d) = p(w) \sum_{t \in T} \frac{\varphi_{tw} \theta_{td}}{p(t)}.$$

Параметры $\Phi$ и $\Theta$ оцениваются максимизацией регуляризованного логарифма правдоподобия

$$\max_{\Phi,\Theta} \quad L(\Phi,\Theta) + R(\Phi,\Theta), \qquad L(\Phi,\Theta) = \sum_{d \in D} \sum_{w \in d} n_{dw} \log p(w \mid d),$$

при ограничениях

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geq 0, \qquad \sum_{t \in T} \theta_{td} = 1, \quad \theta_{td} \geq 0.$$

В ARTM регуляризатор $R(\Phi,\Theta)$ обычно представляется взвешенной суммой частных регуляризаторов и используется для стабилизации некорректной задачи матричной факторизации.

**Определение 2** (Оператор нормировки)**.** Для вектора $x = (x_i)_{i \in I}$ положим

$$\operatorname{norm}_{i \in I}(x_i) = \frac{\max\{0,x_i\}}{\sum_{k \in I} \max\{0,x_k\}}, \qquad i \in I,$$

если знаменатель положителен. Если все компоненты $x_i$ неположительны, опе-

ратор считается неопределенным; в алгоритмической реализации такой случай устраняется добавлением малой положительной константы в знаменатели и инициализацией параметров внутри симплекса.

**Лемма 1** (Стационарность на произведении симплексов, [9]). Пусть $f(\Omega)$ непрерывно дифференцируема, а переменные разбиты на блоки $\omega_j = (\omega_{ij})_{i \in I_j}$, каждый из которых лежит в симплексе

$$\Delta_j = \left\{ \omega_j : \sum_{i \in I_j} \omega_{ij} = 1,\ \omega_{ij} \geq 0 \right\}.$$

Если $\omega^*$ является локальным максимумом $f$ и для некоторого блока выполнено

$$\omega^*_{ij} \frac{\partial f}{\partial \omega_{ij}}(\omega^*) \geq 0 \quad i \in I_j$$

и хотя бы одна из этих величин положительна, то этот блок удовлетворяет условию неподвижной точки

$$\omega^*_{ij} = \mathrm{norm}_{i \in I_j} \left( \omega^*_{ij} \frac{\partial f}{\partial \omega_{ij}}(\omega^*) \right).$$

**Теорема 1** (EM-алгоритм для ARTM, [12]). Пусть $R(\Phi,\Theta)$ непрерывно дифференцируема. Тогда локальный экстремум регуляризованной задачи тематической модели мешка слов удовлетворяет системе уравнений с вспомогательными переменными

$$p_{tdw} = p(t \mid d,w) = \mathrm{norm}_{t \in T}(\varphi_{wt}\theta_{td}),$$

$$n_{wt} = \sum_{d \in D} n_{dw} p_{tdw}, \qquad n_{td} = \sum_{w \in d} n_{dw} p_{tdw},$$

$$\varphi_{wt} = \mathrm{norm}_{w \in W} \left( n_{wt} + \varphi_{wt} \frac{\partial R}{\partial \varphi_{wt}} \right), \qquad \theta_{td} = \mathrm{norm}_{t \in T} \left( n_{td} + \theta_{td} \frac{\partial R}{\partial \theta_{td}} \right).$$

Простая итерация этой системы задает EM-подобный алгоритм: сначала вычисляются апостериорные распределения тем $p_{tdw}$, затем обновляются ожидаемые счетчики и параметры модели. При $R = 0$ получается PLSA, а ARTM добавляет к этим уравнениям регуляризационные слагаемые.

Сходимость такого EM-подобного алгоритма к стационарной точке при проверяемых условиях на регуляризатор доказана в работе [13]. В этой же линии

используется замена правых частей обновлений несмещенными оценками частот, что позволяет ускорять обучение без дополнительных затрат памяти.

### 2.2. Тематическая модель локального контекста

Модель локального контекста заменяет документную смесь $\theta_{td} = p(t \mid d)$ позиционным распределением $p(t \mid C_i)$. Тем самым вероятность токена объясняется не всем документом, а локальным окружением его вхождения.

**Определение 3** (Локальный контекст)**.** Для позиции $i$ локальным контекстом называется множество индексов

$$C_i \subseteq \{b_{d-1} + 1, \ldots, b_d\},$$

где $d$ — документ, содержащий позицию $i$. Это условие запрещает контекстным окнам пересекать границы документов. В экспериментах $C_i$ задается окном радиуса $len$.

Каждой позиции $w$ сопоставим неотрицательные коэффициенты $\alpha_{ci}$, $c \in C_i$, такие что

$$\sum_{c \in C_i} \alpha_{ci} = 1.$$

В базовой AARTM эти коэффициенты фиксируются заранее. В предлагаемой AAARTM они являются функцией обучаемых весов контекста и поэтому могут адаптироваться к данным. Контекстное распределение тем определяется как выпуклая комбинация тематических эмбеддингов слов из локального окна:

$$p(t \mid C_i) = \theta_{ti} = \sum_{c \in C_i} \alpha_{ci} \varphi_{tw_c}.$$

Тогда вероятность центрального токена $w_i$ при заданном контексте равна

$$p(w_i \mid C_i) = p(w_i) \sum_{t \in T} \frac{\varphi_{tw_i} \theta_{ti}}{p(t)}.$$

При фиксированных контекстах и коэффициентах $\alpha$ параметры $\Phi$ оценива-

ются максимизацией регуляризованного логарифма правдоподобия

$$L(\Phi) = \sum_{i=1}^{N} \log p(w_i \mid C_i) + R(\Phi),$$

где строки $\Phi$ лежат в единичных симплексах, а $R(\Phi)$ обозначает дифференцируемый регуляризатор.

