Министерство науки и высшего образования Российской Федерации
Московский физико-технический институт
(государственный университет)
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»
при Вычислительном центре им. А. А. Дородницына РАН

Алексеев Василий Антонович

Внутритекстовая когерентность как мера интерпретируемости тематических моделей текстовых коллекций

03.03.01 — Прикладные математика и физика

Бакалаврская диссертация

Научный руководитель

д. ф.-м. н.
Воронцов Константин Вячеславович

Москва
2018

Содержание

Аннотация

Работа посвящена проблеме автоматической оценки качества тем, найденных с помощью тематических моделей. Предлагается новый подход к оценке качества тем — *внутритекстовая когерентность*. Суть подхода в том, что при оценке качества темы учитывается весь текст коллекции документов целиком, анализируется распределение темы по всем словам корпуса. Эта идея согласуется с гипотезой о сегментной структуре текстов: все темы представлены в тексте в виде сегментов, слова каждой темы образуют в тексте группы. Когерентности же по топ-словам, с которыми сравниваются предлагаемые методы, учитывают совместные встречи лишь некоторого количества самых частых слов темы (топ-слов). Но из частых совстречаемостей нескольких топовых слов темы не следует, что и все слова темы представлены в тексте группами. Помимо новых методов подсчёта когерентности темы, также предлагается полуавтоматический способ оценки качества функций когерентности: по корреляции с качеством сегментации текста тематическими моделями. Вычислительные эксперименты проводятся на коллекции статей научно-популярного контента «ПостНаука». По введённой функции качества некоторые из предложенных когерентностей превосходят методы, основанные на топ-словах.

Ключевые слова: *тема, когерентность темы, интерпретируемость темы, тематическое моделирование, тематическая модель, сегментация текста, совстречаемость, топ-слово, BigARTM, анализ текстов, машинное обучение*

1. Введение

Тематическое моделирование — это область анализа текстов, которая нацелена на то, чтобы найти и описать темы, которые затрагиваются в коллекции текстовых документов. Теме соответствует набор слов, которые часто совместно встречаются в тексте. Тематическое моделирование используется в информационном поиске [9], категоризации документов [11], анализе данных социальных сетей [17, 16], рекомендательных системах [9, 7], разведочном поиске [22].

Тематическая модель принимает на вход коллекцию документов. На выходе она выдаёт набор тем, которые затрагиваются в документах, информацию о распределении тем в каждом документе. Каждая тема, найденная моделью, характеризуется списком слов [10].

Интерпретируемость — свойство темы, состоящее в том, как хорошо слова темы описывают некоторую реальную область [24], может ли человек по списку слов темы дать ей конкретное название. Темы, полученные тематическими моделями, могут быть не понятны человеку, состоять из слов разных слабо связанных областей [8]. Таким образом, интерпретируемость темы оценивается человеком по списку составляющих её слов. Но это долго и затратно.

Был предложен автоматический способ оценки интерпретируемости темы [15, 6], названный *когерентностью*. При подсчёте когерентности оцениваются вероятности совместных встреч в тексте топовых (самых частых) слов темы. Было показано, что когерентность темы коррелирует с экспертными оценками её интерпретируемости [6]. Однако способ подсчёта когерентности темы по её топ-словам имеет недостатки.

В данной работе ставится две цели.

Во-первых, показать, в чём состоит проблема когерентностей по топ-словам. Проблема же этих методов в том, что топ-слова составляют лишь малую часть от общего запаса слов коллекции; доля текста, покрываемая топ-словами, никак не контролируется. Поэтому ошибочно считать, что когерентность темы по топ-словам может служить полноценной оценкой интерпретируемости.

Во-вторых, предложить альтернативный, *внутритекстовый*, подход к подсчёту когерентности, продемонстрировать его состоятельность. Внутритекстовая когерентность темы есть оценка того, насколько близки по смыслу слова, которые тематическая модель относит к данной теме и которые расположены в тексте недалеко друг от друга.

Для достижения второй из поставленных целей попутно решается задача по оценке качества функций когерентности. А именно, качество функции когерентности оценивается по Спирмановской корреляции значений когерентностей и значений качества для ряда тематических моделей. Качество тематической модели при этом оценивается исходя из того, как точно модель способна предсказывать сегментную структуру текста, сшитого из сегментов разных тем.

1.1. Связанные работы

В работах [5], [8] и [15] представлена следующая схема, по которой можно оценить интерпретируемость темы (состоящая из этапов 1 и 2-a в случае экспертной оценки или 1 и 2-b в случае оценки с помощью когерентностей, опирающихся на совстречаемости топовых слов).

1 Для каждой темы выделяется небольшой набор слов (обычно это список десяти самых часто всречающихся слов темы, но возможны и другие варианты [4]).

2-a Приглашается эксперт, чтобы по этому списку слов выдать вердикт о том, хорошая тема или нет.

2-b Собирается информация о частотах совстречаемости данных слов внутри текста. Далее с использованием полученной информации проводятся дополнительные вычисления, считаются когерентности.

Когерентности, которые опираются на анализ совстречаемостей топовых слов темы, далее в работе называются *когерентностями по том-словам* [6, 8]. Эти методы привлекательны тем, что они просты, учитывают лишь небольшое множество слов. Но в этом и их главный недостаток. Много информации теряется, когда тема как вероятностное распределение сводится к списку из пяти или десяти слов. Причём независимо от того, каким именно способом анализируются эти слова: будь то оценка интерпретируемости человеком при просмотре топ-слов или же подсчёт когерентности на основе совстречаемостей этих слов внутри текста.

1.2. Тематическое моделирование

Введём следующие условные обозначения, которы далее будут использоваться в работе

- W — множество слов (словарь)
- D — множество документов
- W_d — упорядоченное мультимножество слов, из которых состоит материальный аналог документа $d \in D$
- T — множество тем
- n_{dw} — количество вхождений слова w в W_d
- ν_{wd} — частота появления слова w в W_d

Приведём несколько важных в тематическом моделировании определений и гипотез

Гипотеза 1. Каждое слово в каждом документе связано с некоторой темой.

Гипотеза 2. $D \times W \times T$ — дискретное вероятностное пространство.

Определение 1. Коллекция — это независимая одинаково распределённая выборка

$$(d_i, w_j, t_k) \sim p(d, w, t)$$

Заметка. При этом d_i и w_j — наблюдаемые, а t_k — скрытые.

Гипотеза 3. Для определения тем, затрагиваемых в коллекции документов, не важен порядок документов в D.

Гипотеза 4 (Мешка слов)**.** Для определения тем, затрагиваемых в коллекции документов, не важен порядок слов в W_d.

Гипотеза 5 (Условной независимости)**.**

$$p(w \mid d, t) = p(w \mid t)$$

Тематическая модель с помощью тем описывает вероятность встретить слово w в документе d

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

Теперь можно ввести ещё несколько обозначений

- $\phi_{wt} \equiv p(w \mid t)$ — вероятность того, что слово w относится к теме t
- $\theta_{td} \equiv p(t \mid d)$ — вероятность того, что тема t относится к документу d
- $\boldsymbol{w}$ — некоторый вектор, который можно поставить в соответствие слову w. В данной работе обозначение $\boldsymbol{w}$ в некоторых местах будет использовано для вектора с компонетнами $p(t \mid d, w)$, иногда же — для вектора с компонентами $p(t \mid w)$.

В контексте тематического моделирования удобно дать следующее определение темы:

Определение 2. О теме t можно думать как о вероятностном распределении на множестве слов: $p(w \mid t),\ w \in W$.

Пример 1. Тема «Солярис», скорее всего, будет иметь распределение, сконцентрированое в словах *океан, фиолетовый, станция, хари, космос, разум, лаборатория, симметриада, фантом, контакт, депрессия*. Напротив, вероятности таких слов, как, например, *фовизм* или *иероглиф*, будут очень малы.

Тематическая модель может быть описана с помощью двух распределений, которые можно представить в виде матриц Φ и Θ:

Определение 3. Матрица слов в темах

$$\Phi \equiv \left(\phi_{wt}\right)_{W\times T} \equiv \left(p(w \mid t)\right)_{W\times T}$$

Определение 4. Матрица тем в документах

$$\Theta \equiv \left(\theta_{td}\right)_{T\times D} \equiv \left(p(t \mid d)\right)_{T\times D}$$

Заметка. Матрицы Φ и Θ имеют неотрицательные нормированные столбцы, представляющие дискретные распределения.