**Теорема 2** (EM-алгоритм для локально-контекстной модели, [11])**.** Пусть функция $R(\Phi)$ непрерывно дифференцируема. Тогда локальный экстремум задачи тематической модели локального контекста удовлетворяет системе уравнений со вспомогательными переменными $p_{ti} = p(t \mid C_i, w_i)$, $\theta_{ti} = p(t \mid C_i)$, $N_{tw}$, $q_{wi}$ и $n_{tw}$:

$$p_{ti} = \underset{t \in T}{\operatorname{norm}} \left( \frac{\varphi_{tw_i} \theta_{ti}}{p(t)} \right);$$

$$\theta_{ti} = \sum_{c \in C_i} \alpha_{ci} \varphi_{tw_c}, \qquad p(t) = \sum_{w \in W} \varphi_{tw} p(w);$$

$$q_{wi} = \sum_{c \in C_i} \alpha_{ci} [w_c = w], \qquad N_{tw} = \sum_{i=1}^{n} q_{wi} \frac{p_{ti}}{\theta_{ti}};$$

$$n_{tw} = \sum_{i=1}^{n} p_{ti} [w_i = w], \qquad n_t = \sum_{i=1}^{n} p_{ti};$$

$$\varphi_{tw} = \underset{t \in T}{\operatorname{norm}} \left( n_{tw} + \varphi_{tw} \frac{\partial R}{\partial \varphi_{tw}} + \varphi_{tw} N_{tw} - \varphi_{tw} n_t \frac{p(w)}{p(t)} \right).$$

Третий член внутри нормировки является контекстно-индуцированным: он усиливает слова, часто встречающиеся в локальных окружениях темы. Четвертый член связан с тем, что $p(t)$ определяется через $\Phi$ и поэтому также зависит от оптимизируемых параметров.

## 3. Предлагаемый метод

Предлагаемый метод обучает словесные тематические эмбеддинги $\Phi$ и коэффициенты локального внимания в одной итерационной процедуре. По сравнению с моделью мешка слов вероятность токена зависит не от всего документа, а от локального контекста $C_i$. По сравнению с фиксированной локально-контекстной AARTM меняется способ задания коэффициентов $\alpha_{ci}$: вместо заранее выбранного экспоненциального затухания используются обучаемые веса смещений.

Ключевая идея состоит в том, что важность позиции в контексте должна зависеть не только от расстояния до центрального токена, но и от самого слова. Одни слова могут быть описаны левыми соседями, другие правыми. Поэтому веса контекста параметризуются одновременно смещением и словом, но остаются общими для всей коллекции. Такая параметризация существенно меньше, чем задание отдельного вектора $\alpha_i$ для каждой позиции, и при этом достаточно гибка, чтобы извлекать устойчивые локальные закономерности из данных.

### 3.1. Параметризация коэффициентов контекста

Пусть $K$ обозначает радиус локального окна, а

$$k_i \subseteq \{-K, \ldots, K\}$$

— множество смещений, допустимых для позиции $i$: индекс $i + k$ должен принадлежать тому же документу, что и $i$, и удовлетворять выбранному типу окна. Для каждого смещения $k$ и слова $w$ задается вес $\alpha_{wk} \geq 0$. Веса нормируются по смещениям базового окна для каждого слова:

$$\sum_{k=-K}^{K} \alpha_{wk} = 1, \qquad w \in W.$$

Для конкретной позиции $i$ часть смещений может быть недоступна из-за границ документа или выбранного режима окна. Для недопустимых смещений вклад полагается равным нулю. Тематическое распределение контекста принимает вид

$$\theta_{ti} = \sum_{k=-K}^{K} \alpha_{w_i k} \varphi_{t w_{i+k}}.$$

Тогда вероятность центрального токена $w_i$ при заданном контексте равна

$$p(w_i \mid C_i) = \sum_{t \in T} p(w_i \mid t) p(t \mid C_i) = \sum_{t \in T} \varphi_{tw_i} \frac{p(w_i)}{p(t)} \sum_{k=-K}^{K} \alpha_{w_i k} \varphi_{tw_{i+k}}$$

Обозначим

$$A = (\alpha_{wk})_{w \in W, k \subseteq \{-K,\ldots,K\}},$$

Параметры $\Phi$ и $A$ оцениваются максимизацией регуляризованного логарифма правдоподобия

$$\max_{\Phi,A} \quad L(\Phi,A) + R(\Phi,A), \qquad L(\Phi,A) = \sum_{i=1}^{N} \log p(w_i \mid C_i),$$

при ограничениях

$$\sum_{w \in W} \varphi_{wt} = 1, \quad \varphi_{wt} \geq 0, \qquad \sum_{k=-K}^{K} \alpha_{wk} = 1, \quad \alpha_{wk} \geq 0.$$

В отличие от фиксированной AARTM, здесь веса $\alpha_{wk}$ входят в оптимизационную задачу и обновляются по данным. В отличие от нейросетевого self-attention, они остаются неотрицательными, нормированными и интерпретируемыми: для каждого слова можно посмотреть, какие относительные позиции локального окна получили наибольший вес.

### 3.2. Связь с фиксированной AARTM

Если положить $\alpha_{wk}$ одинаковыми для всех слов и задать их заранее как затухающую функцию расстояния $|k|$, предлагаемая модель сводится к фиксированной локально-контекстной схеме. Если дополнительно исключить центральный токен из множества допустимых смещений, получается self-excluding режим; если оставить его, получается self-aware режим. В статье по AARTM показано, что включение центрального слова может приводить к предсказательной утечке: модель получает возможность слишком сильно опираться на сам предсказываемый токен. Поэтому в данной работе вводится дополнительный контроль самовлияния через self_aware режим, а велична переобучения изучается в отдельном эксперименте.

### 3.3. Обновление коэффициентов контекста

**Теорема 3** (EM-алгоритм внимательной локально-контексной модели)**.** Пусть функция $R(\Phi, A)$ непрерывно дифференцируема. Тогда локальный экстремум задачи тематической модели локального контекста с обучаемыми коэффициентами внимания удовлетворяет системе уравнений со вспомогательными переменными $p_{ti} = p(t \mid C_i,w_i)$, $\theta_{ti} = p(t \mid C_i)$, $N_{tw}$, $q_{wi}$ и $n_{tw}$:

$$p_{ti} = \underset{t\in T}{\operatorname{norm}} \left( \frac{\varphi_{tw_i}\theta_{ti}}{p(t)} \right);$$

$$\theta_{ti} = \sum_{k=-K}^{K} \alpha_{w_i k}\varphi_{tw_{i+k}}, \qquad p(t) = \sum_{w\in W} \varphi_{tw}p(w);$$

$$q_{wi} = \sum_{k=-K}^{K} \alpha_{w_i k}[w_{i+k} = w], \qquad N_{tw} = \sum_{i=1}^{n} q_{wi}\frac{p_{ti}}{\theta_{ti}};$$

$$n_{tw} = \sum_{i=1}^{n} p_{ti}[w_i = w], \qquad n_t = \sum_{i=1}^{n} p_{ti};$$

$$\alpha_{wh} = \underset{h\in\{-K,\ldots,K\}}{\operatorname{norm}} \left( \alpha_{wh} \sum_{i=1}^{n}\sum_{t\in T}\sum_{k=-K}^{K} p_{ti}\frac{\varphi_{tw_{i+k}}}{\theta_{ti}}[w_i = w][k = h] + \alpha_{wh}\frac{\partial R}{\partial \alpha_{wh}} \right).$$

$$\varphi_{tw} = \underset{t\in T}{\operatorname{norm}} \left( n_{tw} + \varphi_{tw}\frac{\partial R}{\partial \varphi_{tw}} + \varphi_{tw}N_{tw} - \varphi_{tw}n_t\frac{p(w)}{p(t)} \right).$$

**Доказательство.** Из определения модели

$$p(w_i \mid C_i) = p(w_i)\sum_{t\in T}\frac{\varphi_{tw_i}\theta_{ti}}{p(t)}, \qquad \theta_{ti} = \sum_{k=-K}^{K} \alpha_{w_i k}\varphi_{tw_{i+k}}.$$

Так как $\partial\theta_{ti}/\partial\alpha_{wh} = \varphi_{tw_{i+k}}[w_i = w][k = h]$, имеем

$$\frac{\partial L}{\partial \alpha_{wh}} = \sum_{i=1}^{n}\frac{1}{p(w_i \mid C_i)}p(w_i)\sum_{t\in T}\sum_{k=-K}^{K}\frac{\varphi_{tw_i}}{p(t)}\varphi_{tw_{i+k}}[w_i = w][k = h].$$

С другой стороны,

$$p_{ti} = \frac{\varphi_{tw_i}\theta_{ti}/p(t)}{\sum_{s\in T}\varphi_{sw_i}\theta_{si}/p(s)}$$

и

$$p(w_i \mid C_i) = p(w_i) \sum_{s \in T} \frac{\varphi_{sw_i} \theta_{si}}{p(s)}.$$

Подстановка этих равенств в выражение для производной дает

$$\frac{\partial L}{\partial \alpha_{wh}} = \sum_{i=1}^{n} \sum_{t \in T} \sum_{k=-K}^{K} p_{ti} \frac{\varphi_{tw_{i+k}}}{\theta_{ti}} [w_i = w][k = h].$$

Неподвижное уравнение следует из леммы о стационарности на симплексе, примененной к блоку $\alpha_{wh}$. □

**Теорема 4** (Сходимость обновления $\alpha_{wk}$)**.** Зафиксируем $\Phi$, $p(t)$ и токен $w$. Пусть

$$\ell_w(\alpha_w) = \sum_{w_i = w} \log p(w_i \mid C_i), \qquad \alpha_w \in \Delta_w,$$

и пусть начальная точка имеет полный носитель: $\alpha_{wh}^{(0)} > 0$ для всех $h \in \{-K, \ldots, K\}$. Предположим также, что на всех итерациях $p(w_i \mid C_i) > 0$. Рассмотрим итерацию

$$\alpha_{wh}^{(k+1)} = \underset{h \in \{-K,\ldots,K\}}{\operatorname{norm}} \left( \alpha_{wh}^{(k)} \sum_{i=1}^{n} \sum_{t \in T} \sum_{k=-K}^{K} p_{ti}^{(k)} \frac{\varphi_{tw_{i+k}}}{\theta_{ti}^{(k)}} [w_i = w][k = h] \right),$$

где $p_{ti}^{(k)}$ и $\theta_{ti}^{(k)}$ вычислены при $\alpha_w = \alpha_w^{(k)}$. Тогда:

$$\ell_w(\alpha_w^{(k+1)}) \geq \ell_w(\alpha_w^{(k)}),$$

последовательность $\ell_w(\alpha_w^{(k)})$ сходится, а каждая предельная точка $\alpha_w^*$ последовательности $\{\alpha_w^{(k)}\}$ удовлетворяет условиям Каруша–Куна–Таккера для задачи

$$\max_{\alpha_w \in \Delta_w} \ell_w(\alpha_w).$$

Поскольку $\ell_w$ вогнута по $\alpha_w$, каждая такая точка является глобальным максимумом блока. Если глобальный максимум единственен, то вся последовательность $\alpha_w^{(k)}$ сходится к нему.