Ещё одна матрица, которая несёт информацию о коллекции и известна с самого начала:

Определение 5. Матрица частот слов в документах

$$F \equiv \left(\nu_{td}\right)_{W\times D}$$

В тематическом моделировании решается задача матричного разложения (рисунок 1)

$$F = \left(\nu_{td}\right)_{W\times D} \approx \left(p(w \mid d)\right)_{W\times D} = \Phi \times \Theta \tag{1}$$

Существует несколько тематических моделей. Например, модель PLSA [2] при решении задачи (1), максимизирует правдоподобие выборки, или её плотность распределения $p(D, \Phi, \Theta)$:

$$p(D,\Phi,\Theta) = \prod_{d\in D}\prod_{w\in W_d} p(d,w)^{n_{dw}} = \prod_{d\in D}\prod_{w\in W_d} p(w \mid d)^{n_{dw}}\, \underbrace{p(d)^{n_{dw}}}_{\text{const}} \to \max_{\Phi,\Theta}$$

Вместо правдоподобия удобнее максимизировать его логарифм:

$$L(\Phi,\Theta) \equiv \ln p(D,\Phi,\Theta) = \sum_{d\in D}\sum_{w\in W_d} n_{dw} \underbrace{\ln \sum_{t\in T} \phi_{wt}\theta_{td}}_{p(w|d)} \to \max_{\Phi,\Theta}$$

Задача максимизации решается при ограничениях нормированности и неотрицательности столбцов матриц Φ и Θ:

$$\begin{cases} \sum\limits_{w\in W} \phi_{wt} = 1 & \phi_{wt} \geq 0 \\ \sum\limits_{t\in T} \theta_{td} = 1 & \theta_{td} \geq 0 \end{cases}$$

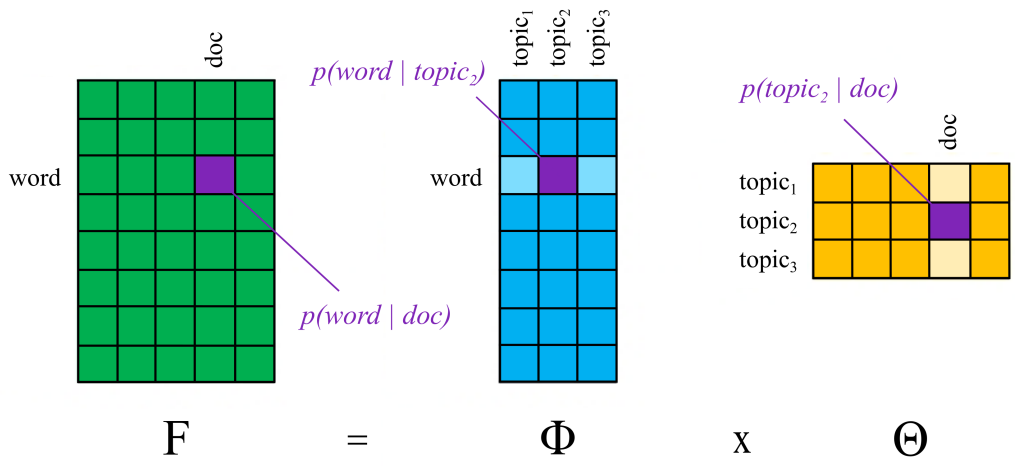

Рис. 1: Тематическая модель по известным частотам слов в документах выявляет скрытые (*латентные*) темы: для каждой темы становится известной информация о том, какие слова к ней относятся и в каких документах она встречается.

Известно [20], что точка (Φ, Θ) локального экстремума поставленной задачи удовлетворяет системе уравнений (1.2) (см. также рисунок 2):

$$\begin{cases} p_{tdw} = \dfrac{\phi_{wt}\theta_{td}}{\sum\limits_{s} \phi_{ws}\theta_{sd}} & p_{tdw} \equiv p(t \mid d, w) \\ \phi_{wt} = \dfrac{n_{wt}}{n_t} & n_{wt} = \sum\limits_{d \in D} n_{dw} p_{tdw} \qquad n_t = \sum\limits_{w \in W} n_{wt} \\ \theta_{td} = \dfrac{n_{td}}{n_d} & n_{td} = \sum\limits_{w \in W_d} n_{dw} p_{tdw} \qquad n_d = \sum\limits_{t \in T} n_{td} \end{cases}$$

Решение системы (1.2) находится с помощью метода простых итераций, который в данном случае совпадает с EM-алгоритмом [20]. E-шаг — вычисление p_{tdw}, M-шаг — ϕ_{wt} и θ_{td}.

Заметка. Параметры ϕ_{wt} и θ_{td} можно инициализировать случайными значениями.

В модели LDA [3] дополнительно вводится предположение, что столбцы матриц Φ и Θ являются случайными векторами из распределения Дирихле.

При аддитивной регуляризации тематических моделей [20, 19, 18] к критерию, по которому проводится оптимизация, добавляется взвешенная сумма произведений регуляризаторов $R_i(\Phi, \Theta)$ при $\sum_{i=1}^{k} \tau_i = 1$, $\tau_i \geq 0$

$$L(\Phi, \Theta) + \sum_{i=1}^{k} \tau_i R_i(\Phi, \Theta) \to \max_{\Phi, \Theta}$$

Модель LDA интерпретируется в ARTM через введение определённых регуляризаторов. Модель PLSA соответствует случаю, когда регуляризаторов нет.

Существуют разные виды регулязаторов $R(\Phi, \Theta)$, в зависимости от целей, которых хотят достичь их введением [20]. Так, они могут могут использоваться для

Рис. 2: Эволюция тематической модели в зависимости от числа итераций при решении системы (1.2) (в зависимости от числа обновлений матрицы Ф)

- сглаживания фоновой темы, концентрации её вероятностого распределения в словах общей лексики
- разреживания предметной темы, сгущения её вероятностого распределения в небольшом количестве характеризующих её слов
- декорреляции тем как столбцов матрицы Φ

Перплексия — один из *внутренних критериев* качества тематических моделей [18]:

$$\mathscr{P}(D, W, \Phi, \Theta) = \exp\left(-\frac{1}{N} L(\Phi, \Theta)\right) = \exp\left(-\frac{1}{N} \sum_{d \in D} \sum_{w \in W_d} n_{dw} \ln p(w \mid d)\right) \tag{2}$$

где $N = \sum_{d \in D} \sum_{w \in W_d} n_{dw} = \sum_{d \in D} |W_d|$ — длина коллекции. Чем больше правдоподобие коллекции $p(D, W, T)$, тем меньше перплексия. Иными словами, чем лучше модели удаётся предсказывать появление слов w в документах $d \in D$ коллекции, тем меньше перплексия.

1.3. Качество тем

Гипотеза, на которую опирается всё дальнейшее изложение про оценку качества тем:

Гипотеза 6 (О сегментной структуре текста)**.** Тексты естественного языка сегментированы, состоят из сегментов разных тем.

***Следствие* 6.1.** *Слова каждой темы расположены группами, а не разбросаны по тексту беспорядочно.*

***Следствие* 6.2.** *Цель тематического моделирования — правильная сегментация исходного текста на тематически однородные фрагменты, состоящие из небольшого числа тем (одной, или нескольких смежных).*

Тема $t \in T$, полученная с помощью тематической модели, может характеризоваться словами, которые

- вместе ни с чем не ассоциируются у человека
- разбросаны по тексту в случайном порядке

Мы приходим к понятию *качества темы*.

Определение 6. Качество темы — абстрактное понятие, отражающее то, насколько хорошо распределение темы в тексте соответствует гипотезе о сегментной структуре текста.

Определение 7. Функция качества темы $q(t)$

$$q(t_1) < q(t_2) \leftrightarrow \text{ тема } t_1 \text{ менее качественная, чем тема } t_2$$

Проблема в том, что истинная функция качества $q(\cdot)$ не известна. Поэтому для оценки качества было предложено оценивать *интерпретируемость* темы: как хорошо человеку удаётся по

самым частым словам темы (топ-словам) дать ей подходящее название. Недостаток такого способа в том, что для оценки интерпретируемости темы необходимо привлекать экспертов. А также в том, что несколько самых частых слов несут лишь часть информации о теме как о вероятностном распределении на множестве слов и совсем не несут информации о распределении тем в документах.