**Доказательство.** Обозначим $H = \{-K, \ldots, K\}$, $I_w = \{i : w_i = w\}$ и положим

$a_h = \alpha_{wh}$. При фиксированных $\Phi$ и $p(t)$ слагаемые, зависящие от блока $a = (a_h)_{h \in H}$, имеют вид

$$\ell_w(a) = \sum_{i \in I_w} \log \left( p(w_i) \sum_{t \in T} \frac{\varphi_{tw_i}}{p(t)} \sum_{h \in H} a_h \varphi_{tw_{i+h}} \right).$$

Множитель $p(w_i)$ не влияет на максимизацию, поэтому далее его можно считать частью константы. Введем

$$\theta_{ti}(a) = \sum_{h \in H} a_h \varphi_{tw_{i+h}}, \qquad Z_i(a) = \sum_{t \in T} \frac{\varphi_{tw_i}}{p(t)} \theta_{ti}(a).$$

Тогда

$$p_{ti}^{(r)} = \frac{\varphi_{tw_i} \theta_{ti}(a^{(r)})/p(t)}{\sum_{s \in T} \varphi_{sw_i} \theta_{si}(a^{(r)})/p(s)}.$$

Построим минорирующую функцию. По неравенству Йенсена,

$$\log Z_i(a) = \log \sum_{t \in T} p_{ti}^{(r)} \frac{\varphi_{tw_i} \theta_{ti}(a)/p(t)}{p_{ti}^{(r)}} \geq \sum_{t \in T} p_{ti}^{(r)} \log \frac{\varphi_{tw_i} \theta_{ti}(a)/p(t)}{p_{ti}^{(r)}}.$$

Далее, поскольку $a_h^{(r)} > 0$, снова применяем Йенсена:

$$\theta_{ti}(a) = \sum_{h \in H} a_h \varphi_{tw_{i+h}} = \theta_{ti}^{(r)} \sum_{h \in H} \frac{a_h^{(r)} \varphi_{tw_{i+h}}}{\theta_{ti}^{(r)}} \frac{a_h}{a_h^{(r)}}.$$

Следовательно,

$$\log \frac{\theta_{ti}(a)}{\theta_{ti}^{(r)}} \geq \sum_{h \in H} \frac{a_h^{(r)} \varphi_{tw_{i+h}}}{\theta_{ti}^{(r)}} \log \frac{a_h}{a_h^{(r)}}.$$

Итак,

$$\ell_w(a) \geq Q(a \mid a^{(r)}) + \text{const},$$

где существенная часть $Q$ равна

$$Q(a \mid a^{(r)}) = \sum_{h \in H} B_h^{(r)} \log a_h,$$

а

$$B_h^{(r)} = a_h^{(r)} \sum_{i \in I_w} \sum_{t \in T} p_{ti}^{(r)} \frac{\varphi_{tw_{i+h}}}{\theta_{ti}^{(r)}}.$$

Причем миноризация точна в точке $a = a^{(r)}$:

$$Q(a^{(r)} \mid a^{(r)}) + \text{const} = \ell_w(a^{(r)}).$$

Максимизация $Q$ по симплексу $\Delta_w$ дает

$$a_h^{(r+1)} = \frac{B_h^{(r)}}{\sum_{s \in H} B_s^{(r)}}.$$

Так как

$$\sum_{s \in H} B_s^{(r)} = \sum_{i \in I_w} \sum_{t \in T} p_{ti}^{(r)} \frac{\sum_{s \in H} a_s^{(r)} \varphi_{tw_{i+s}}}{\theta_{ti}^{(r)}} = \sum_{i \in I_w} \sum_{t \in T} p_{ti}^{(r)} = |I_w|,$$

получаем ровно указанное обновление:

$$a_h^{(r+1)} = \underset{h \in H}{\text{norm}} \left( a_h^{(r)} \sum_{i=1}^{n} \sum_{t \in T} \sum_{k=-K}^{K} p_{ti}^{(r)} \frac{\varphi_{tw_{i+k}}}{\theta_{ti}^{(r)}} [w_i = w][k = h] \right).$$

Теперь из свойства миноризации следует монотонность:

$$\ell_w(a^{(r+1)}) \geq Q(a^{(r+1)} \mid a^{(r)}) + \text{const} \geq Q(a^{(r)} \mid a^{(r)}) + \text{const} = \ell_w(a^{(r)}).$$

Значит,

$$\ell_w(\alpha_w^{(r+1)}) \geq \ell_w(\alpha_w^{(r)}).$$

Так как $\Delta_w$ компактен, а $\ell_w$ является суммой логарифмов аффинных функций и сверху ограничена на $\Delta_w$, монотонная последовательность $\ell_w(\alpha_w^{(r)})$ имеет конечный предел.

Осталось показать условие стационарности предельных точек. Градиент блока имеет компоненты

$$\frac{\partial \ell_w}{\partial a_h}(a) = \sum_{i \in I_w} \sum_{t \in T} p_{ti}(a) \frac{\varphi_{tw_{i+h}}}{\theta_{ti}(a)} =: g_h(a).$$

Кроме того,

$$\sum_{h \in H} a_h g_h(a) = |I_w|.$$

Поскольку построенный MM-шаг является точным в первом порядке, стандартный аргумент сходимости MM-алгоритмов дает: если предельная точка $a^*$ не удовлетворяет условиям ККТ, то существует допустимое направление, вдоль которого производная $\ell_w$ положительна. Тогда в некоторой окрестности $a^*$ максимизация миноранты давала бы строгое, отделенное от нуля увеличение $\ell_w$, что противоречит сходимости значений $\ell_w(a^{(r)})$. Следовательно, всякая предельная точка $a^*$ удовлетворяет условиям ККТ: существует $\lambda$, такое что

$$g_h(a^*) \le \lambda, \qquad h \in H,$$

$$a_h^* \ge 0, \qquad \sum_{h \in H} a_h^* = 1,$$

$$a_h^*\big(g_h(a^*) - \lambda\big) = 0, \qquad h \in H.$$

Умножая равенства для активных координат на $a_h^*$ и суммируя, получаем $\lambda = |I_w|$.

Наконец, $\ell_w$ вогнута по $a$, поскольку это сумма функций вида log от неотрицательной аффинной функции. Поэтому любые условия ККТ для задачи максимизации на выпуклом множестве $\Delta_w$ являются достаточными: каждая предельная точка есть глобальный максимум блока.