2. Когерентности по топ-словам

2.1. Ньюман, Мимно

В работе рассматриваются два способа посчитать когерентность темы по топ-словам: по Ньюману [6] и по Мимно [8]. Оба способа можно описать одной формулой (3), оценивающей неслучайность того, что топ-слова темы встречаются недалеко друг от друга в тексте:

$$\operatorname{coh}\left(D, W, \boldsymbol{\phi}_{\cdot t}\right) = \underset{w_i, w_j \text{ from } k \text{ top-words}}{\text{Average}} \operatorname{PMI}(w_i, w_j) \tag{3}$$

где $\operatorname{PMI}(w_i, w_j)$ (Pointwise Mutual Information) — показатель неслучайности совместной встречи слов w_i и w_j в тексте внутри некоторого окна слов. Отличие между Ньюманом и Мимно — в способе оценить $\operatorname{PMI}(w_i, w_j)$:

$$\operatorname{PMI}_{\text{Newman}}(w_i, w_j) = \ln \frac{p(w_i, w_j)}{p(w_i)p(w_j)} \tag{4}$$

$$\operatorname{PMI}_{\text{Mimno}}(w_i, w_j) = \ln \frac{D(w_i, w_j) + 1}{D(w_i)} \tag{5}$$

Поясним обозначения в формулах выше

- $p(w_i)$, $p(w_i, w_j)$ — вероятности встретить слово w_i и два слова w_i, w_j в одном окне некоторого размера в тексте
- $D(w_i)$, $D(w_i, w_j)$ — количество документов, содержащих слово w_i и два слова w_i, w_j в окне слов некоторого размера

Для обозначения встречи двух топовых слов в одном окне внутри текста можно ввести специальное определение

Определение 8. Совстречаемость слов $U \subseteq W$ — факт нахождения двух слов $w, u \in U$ в одном текстовом окне в W_d для некоторого $d \in D$

$$w \mapsto (w, i) \in W_d,\ u \mapsto (u, j) \in W_d, \quad |j - i| \leq \text{window}$$

Зная когерентности отдельных тем, можно оценить и качество всей тематической модели:

Определение 9 (Когерентность тематической модели)**.** Среднее значение когерентности по темам T модели.

Особенность подхода к оценке интерпретируемости через совстречаемости топ-слов в том, что учитывается лишь фиксированное число топовых слов темы (в формуле (3) это число k). Это существенный недостаток. Некоторые последствия такого ограничения будут проиллюстрированы далее.

Также отметим, что

Заметка. Из частых совстречаемостей *определённого числа* слов темы не следует, что тема представлена в тексте в виде однородных сегментов (то есть распределение темы по тексту согласуется с гипотезой о сегментной структуре текста (6)).

2.2. Недостатки когерентностей по топ-словам

Когерентности по топ-словам имеют недостатки, которые и призваны исправить *внутритекстовые* когерентности, которые будут обсуждаться в соответствующем разделе (3).

Когерентности по Ньюману и Мимно (3) делают подсчёты лишь с учётом фиксированного числа слов. Обычно рассматривается набор из 10 самых частых слов темы. Этого не достаточно для качественного покрытия всей текстовой коллекции. Например, на рисунке 3 наглядно показано, насколько мала часть топовых слов темы от общего числа слов в документе.

Малое покрытие коллекции десятью словами приводит к тому, что часть информации о распределении темы в тексте не учитывается, теряется (рисунок 4).

Можно подсчитать доли, занимаемые совстречаемостями топ-слов. В таблице 1 представлены результаты для подсчётов долей текста, занимаемых близкими словопозициями топ-слов (совстречаемостями) для тем двух тематических моделей. Первая — модель «ПостНауки» с 19 темами — будет обсуждаться подробно в разделе 4.2. Вторая — модель Википедии — как в [5], построена по выборке статей из Википедии. Экспертами были просмотрены десять топ-слов каждой из тем этой модели, и она была признана лучшей из рассматривавшихся в [5]. В модели Википедии 50 тем. Из таблицы 1 видно, что топ-слова могут покрывать часть корпуса, сравнимую лишь с **1%–2%**.

Таблица 1: Доля слов корпуса, участвующих при подсчёте совстречаемостей десяти топовых слов тем для двух моделей: «ПостНаука» (19 тем) и Википедия (50 тем). Видно, что процент совстречаемостей для топ-слов всех тем от всех словопозиций составляет **1%–2%**

	ПостНаука, %	Википедия, %
Min	0.016	0.0065
Max	0.28	0.11
Median	0.048	0.029
Mean	0.062	0.036
Total	**1.2**	**1.7**

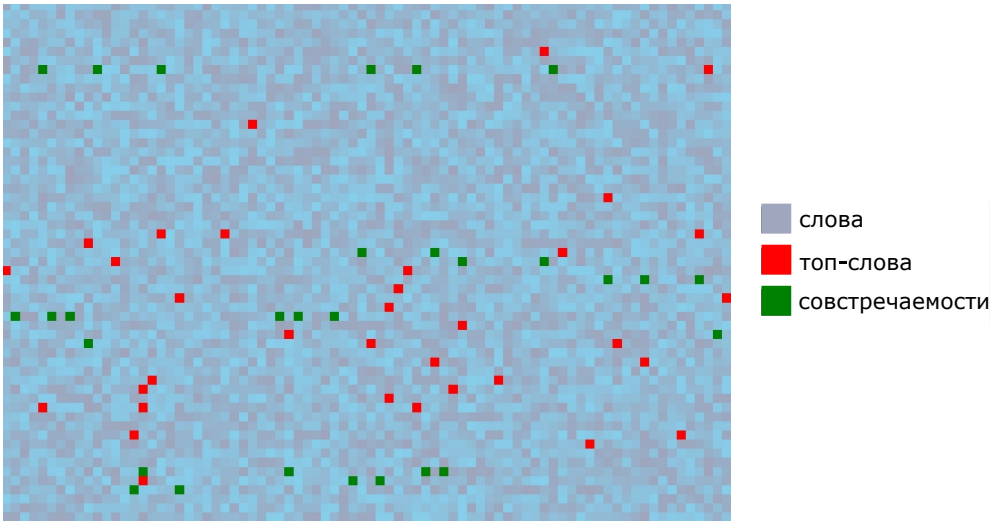

Рис. 3: Выделены позиции, занимаемые десятью самыми частыми словами (топ-словами) некоторой темы. Видно, что эти десять топ-слов покрывают очень малую часть текста. Совстречаемостей этих слов (то есть позиций слов в документах, когда рядом с ними находятся другие топ-слова) ещё меньше.

Напротив, если предположить существование суперсимметрии, то введение новых **частиц** приводит как раз к такому объединению. Оказывается, что суперсимметрия не только обеспечивает объединение взаимодействий, но и стабилизирует объединённую теорию, в которой присутствуют два совершенно разных масштаба: масштаб масс обычных **частиц** (порядка 100 масс протона) и масштаб великого объединения (порядка 10^{16} масс протона). Последний масштаб уже близок к так называемому планковскому масштабу, равному обратной ньютоновской константе тяготения, что составляет порядка 10^{19} масс протона. На этом масштабе мы ожидаем проявление эффектов квантовой гравитации. В этом моменте нас ожидает приятный сюрприз. Дело в том, что гравитация всегда стояла несколько особняком по отношению к остальным взаимодействиям. Переносчик гравитации, гравитон, имеет спин 2, в то время как переносчики остальных взаимодействий имеют спин 1. Однако суперсимметрия перемешивает спины.