Если глобальный максимум единственен, то все предельные точки последовательности совпадают с ним. Компактность $\Delta_w$ тогда означает, что вся последовательность $\alpha_w^{(r)}$ сходится к этому единственному максимуму. □

### 3.4. Итерационный алгоритм

Обучение чередует три шага: построение контекстных тематических распределений, вычисление апостериорных распределений тем и обновление параметров модели. Несколько проходов внимания $L$ позволяют согласовать $\theta_{ti}$ и $p_{ti}$ перед накоплением статистик.

**Algorithm 1** ЕМ-алгоритм для AAARTM

**Вход:** Батчи $B_1, \ldots, B_M$; параметры $|T|, K, L$

**Выход:** $(\varphi_{tw})_{T \times W}$, $(p_{ti})_{T \times n}$

Инициализировать $\varphi_{tw} := \text{norm}_t(\text{rand})$; $n_t := 1$

**for all** итераций $k = 1, \ldots, K$ **do**

  Инициализировать $n_{tw} := 0$; $N_{tw} := 0$

  **for all** батчей $B_m$, $m = 1, \ldots, M$ **do**

    $p_{ti} := \varphi_{tw_i}, \quad i \in B_m$

    **for all** проходов $\ell = 1, \ldots, L$ **do**

      $\theta_{ti} := \sum\limits_{k=-K}^{K} \alpha_{w_i k} p_{tw_{i+k}}, \quad i \in B_m$

      $p_{ti} := \text{norm}_{t \in T} \left( \frac{p_{ti} \theta_{ti}}{n_t} \right), \quad i \in B_m$

    **end for**

    $N_{tw} := N_{tw} + \sum\limits_{i \in B_m} \frac{p_{ti}}{\theta_{ti}} \sum\limits_{k=-K}^{K} \alpha_{w_i k} [w_{i+k} = w]$

    $n_{tw} := n_{tw} + \sum\limits_{i \in B_m} [w_i = w] p_{ti}$

  **end for**

  $\varphi^*_{tw} := \text{norm}_{t \in T}(n_{tw})$

  $\varphi_{tw} := \text{norm}_{t \in T} \left( n_{tw} + \varphi^*_{tw} \frac{\partial R(\Phi^*)}{\partial \varphi^*_{tw}} + \varphi^*_{tw} N_{tw} \right)$

  $\alpha_{wh} := \text{norm}_{h \in \{-K, \ldots, K\}} \left( \alpha_{wh} \sum\limits_{i=1}^{n} \sum\limits_{t \in T} \sum\limits_{k=-K}^{K} p_{ti} \frac{\varphi_{tw_{i+k}}}{\theta_{ti}} [w_i = w][k = h] + \alpha_{wh} \frac{\partial R}{\partial \alpha_{wh}} \right)$

  $n_t := \sum\limits_{w \in W} \varphi_{tw} n_w$

**end for**

# 4. Данные

## 4.1. Датасеты

Экспериментальная проверка проводится на трех англоязычных корпусах: 20 Newsgroups и AG News. Они различаются жанром, длиной документов и характером тематической неоднородности, поэтому позволяют оценивать модель не только на одном типе текстов. В частности, 20 Newsgroups содержит сообщения дискуссионных групп, а AG News содержит короткие новостные тексты. Такое сочетание корпусов важно для рассматриваемой постановки, поскольку локально-контекстная модель должна корректно работать как при длинных документах с несколькими смысловыми фрагментами, так и при коротких текстах, где число наблюдаемых слов ограничено.

Все корпуса имеют фиксированное разбиение на обучающую и тестовую части, и в экспериментах используется именно это разбиение. Метки классов присутствуют в исходных наборах данных, однако не используются при обучении тематических моделей: рассматривается ненаблюдаемая постановка, в которой модель получает только тексты документов. Разметка служит для документирования происхождения данных и может быть использована отдельно при анализе интерпретируемости тем. Основные характеристики используемых версий корпусов приведены в таблице 1.

Таблица 1: Корпуса, использованные в экспериментальной оценке. Указаны размеры обучающей и тестовой частей в используемой версии данных.

| Корпус | Классы | Train | Test | Источник |
|---|---|---|---|---|
| 20news | 20 | 11314 | 7532 | scikit-learn |
| AG news | 4 | 96000 | 7600 | Hugging Face Datasets |

### 4.1.1. 20 Newsgroup

20news является стандартным корпусом сообщений из 20 тематических групп Usenet [14]. В данной работе используется версия, распространяемая в составе scikit-learn; из текстов предварительно удаляются заголовки, подписи и цитируемые фрагменты писем, чтобы уменьшить влияние служебной информации и повторов из переписки. Обучающая часть содержит 11314 документов, тестовая – 7532 документа.

Корпус удобен для тематического моделирования по двум причинам. Во-первых, тематики групп достаточно различимы на уровне лексики, что позволяет ожидать интерпретируемых распределений “тема–слово”. Во-вторых, отдельные сообщения часто содержат служебные фрагменты, цитирование предыдущих сообщений и неформальный стиль. Даже после удаления заголовков, подписей и цитат данные остаются неоднородными, поэтому 20news проверяет устойчивость модели к шумной пользовательской лексике и нерегулярной структуре документов.

### 4.1.2. Ag news

AG news является новостным корпусом для четырехклассовой тематической классификации, построенным на основе AG's News Corpus и широко используемым в экспериментах по классификации текста [15]. В экспериментах используется 96000 обучающих и 7600 тестовых документов. Тексты в этом корпусе короче, чем в 20news, поэтому для него особенно важна корректная обработка локальных совместных встречаемостей слов при ограниченном документальном контексте.

Новостной стиль делает AG news менее разговорным, чем предыдущий корпус, однако короткая длина документов усложняет оценивание тематических смесей: в каждом документе наблюдается меньше слов, а значит, отдельные случайные токены сильнее влияют на локальную статистику. Этот корпус используется как проверка того, насколько метод сохраняет устойчивость при уменьшении объема контекста внутри документа.