Первые топ-слова темы *«физика»*, включая первые **топ-10-слов**, с вероятностями (%):

частица (2.7), **электрон** (1.5), **кварк** (1.5), **атом** (1.3), **энергия** (1.2), **вселенная** (1.1), **фотон** (1.0), **физика** (0.9), **физик** (0.9), **эксперимент** (0.9), масса (0.7), теория (0.7), свет (0.7), симметрия (0.7), протон (0.7), эйнштейн (0.5), нейтрино (0.5), вещество (0.5), квантовый (0.5), ускоритель (0.5), детектор (0.4), волна (0.4), эффект (0.4), свойство (0.4), спин (0.4), гравитация (0.4), материя (0.4), адрон (0.4), поль (0.4), частота (0.4)

Рис. 4: В приведённом тексте встречается единственный топ-токен «частиц» из первых десяти топовых слов темы «физика». При этом широкий спектр менее тематичных слов будет проигнорирован когерентностями по топ-словам. Более того, само слово «частиц» не будет учтено, потому что оно не участвует в совстречаемостях топ-слов: рядом с ним нет других топовых слов. Таким образом, когерентности по топ-словам не увидят в приведённом отрывке темы «физика».

3. Внутритекстовая когерентность

3.1. На пути к внутритекстовому подходу

Идея, на которой основаны подходы к подсчёту когерентности темы t (как основанные на топ-словах, так и внутритекстовые, которые будут представлены далее) — выяснить, как часто слова, которые модель относит к теме t, встречаются в тексте вблизи друг от друга. Тогда когерентность темы будет высокой, если её слова образуют кластеры внутри текстов коллекции, а не разбросаны по тексту случайно. Это похоже на лингвистическое понятие *когезии* (связности) текста [14]: предложения связаны между собой синтаксически и лексически с помощью повторов слов, синонимов, гипонимов и других языковых средств, единиц и форм.

Качество темы (7) надо оценивать не по совстречаемостям топ-слов темы, но с учётом того, как *вся тема* распределена в тексте, образуют ли её слова однородные фрагменты. Таким образом, чтобы посчитать *внутритекстовую когерентность* для темы t, можно последовательно перебирать слова текста и сравнивать вероятности этой темы t на близких словах. В зависимости от способа сравнения слов и самого подсчёта когерентности можно предлагать разные функции внутритекстовой когерентности.

3.2. Предлагаемые функции внутритекстовой когерентности

В данной работе представляется несколько вариантов подсчёта внутритекстовой когерентности $\operatorname{coh}\big(D, W, \Phi, \Theta\big)$.

Далее в формулах знаком $\langle\cdot\rangle$ обозначено усреднение, знаком $[\cdot]$ — единица в случае истинности условия в скобках и ноль иначе.

Первый метод — SemantiC (Semantic Closeness) (рисунок 5) — считает тематическую близость слов, расположенных недалеко друг от друга в внутри текста, как векторов с компонентами $p(t \mid d, w)$ (в [13] предлагается способ оценки когерентности по близости топ-слов темы, где каждому слову также ставится в соответствие вектор в некотором пространстве, но компоненты векторов представляют не вероятности тем, а зависят от совстречаемостей слов словаря в некотором корпусе). Для оценки схожести векторов, соответствующих словам, в одной разновидности SemantiC используется минус l_2 норма вектора разности (6), в другой — косинус угла между векторами (7)

$$\operatorname{SemantiC}_{l_2}\Big|_t = \Big\langle \big[0 < \rho(w_i, w_j) < \text{window}\big] \; -\big\|\boldsymbol{w_i} - \boldsymbol{w_j}\big\|_2 \Big\rangle_{\substack{w_i, w_j \in \bigcup_d W_d \\ \arg\max_s \boldsymbol{w}_i(s) = t \\ \arg\max_s \boldsymbol{w}_j(s) = t}} \tag{6}$$

$$\operatorname{SemantiC}_{\text{Cos}}\Big|_t = \Big\langle \big[0 < \rho(w_i, w_j) < \text{window}\big] \; \cos\left(\boldsymbol{w_i}, \boldsymbol{w_j}\right) \Big\rangle_{\substack{w_i, w_j \in \bigcup_d W_d \\ \arg\max_s \boldsymbol{w}_i(s) = t \\ \arg\max_s \boldsymbol{w}_j(s) = t}} \tag{7}$$

где $\rho(w_i, w_j) = |j - i|$ — расстояние по тексту между словами (количество слов между ними), window — число слов в окне, внутри которого слова считаются близкими по тексту.

При этом каждому слову ставится в соответствие вектор по правилу

$$W_d \ni w \mapsto \boldsymbol{w} \equiv \big(p(t \mid d, w)\big)_{t \in T} \tag{8}$$

Заметка. Когерентности SemantiC_{l_2}, $\text{SemantiC}_{\text{Cos}}$ тем выше, чем более похожи векторы слов, которые участвуют в подсчёте.

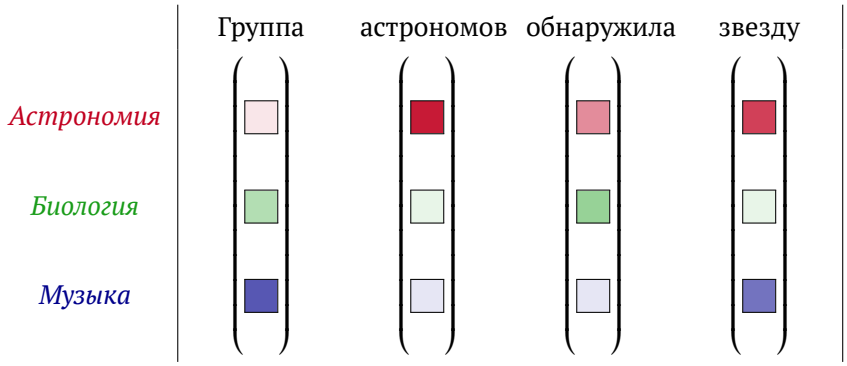

Рис. 5: Когерентности SemantiC_{l_2}, $\text{SemantiC}_{\text{Cos}}$ считают близость близко расположенных в тексте слов, сравнивая пары векторов слов из одного текстового окна и относящихся к одной теме. В приведённом примере для темы «Астрономия» сравнивались бы векторы слов «астрономов» и «звезду».

Третья разновидность SemantiC — $\text{SemantiC}_{\text{Var}}$ для темы t считает дисперсию между соответствующими теме t компонентами векторов слов, расположенных в одном текстовом окне (формула (9) и рисунок 6), при этом каждому слову ставится в соответствие вектор по правилу (8).

$$\text{SemantiC}_{\text{Var}}\Big|_t = \Big\langle -\text{Var}\Big(\boldsymbol{w}_i(t), \boldsymbol{w}_{i+1}(t), \dots, \boldsymbol{w}_{i+\text{window}}(t)\Big)\Big\rangle_{w_i \in \bigcup_d W_d} \tag{9}$$

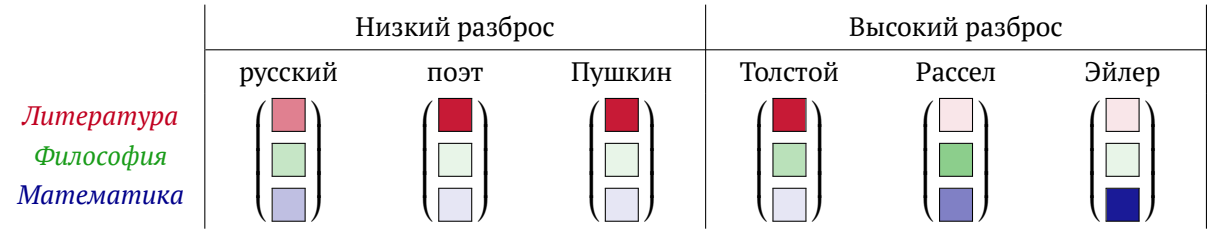

Рис. 6: $\text{SemantiC}_{\text{Var}}$ оценивает, как сильно разбросана тема t по словам из одного текстового окна.

Заметка. В данной работе перед вычислением когерентностей по способу SemantiC векторы слов умножались на 1000, чтобы увеличить по модулю итоговое значение когерентности.

Другая функция когерентности — TopLen (Topic Length) (формулы (10) и (11) и рисунок 7) — считает среднюю длину темы внутри текста, на каждом слове вычисляя разницу между компонентой вектора, соответствующей теме t и максимальной компонентой их оставшихся. Неотрицательный параметр threshold сглаживает эффект, когда встречаются слова не темы t: подсчёт продолжается, пока сумма порога threshold и разностей компонент по просмотренным словам неотрицательна. И опять каждому слову ставится в соответствие вектор по правилу (8).

$$\text{TopLen}\Big|_t = \Big\langle \overbrace{l(0)}^{l_1}, \overbrace{l(l_1)}^{l_2}, l(l_1 + l_2), \dots, l\left(\sum_{r=0}^{k-1} l_r\right), \dots \Big\rangle_{l_j > 0} \tag{10}$$

При этом

$$l(\cdot):\ \begin{cases} l(i) = \max\left\{ L \,:\, \text{threshold} + \sum\limits_{j=i}^{i+L} \left(\boldsymbol{w}_j(t) - \max\limits_{\substack{1\le\tau\le|T|\\ \tau\neq t}} \boldsymbol{w}_j(\tau) \right) \ge 0 \right\} \\ \arg\max_s \boldsymbol{w}_i(s) = t \end{cases} \qquad (11)$$
$$\begin{cases} l(i) = 0 \\ \arg\max_s \boldsymbol{w}_i(s) \neq t \end{cases}$$

Группе $\underbrace{\text{астрономов удалось}}_{l_1=2}$ обнаружить $\underbrace{\text{звезду, обращающуюся}}_{l_2=2}$
вокруг $\underbrace{\text{чёрной дыры на рекордно близком}}_{l_3=4}$ расстоянии.

Рис. 7: Метод TopLen считает среднюю длину темы внутри текста. Пример возможной работы метода для темы t = «Чёрные дыры». Ведётся подсчёт слов. На каждом слове темы t метод получает положительное поощрение. Если встречается слово w не темы t, то на нём метод штрафуется. Поощрение/штраф тем больше по модулю, чем больше/меньше слово w относится к теме t. Когда сумма штрафов по модулю становится больше, чем сумма поощрений, на величину threshold, процесс подсчёта слов останавливается. Количество посчитанных слов есть одно из зарегистрированных значений длины темы t, которые в конце усредняются, давая итоговое значение когерентности TopLen.

Последний способ — FoCon (Focus Consistency) (формула (12) и рисунок 8) — смотрит, как сильно изменяется тема t среди смежных слов, суммируя разности компонент $p(t \mid d, w)$ (пара компонент в формуле (12), по которым считаются разности, соответствуют максимумам по компонентам в смежных векторах). Знак минус играет ту роль, что когерентность возрастает, когда близкие слова меньше отличаются друг от друга по темам. Соответствие между словами и векторами — по правилу (8).

$$\text{FoCon} = -\sum_{d\in D} \sum_{\substack{w_i, w_j \in W_d \\ j-i=1}} \left|\boldsymbol{w}_i(t) - \boldsymbol{w}_j(t)\right| + \left|\boldsymbol{w}_i(\tau) - \boldsymbol{w}_j(\tau)\right| \qquad (12)$$
$$\begin{cases} t = \arg\max_s \boldsymbol{w}_i(s) \\ \tau = \arg\max_s \boldsymbol{w}_j(s) \end{cases}$$

Заметка. Метод FoCon не привязан к конкретной теме, он выдаёт результат сразу для тематической модели как целого (*когерентность тематической модели*).

4. Качество функций когерентности

4.1. Интерпретация

Автоматические методы посчёта когерентности опираются на совстречаемости слов. Если слова из некоторого множества слов часто встречаются вместе в одном окне внутри текста, то это мно-

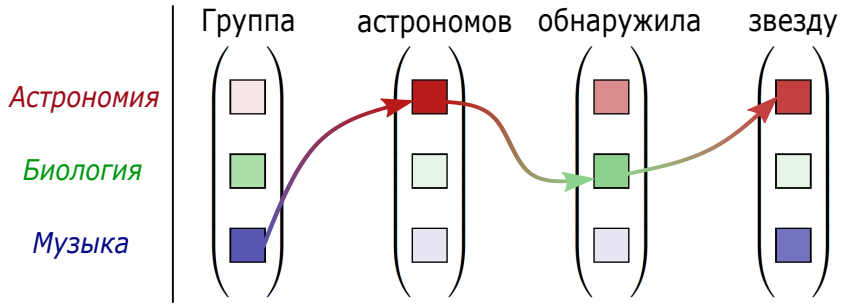

Рис. 8: FoCon оценивает, как сильно меняются темы смежных слов.

жество слов считается когерентным, то есть эти слова в некотором смысле подходят друг другу. При этом неявно подразумевается, что если множество топовых слов когерентно, то когерента и вся тема, к которой они относятся. Такое предположение уже подвергалось критике ранее [12]. В данной же работе, помимо прочего, осуществляется количественная оценка доли текста рассматриваемой коллекции, которая покрывается совстречаемостями топовых слов.

Приведём ещё одно следствие из гипотезы о сегментной структуре текста (6), которое приведёт нас к способу оценки качества функций когерентности.

***Следствие* 6.3.** *Чем лучше функция когерентности, тем лучше она должна описывать способность тематической модели угадывать сегментную структуру текста.*

Но позиции сегментов тем в текстах естественного языка заранее не известны. Нежелание связываться с долгой, не выполнимой за разумные сроки, экспертной разметкой некоторого корпуса на тематические сегменты, привело к идее о *полусинтетическом датасете*: имея в распоряжении монотематические документы, можно самому разрезать их на сегменты, которые потом сшить в новые документы. И получится датасет с известной тематической разметкой, который можно будет потом использовать для оценки качества функций когерентности. Высказанные идеи обсуждаются подробнее в следующих разделах (4.2) и (4.3).

4.2. Исходный датасет

В распоряжении имеется около 2000 монотематических статей из научно-популярного контента «ПостНаука»[1]. Всего в статьях затрагивается 19 тем (таблица 9 и рисунок 2).

4.3. Полусинтетический датасет

Как уже отмечалось, оценка интерпретируемости тем трудозатратна. Преимущество когерентностей по топ-словам в том, что в них каждая тема представляется небольшим списком слов. Но и в этом случае сбор экспертных оценок интрепретируемости представляет сложности.

Внутритекстовые когерентности — попытка придумать функцию, учитывающую целиком матрицы слов в темах и тем в документах (Φ и Θ) и весь текстовый корпус. Встаёт вопрос о том,

[1] https://postnauka.ru

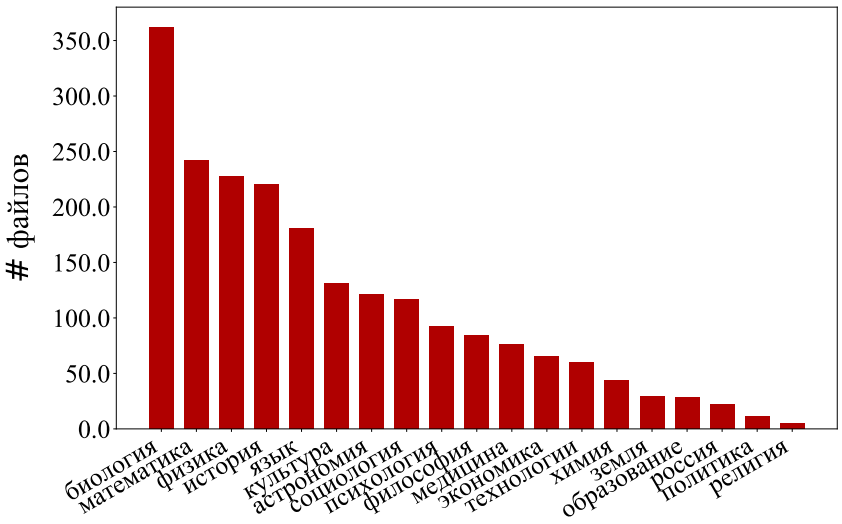

Рис. 9: Распределение документов датасета «ПостНауки» по темам.

как оценивать качество предложенных методов. Кажется, что для этого нужно каким-то образом получить экспертные мнения о вероятностных распределениях слов в темах и тем в документах, что представляется неосуществимым в разумных временных рамках.

Ранее уже была выдвинута гипотеза о том, что все тексты естественного языка сегментированы (6). Чем лучше тематическая модель, построенная по некоторой коллекции, тем лучше она должна предсказывать эту сегментную стуктуру. Таким образом, качество модели можно оценивать по *качеству сегментации* ею некоторого текста. Когерентность тоже призвана оценивать качество тем и порождающих их тематических моделей. Поэтому если знать разметку на сегменты у статей корпуса, то можно, посчитав как-нибудь качество сегментации текста моделью и считая это «золотым стандартом», оценить качество функции когерентности.