## 4.2. Предобработка

Для всех корпусов используется один и тот же конвейер предобработки, чтобы различия в результатах не были следствием различных правил очистки текста. Пусть $d$ обозначает исходный документ. Сначала текст приводится к нижнему регистру. Затем все символы, не являющиеся латинскими буквами, заменяются пробелами; таким образом, числа, пунктуация, URL и прочие специальные символы не попадают в словарь как отдельные термы. После этого выполняется токенизация средствами NLTK, стемминг Портера и удаление английских стоп-слов из списка NLTK. На выходе каждый документ представлен

последовательностью токенов

$$d = (w_1, \ldots, w_{n_d}), \qquad w_i \in W,$$

где $W$ – словарь корпуса после предобработки.

Словарь строится только по обучающей части соответствующего корпуса. При преобразовании тестовой части применяется тот же конвейер очистки и токенизации, но токены, отсутствующие в обучающем словаре, отбрасываются. Этот протокол исключает утечку информации из тестовой выборки на этапе построения словаря и обеспечивает одинаковое пространство термов для обучения и оценки. Если после удаления вне словаря тестовый документ не содержит ни одного токена, он остается пустым с точки зрения последовательности термов; такие случаи учитываются в диагностике предобработки, но не требуют изменения словаря.

Границы документов сохраняются явно как для обучающей, так и для тестовой части. Это существенно для локально-контекстной модели: окна контекста строятся внутри одного документа и не переносятся через границу соседних текстов. Тем самым статистика локальной совместной встречаемости не смешивает независимые документы, что особенно важно для коротких текстов AG news.

## 5. Вычислительный эксперимент

### 5.1. Метрики

*Согласованность тем.* Для оценки согласованности тем используются метрики NPMI@10 и $C_v$@10, вычисляемые по десяти наиболее вероятным словам каждой темы. Значение NPMI рассчитывается на основе совместной встречаемости слов на уровне документов в обучающей выборке. Метрика $C_v$ вычисляется по токенизированным обучающим документам.

*Разнообразие тем.* Для измерения разнообразия тем используется TD@25 — доля уникальных слов среди всех top-25 слов тем:

$$\text{TD@25} = \frac{\left|\bigcup_{t=1}^{T} \mathcal{W}_t^{(25)}\right|}{\sum_{t=1}^{T} \left|\mathcal{W}_t^{(25)}\right|}.$$

*Классификация документов.* Для каждого документа строится тематический вектор, после чего на этих представлениях обучается классификатор логистической регрессии. В вариантах AARTM документные векторы получаются усреднением апостериорных распределений тем на уровне токенов внутри соответствующего документа. Качество классификации оценивается с помощью macro-F1.

*Perplexity.* Perplexity используется для оценки того, насколько хорошо модель предсказывает слова в отложенных документах. Чем ниже значение perplexity, тем выше предсказательная способность модели.

*Время работы.* Время обучения измеряется как фактическое астрономическое время в секундах. Так как процедуры предобработки различаются для разных базовых методов, они не включаются в расчет времени работы.

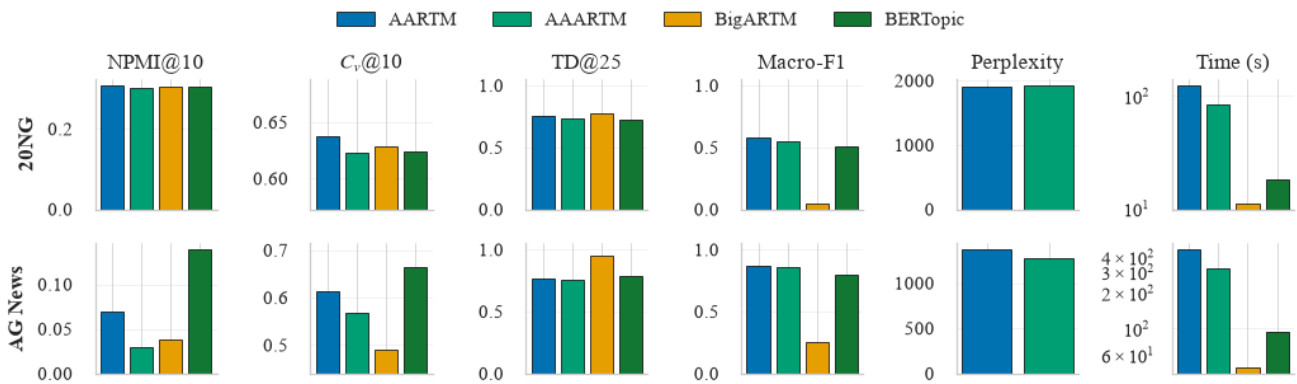
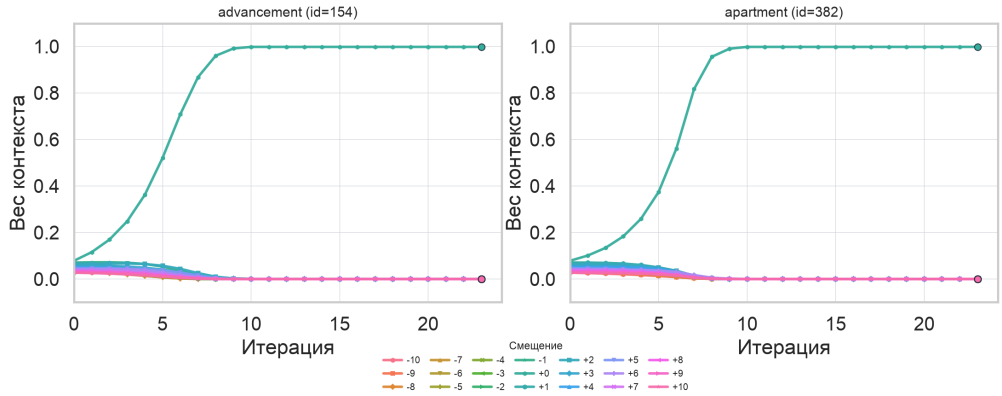



Рис. 1: Сравнение моделей по качеству и разнообразию тем.