Озвученные выше мысли приводят к следующему решению. По датасету статей «ПостНауки» создаётся новый *полусинтетический* датасет. Исходные 2000 *монотематических* статей «ПостНауки» разрезаются на сегменты одинаковой длины, которые потом сшиваются в новые документы D' (рисунок 10). Для чего?

Дело в том, что статьи «ПостНауки» можно с большой уверенность считать *монотематическими*: большинство слов в статье относятся к одной предметной теме — сама коллекция «ПостНауки» устроена таким образом, что каждая статья в ней посвящена какой-то одной теме. К тому же, 19 тем модели в основном слабо связаны друг с другом. Таким образом, нарезав и сшив статьи, получаем датасет с *известной разметкой* на сегменты (рисунок 11).

Осталось разобраться со способом оценки качества сегментации такого сшитого из разных сегментов текста тематической моделью.

4.4. Качество сегментации

Итак, при создании нового датасета из сегментов статей исходного происходит преобразование

$$W_d \ni w \mapsto w' \in W'_{d'},$$

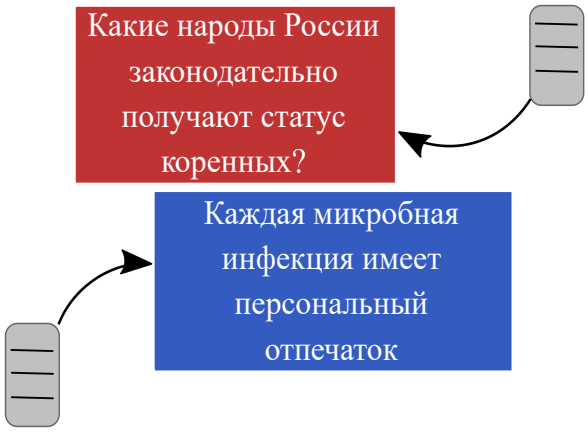

Рис. 10: Документ, состоящий из двух сегментов: один по теме «социология», другой — «медицина». Сегменты есть фрагменты из статей по соответствующим темам.

Рис. 11: Сегменты заданного размера объединяются в документы. Каждый сегмент относится к некоторой теме. Фиксируются размер документа и число тем, представленных в документе сегментами.

Таблица 2: Темы статей «ПостНауки». Каждая представлена своими тремя топ-словами.

Тема	Топ-слово #1, %	Топ-слово #2, %	Топ-слово #3, %
1: математика	математика (1.6)	задача (0.8)	декарт (0.8)
2: технологии	технология (1.5)	робот (1.2)	сеть (1.0)
3: физика	частица (2.7)	электрон (1.5)	кварк (1.5)
4: химия	химия (2.1)	молекула (1.9)	материал (1.6)
5: земля	земля (2.9)	планета (2.8)	атмосфера (1.2)
6: астрономия	звезда (3.9)	галактика (3.1)	вселенная (1.9)
7: биология	клетка (2.7)	организм (1.1)	мозг (1.0)
8: медицина	пациент (1.6)	препарат (1.2)	заболевание (1.2)
9: психология	психология (0.9)	мозг (0.9)	психолог (0.8)
10: экономика	экономика (1.6)	страна (1.0)	цена (0.8)
11: история	история (1.0)	историк (0.7)	власть (0.6)
12: политика	государство (1.4)	политика (1.2)	политический (1.1)
13: социология	социология (1.3)	социолог (0.9)	социальный (0.8)
14: культура	культура (1.5)	фильм (0.7)	искусство (0.6)
15: образование	университет (2.1)	образование (1.4)	школа (1.3)
16: язык	язык (7.7)	слово (3.7)	словарь (1.1)
17: философия	философия (1.8)	философ (1.3)	философский (0.8)
18: религия	святилище (1.0)	религия (0.7)	царь (0.6)
19: россия	россия (2.8)	страна (0.9)	русский (0.9)

При этом, каждому слову, как из W_d, так и из $W'_{d'}$, соответствует вектор по правилу (8):

$$\begin{cases} W_d \ni w \mapsto \boldsymbol{w} \equiv \big(p(t \mid d, w)\big)_{t \in T} \\ W'_{d'} \ni w' \mapsto \boldsymbol{w}' \equiv \big(p(t' \mid d', w')\big)_{t' \in T'} \end{cases}$$

Между темами исходного датасета T и полусинтетического T' есть взаимно однозначное соответствие

$$T \ni t \leftrightarrow t' \in T'$$

которое устанавливается по качеству сегментации (13) (о котором будет далее) с помощью венгерского алгоритма [1]: находится оптимальное соответствие по матрице размера $|T| \times |T'|$, $|T'| = |T|$, где элемент (i, j) — качество сегментации новой темы $t'_j \in T'$ при условии, что она есть образ исходной темы $t_i \in T$.

Предлагается два способа оценки качества сегментации (segmentation quality — sq).

Первый — Soft — для каждой темы t считает сумму вероятностей $p(t \mid d, w)$ по словам *сегментов темы* t, а потом суммирует все результаты по темам (формула (13) и рисунки 12 и 14).

$$sq(D', W, \Phi', \Theta')_{\text{Soft}} \Big|_{t'} = \sum_{d' \in D'} \sum_{\substack{w' \in W'_{d'} \\ \arg\max_s \boldsymbol{w}(s) = t}} p(t' \mid d', w') \tag{13}$$

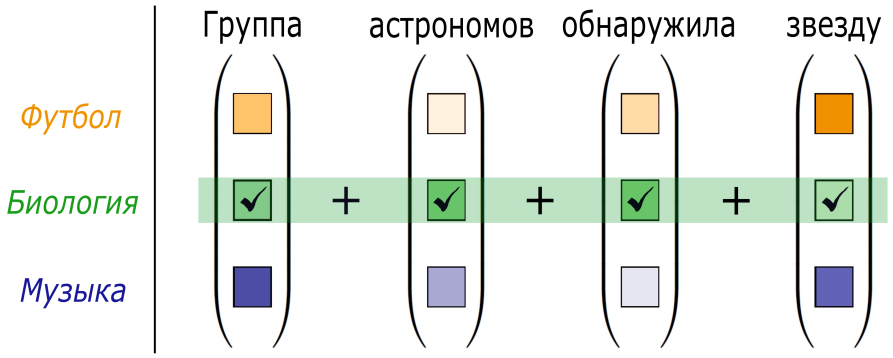

Рис. 12: Иллюстрация подсчёта качества сегментации по методу Soft (13).

Второй — Hard — для каждой темы t считает число правильных угадываний этой темы моделью, то есть для скольких слов *из сегментов темы* t предсказанная моделью для этого слова тема $\arg\max_\tau p(\tau \mid d, w)$ совпала с темой t (формула (14) и рисунок 13).

$$sq(D', W, \Phi', \Theta')_{\text{Hard}}\Big|_{t'} = \sum_{d' \in D'} \sum_{\substack{w' \in W'_{d'} \\ \arg\max_s \boldsymbol{w}(s) = t}} \left[\arg\max_s \boldsymbol{w}'(s) = t\right] \tag{14}$$

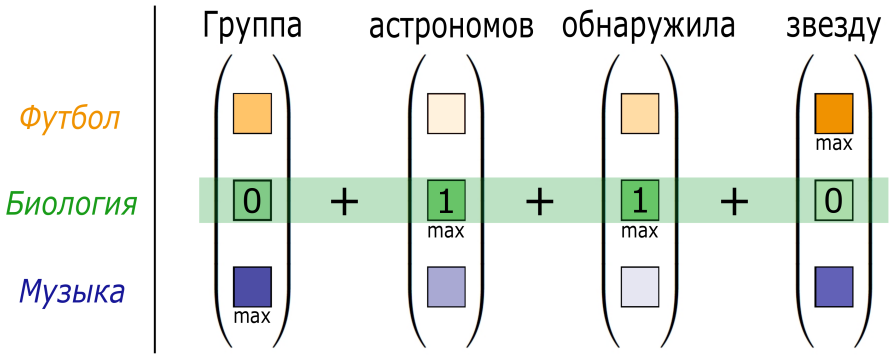

Рис. 13: Иллюстрация подсчёта качества сегментации по методу Hard (14).