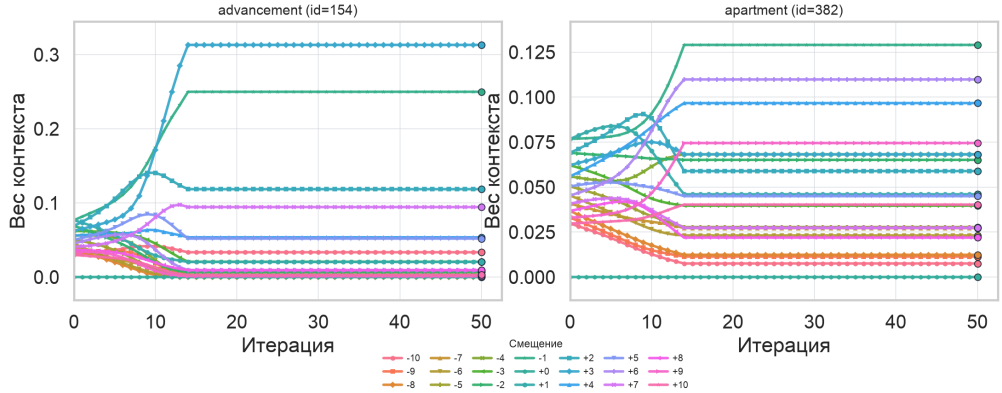
### 5.2. Переобучение и самовнимание

Результаты на рисунке 1 демонстрируют принципиальное различие между двумя режимами формирования контекста. В режиме self-aware, где центральный токен входит в контекстное окно, обучаемые веса быстро становятся почти вырожденными: основная масса распределения переносится на позицию $k = 0$. Это объясняется тем, что сам центральный токен является наиболее сильным предиктором собственного значения. В результате модель получает возможность максимизировать правдоподобие за счет тривиального самообъяснения, практически не используя информацию из соседних слов.

Напротив, в режиме self-exclude центральная позиция исключается из множества доступных элементов контекста. Поэтому модель не может опираться на самоописание токена и вынуждена распределять вес между окружающими словами. На графиках видно, что в этом случае веса не схлопываются в одну позицию, а формируют нетривиальные профили, зависящие от слова и набора данных. Это указывает на то, что обучаемые коэффициенты действительно отражают относительную важность различных позиций локального контекста для предсказания центрального слова.




(a) Веса контекста в режиме self-aware.




(b) Веса контекста в режиме self-exclude.

Рис. 2: Сравнение весов контекста при разных режимах самовнимания.

### 5.3. Коэффициенты внимания

На рисунке 3 показано усредненное распределение весов внимания по позициям локального контекста. Видно, что максимальные веса получают слова, расположенные непосредственно рядом с центральным токеном, а по мере увеличения расстояния вклад позиции постепенно убывает. Такая форма распределения напоминает геометрическое затухание: каждая следующая позиция в среднем оказывается менее информативной, чем предыдущая.

Этот результат согласуется с естественной лингвистической интуицией: ближайшие слова обычно сильнее связаны с центральным словом синтаксически и семантически, тогда как более удаленные элементы контекста содержат меньше информации для его предсказания. При этом распределение остается приблизительно симметричным относительно центральной позиции: как левые,

так и правые соседи вносят заметный вклад, но этот вклад уменьшается при удалении от рассматриваемого токена.

Таким образом, обученные коэффициенты внимания отражают не случайный шум, а осмысленную зависимость от расстояния до центрального слова. Модель автоматически восстанавливает ожидаемую структуру локального контекста: наибольшее значение имеют ближайшие позиции, а дальние слова используются существенно слабее.

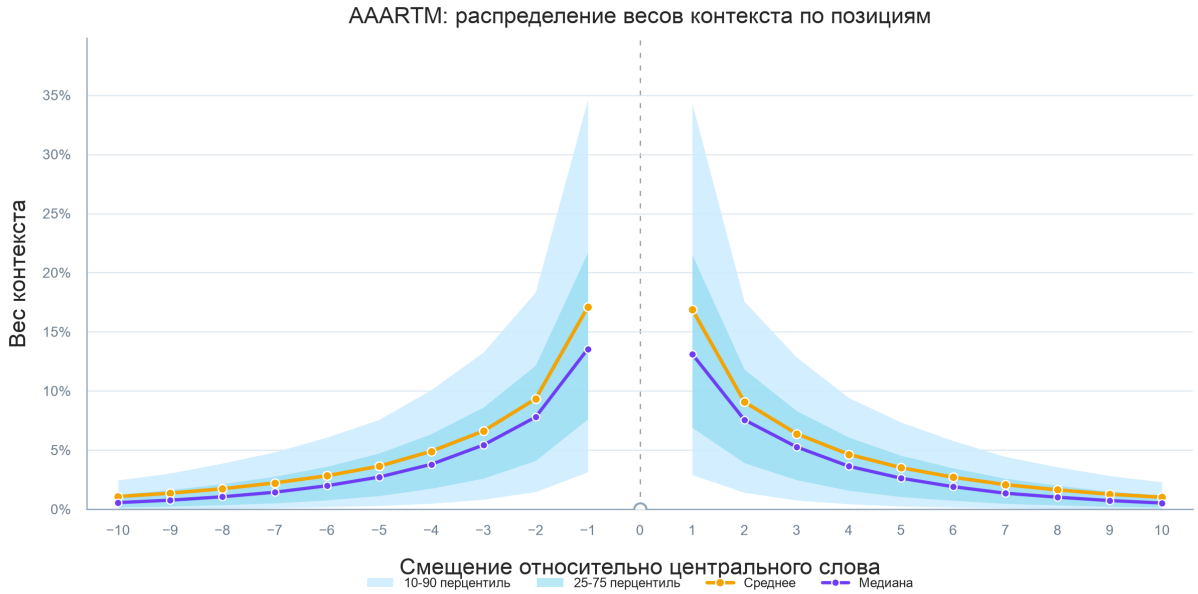



Рис. 3: Распределение весов по контексту.