Оценка качества функций когерентности с помощью сравнения с качеством сегментации опирается на

Гипотеза 7. Функция $sq(\cdot)$ задаёт тот же порядок на образах тем T', что и неизвестная функция качества тем $q(\cdot)$ (см. определение 7) на исходных темах T

$$sq(t'_1) < sq(t'_2) \leftrightarrow q(t_1) < q(t_2)$$

5. Эксперименты

Считая предложенные в предыдущей секции способы оценки качества сегментации «золотым стандартом», можно провести оценку качества функций когерентности по их соответсвию этому стандарту. Так, далее в работе считаются Спирмановские корреляции между когерентностями и

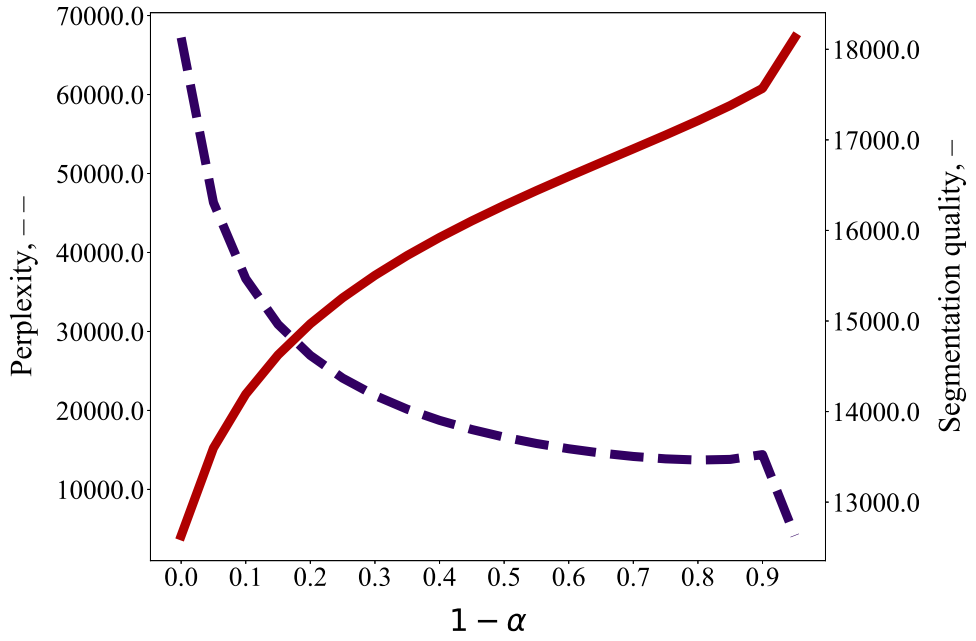

Рис. 14: Зависимость между Soft качеством сегментации и перплексией модели от качества модели. Эксперимент проведён на датасете с размером сегмента 200 слов и с 5 темами в каждом документе. По оси X отложена доля хорошей матрицы Φ_{good}: один минус α (степень деградации Ф). Факт того, что качество сегментации монотонно возрастает с уменьшением перплексии не противоречит тому, что Soft качество сегментации может использоваться для оценки качества тематической модели.

качествами сегментации для ряда тематических моделей. Таким образом, чем ближе к единице корреляция, тем выше качество когерентности.

Чтобы получить ряд тематических моделей, на которых будут считаться качества сегментации и когерентности для последующего получения Спирмановских коэффициентов корреляции, применяется следующая процедура. Для каждой модели матрица слов в темах Ф задаётся равной взвешенной сумме «хорошей» матрицы Φ_{good} (матрицы модели «ПостНауки», с оригинальными, ненарезанными статьями) и «плохой» матрицы Φ_{bad}, столбцы которой получены из распределения $\text{Dirichlet}\left(\mathbf{0.01}_{|W|}\right)$.

$$\Phi(\alpha) = \alpha \cdot \Phi_{bad} + (1 - \alpha) \cdot \Phi_{good} \tag{15}$$

Таким образом, α — доля «плохой» матрицы.

На рисунке 15 для «плохой» и «хорошей» моделей представлены примеры того, как сегментируются тексты (то есть как модель угадывает темы слов в сегментах).

Для каждого α, который брался из полуинтервала $[0, 1)$ с некоторым шагом, по формуле (15) строилась тематическая модель. Для неё считались качества сегментации и когерентности. Далее между рядами значений когерентностей и качества сегментации считалась Спирменовская корреляция (16).

$$\text{Corr}\left\{\left(\text{coh}\left(\Phi'(\alpha)\right)\right)_{\alpha}, \left(sq\left(\Phi'(\alpha)\right)\right)_{\alpha}\right\} \tag{16}$$

На четырёх полусинтетических датасетах, с размерами сегментов 50, 100, 200 и 400 слов, бы-

ло проведено по четыре серии таких измерений (с разными «плохими» матрицами Φ_{bad} из одного распределения Дирихле). Результаты экспериментов представлены в таблице 3 и на рисунке 16. Для когерентностей по топ-словам количество топ-слов равно 10. Размер окна 20 слов. Порог threshold метода TopLen взят равным 0.01. При подсчёте корреляций бралось только Soft качество сегментации: оказалось, что между Soft и Hard принципиальной разницы нет, в том смысле, что при улучшении тематической модели от Φ_{bad} до Φ_{good} по формуле (15) показатели качества сегментации монотонно возрастали и для Soft, и для Hard. Разница была только в величине качества сегментации (величина Hard в 2–3 раза больше соответствующей величины Soft). Отметим, что метод $\text{SemantiC}_{\text{Cos}}$ показывает низкие, близкие к -1 корреляции. Возможно, это происходит из-за того, что в случайной матрице Φ_{bad} со столбцами из распределения Дирихле для каждой темы вероятностное распределение сконцентрировано лишь в нескольких словах. Но по мере изменения модели к Φ_{good} темы как вероятностные распределения захватывают большее количество слов, и средний угол между соответствующими векторами $\big(p(t \mid w)\big)_{t \in T}$ будет больше, чем при матрице Φ_{bad}.

Таблица 3: Корреляции Спирмена между когерентностями и Soft (13) качествами сегментации для трёх полусинтетических датасетов: с 50, 200 и 400 словами в каждом сегменте, и при 5 темах в каждом сшитом из сегментов документе в случае всех трёх датасетов.

Coh	Corr	Coh	Corr	Coh	Corr
Newman	0.75	Newman	0.80	Newman	0.85
Mimno	0.96	Mimno	0.94	Mimno	0.97
SemantiC_{l_2}	0.92	SemantiC_{l_2}	0.70	SemantiC_{l_2}	0.59
$\text{SemantiC}_{\text{Cos}}$	−0.97	$\text{SemantiC}_{\text{Cos}}$	−0.97	$\text{SemantiC}_{\text{Cos}}$	−0.96
$\text{SemantiC}_{\text{Var}}$	**1.00**	$\text{SemantiC}_{\text{Var}}$	**1.00**	$\text{SemantiC}_{\text{Var}}$	**1.00**
TopLen	**1.00**	TopLen	**1.00**	TopLen	**1.00**
FoCon	**1.00**	FoCon	**1.00**	FoCon	**1.00**

6. Результаты

- Проиллюстрирован недостаток когерентностей по топ-словам: покрытие лишь малой части текстовой коллекции.

- Предложен полуавтоматический метод оценки качества функций когерентности: по корреляции с качествами сегментации полусинтетического текста тематическими моделями с разным качеством (от «хорошей» модели исходной коллекции до случайной «плохой» со столбцами матрицы слов в темах Φ из некоторого вероятностного распределения). Полусинтетический датасет получается сшиванием сегментов, на которые разбиваются исходные монотематические статьи. Качество сегментации текста оценивается с учётом известной разметки на сегменты.