## Заключение

В работе предложена и исследована локально-контекстная тематическая модель с обучаемыми коэффициентами внимания. Модель сохраняет вероятностную интерпретацию тем через распределения $p(t \mid w)$ и $p(w \mid t)$, но заменяет глобальное документное представление локальным контекстом позиции. Тем самым она занимает промежуточное место между классическими моделями мешка слов и нейросетевыми контекстными методами: порядок слов используется явно, но темы остаются нормированными и интерпретируемыми распределениями.

Основное отличие от фиксированной AARTM состоит в том, что веса локального окна не задаются заранее через расстояние до центрального слова, а обучаются как параметры смещения и слова. Такая параметризация позволяет модели адаптировать форму контекста к данным: разные слова могут использовать разные части окна, а вклад центрального токена может быть ограничен, чтобы избежать тривиального самопредсказания.

Теоретическая часть работы включает формулировку целевой функции, EM-алгоритм для локально-контексной модели с обучаемым вниманием. Для локального блока коэффициентов показана монотонность нерегуляризованного обновления и сходимость предельных точек к условиям Каруша–Куна–Таккера.

Эксперименты на 20 Newsgroups и AG News показывают, что обучаемые коэффициенты контекста ведут себя содержательно и отражают структуру локального окружения слова. При включении центрального токена в контекст возникает вырожденное поведение: почти весь вес быстро переносится на позицию $k = 0$, поскольку слово лучше всего предсказывает само себя. Исключение центрального элемента устраняет это самовлияние, после чего модель начинает распределять вес между соседними позициями и выделять наиболее информативные слова контекста. В среднем полученное распределение убывает с расстоянием от центрального токена, что согласуется с интуицией о том, что ближайшие слова обычно сильнее связаны с ним синтаксически и семантически, чем удаленные элементы контекста. На коротких текстах новая модель не улучшая значение перплексити, зато на длинных дает небольшой выигрышь. При этом значения когерентностей остаются ниже, чем у AARTM.

Дальнейшее развитие работы связано с совместной оптимизацией перплексии и когерентности, более строгим анализом регуляризованных обновлений для обучаемых коэффициентов контекста, а также расширением сравнения на

корпуса с другой синтаксической и жанровой структурой. Отдельный практический интерес представляет анализ выученных весов $\alpha_{wh}$: они могут служить не только параметрами модели, но и диагностикой того, какие части локального контекста важны для разных тем.

## Список литературы


1. Hofmann Thomas. Probabilistic Latent Semantic Indexing // Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval. — 1999. — Pp. 50–57.

2. Hofmann Thomas. Unsupervised Learning by Probabilistic Latent Semantic Analysis // Machine Learning. — 2001. — Vol. 42, no. 1–2. — Pp. 177–196.

3. Blei David M., Ng Andrew Y., Jordan Michael I. Latent Dirichlet Allocation // Journal of Machine Learning Research. — 2003. — Vol. 3. — Pp. 993–1022. — URL: https://jmlr.org/papers/v3/blei03a.html.

4. Wallach Hanna M. Topic Modeling: Beyond Bag-of-Words // Proceedings of the 23rd International Conference on Machine Learning. — ACM, 2006. — Pp. 977–984.

5. A Biterm Topic Model for Short Texts / Xiaohui Yan, Jiafeng Guo, Yanyan Lan, Xueqi Cheng // Proceedings of the 22nd International Conference on World Wide Web. — 2013. — Pp. 1445–1456.

6. Zuo Yuan, Zhao Jichang, Xu Ke. Word Network Topic Model: A Simple but General Solution for Short and Imbalanced Texts // Knowledge and Information Systems. — 2016. — Vol. 48, no. 2. — Pp. 379–398.

7. Rahimi Marziea, Zahedi Morteza, Mashayekhi Hoda. A Probabilistic Topic Model Based on Short Distance Co-Occurrences // Expert Systems with Applications. — 2022. — Vol. 193. — P. 116518.

8. Vorontsov Konstantin. Additive Regularization for Topic Models of Text Collections // Doklady Mathematics. — 2014. — Vol. 89, no. 3. — Pp. 301–304.

9. Vorontsov Konstantin. Rethinking Probabilistic Topic Modeling from the Point of View of Classical Non-Bayesian Regularization // Data Analysis and Optimization. — Springer, 2023. — Pp. 397–422.

10. Grootendorst Maarten. BERTopic: Neural Topic Modeling with a Class-Based TF-IDF Procedure // arXiv preprint arXiv:2203.05794. — 2022. — URL: https://arxiv.org/abs/2203.05794.

11. Vorontsov Konstantin, Dyakov Ilya. Attentive Additively Regularized Topic Modeling for Local Context. — Manuscript under review.

12. Vorontsov Konstantin, Potapenko Anna. Additive Regularization of Topic Models // Machine Learning. — 2015. — Vol. 101, no. 1. — Pp. 303–323.

13. Irkhin Ilya, Vorontsov Konstantin. Convergence of the Algorithm of Additive Regularization of Topic Models // Trudy Instituta Matematiki i Mekhaniki UrO RAN. — 2020. — Vol. 26, no. 3. — Pp. 56–68.

14. Lang Ken. Newsweeder: Learning to Filter Netnews // Machine Learning Proceedings 1995. — 1995. — Pp. 331–339.

15. Zhang Xiang, Zhao Junbo, LeCun Yann. Character-level Convolutional Networks for Text Classification // Advances in Neural Information Processing Systems. — 2015. — URL: https://proceedings.neurips.cc/paper/2015/hash/250cf8b51c773f3f8dc8b4be867a9a02-Abstract.html.