#top-tokens	10	20	50	100
Newman	0.7 ± 0.5	0.9 ± 0.2	0.9 ± 0.2	0.98 ± 0.03
Mimno	0.9 ± 0.3	0.9 ± 0.1	0.99 ± 0.02	0.98 ± 0.03

Таблица 4: Зависимости корреляций между топ-токен когерентностями Newman и Mimno и качеством сегментации Soft от числа топ-токенов (по совстречаемостям которых считаются когерентности). Результаты приведены для датасета с размером сегмента 100 слов и с 5 темами в каждом документе. Было проведено 5 серий экспериментов (с различными матрицами Φ_{bad}), потом значения корреляций для каждого значения окна усреднялись, а в качестве погрешности бралась полуразность между максимальным и минимальным значениями корреляций для каждого значения окна. Видно, что с увеличением числа топ-токенов от 10 до 100 показатели корреляций возрастают, но единицы не достигают. Также спадает погрешность корреляции. Однако, при увеличении числа топ-слов встаёт другая проблема: становится не понятно, что в таком случае считают когерентности по топ-словам. Когда топ-слов было десять, это были самые ярко выраженные слова темы. Они отражали не всю тему, но большу́ю её часть. Если же взять первые сто топ-слов, то среди них могут быть такие слова, которые относятся к другим темам. В таком случае будут учитываться совстречаемости между словами разных тем, что не имеет смысла при подсчёте когерентности. Можно бы было при подсчёте когерентности *для каждой темы определять число топ-слов*, которые относятся именно к ней, и смотреть уже только их совстречаемости. Но даже в этом случае покрытие коллекции топ-словами будет не полным, к тому же ожидается, что частоты топ-слов будут спадать обратно пропорционально номеру топ-слова [21]. В методе SemantiC_{l_2} тоже определяются темы *для каждого слова* в коллекции, но SemantiC_{l_2} рассматривает слова из W_d, $d \in D$, а не слова из W (учитываются вероятности $p(t \mid d, w)$, а не $p(t \mid w)$).

- Предложены новые методы *внутритекстовой* когерентности: SemantiC_{l_2}, $\text{SemantiC}_{\text{Cos}}$, $\text{SemantiC}_{\text{Var}}$, TopLen, FoCon. При подсчёте они учитывают весь текст коллекции и матрицы слов в темах Φ и тем в документах Θ. По предложенной функции оценки качества (по корреляции с качеством сегментации) новые методы SemantiC_{l_2}, $\text{SemantiC}_{\text{Var}}$, TopLen и FoCon показывают лучшие результаты, чем когерентности по топ-словам Newman и Mimno.

Результаты работы были представлены на 24-ой международной конференции по компьютерной лингвистике и интеллектуальным технологиям «Диалог» (июнь 2018) и на 60-й Научной конференции МФТИ (ноябрь 2017).

7. Благодарности

Выражаю признательность Виктору Булатову и Ирине Ефимовой за помощь в сборе монотематических статей «ПостНауки».

Эксперименты с данными были проведены с помощью библиотеки BigARTM [20, 23].

window	5	10	20	30
SemantiC$_{l_2}$	0.6 ± 0.7	0.6 ± 0.7	0.6 ± 0.6	0.6 ± 0.6
SemantiC$_{\text{Cos}}$	-1 ± 0.005	-1 ± 0.005	-1 ± 0.005	-1
SemantiC$_{\text{Var}}$	0.98 ± 0.03	0.98 ± 0.03	0.98 ± 0.03	0.98 ± 0.03
threshold	0.01	0.1	0.2	0.5
TopLen	1	1	1	1

Таблица 5: Зависимости корреляций между внутритекстовыми когерентностями SemantiC$_{l_2}$, SemantiC$_{\text{Cos}}$, SemantiC$_{\text{Var}}$, и TopLen и качеством сегментации Soft в зависимости от: window — для SemantiC и threshold — для TopLen. Результаты приведены для датасета с размером сегмента 100 слов и с 5 темами в каждом документе. Было проведено 3 серии экспериментов (с различными матрицами Φ_{bad}), потом значения корреляций для каждого значения параметра усреднялись, а в качестве погрешности бралась полуразность между максимальным и минимальным значениями корреляций для каждого значения параметра. Видно, что корреляции не изменялись с изменением window/threshold. При этом у SemantiC$_{l_2}$ наблюдается большая погрешность, сравнимая со значением корреляции — трёх серий экспериментов для усреднения в этом случае оказалось недостаточно. SemantiC$_{\text{Var}}$ и TopLen показывают близкие к единице корреляции, с малыми погрешностями (как в случае 100 топ-токенов для когерентностей Newman и Mimno в таблице 4).

8. Литература

[1] Harold W Kuhn. «The Hungarian method for the assignment problem». В: *Naval Research Logistics (NRL)* 2.1-2 (1955), с. 83—97.

[2] T. Hoffman. «Probabilistic latent semantic indexing». В: *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. New York: ACM Press, 1999, с. 50—57.

[3] David M Blei, Andrew Y Ng и Michael I Jordan. «Latent dirichlet allocation». В: *Journal of machine Learning research* 3.Jan (2003), с. 993—1022.

[4] David M Blei и John D Lafferty. «Topic models». В: *Text mining: classification, clustering, and applications* 10.71 (2009), с. 34.

[5] Jonathan Chang и др. «Reading Tea Leaves: How Humans Interpret Topic Models». В: *Advances in Neural Information Processing Systems 22: 23rd Annual Conference on Neural Information Processing Systems 2009. Proceedings of a meeting held 7-10 December 2009, Vancouver, British Columbia, Canada*. Под ред. Yoshua Bengio и др. Curran Associates, Inc, 2009, с. 288—296. isbn: 9781615679119.

[6] David Newman и др. «Automatic Evaluation of Topic Coherence». В: *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational*

Linguistics. HLT '10. Los Angeles, California: Association for Computational Linguistics, 2010, с. 100—108. isbn: 1-932432-65-5.

[7] Sang Su Lee, Tagyoung Chung и Dennis McLeod. «Dynamic Item Recommendation by Topic Modeling for Social Networks». В: *Eighth International Conference on Information Technology: New Generations, ITNG 2011, Las Vegas, Nevada, USA, 11-13 April 2011*. Под ред. Shahram Latifi. IEEE Computer Society, 2011, с. 884—889. isbn: 978-0-7695-4367-3.

[8] David Mimno и др. «Optimizing Semantic Coherence in Topic Models». В: *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. EMNLP '11. Edinburgh, United Kingdom: Association for Computational Linguistics, 2011, с. 262—272. isbn: 978-1-937284-11-4.

[9] Chong Wang и David M. Blei. «Collaborative topic modeling for recommending scientific articles». В: *Proceedings of the 17th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, San Diego, CA, USA, August 21-24, 2011*. Под ред. Chid Apté, Joydeep Ghosh и Padhraic Smyth. ACM, 2011, с. 448—456. isbn: 978-1-4503-0813-7.

[10] David M. Blei. «Probabilistic topic models». В: *Commun. ACM* 55.4 (2012), с. 77—84.

[11] Timothy N. Rubin и др. «Statistical topic models for multi-label document classification». В: *Machine Learning* 88.1-2 (2012), с. 157—208.

[12] Benjamin M Schmidt. «Words alone: Dismantling topic models in the humanities». В: *Journal of Digital Humanities* 2.1 (2012), с. 49—65.

[13] Nikolaos Aletras и Mark Stevenson. «Evaluating topic coherence using distributional semantics». В: *Proceedings of the 10th International Conference on Computational Semantics (IWCS 2013)–Long Papers*. 2013, с. 13—22.

[14] Michael Alexander Kirkwood Halliday и Ruqaiya Hasan. *Cohesion in english*. Routledge, 2014.

[15] Jey Han Lau, David Newman и Timothy Baldwin. «Machine Reading Tea Leaves: Automatically Evaluating Topic Coherence and Topic Model Quality.» В: *EACL*. 2014, с. 530—539.

[16] Julio Cesar Louzada Pinto и Tijani Chahed. «Modeling Multi-topic Information Diffusion in Social Networks Using Latent Dirichlet Allocation and Hawkes Processes». В: *Tenth International Conference on Signal-Image Technology and Internet-Based Systems, SITIS 2014, Marrakech, Morocco, November 23-27, 2014*. Под ред. Kokou Yétongnon, Albert Dipanda и Richard Chbeir. IEEE Computer Society, 2014, с. 339—346. isbn: 978-1-4799-7978-3.

[17] Devesh Varshney, Sandeep Kumar и Vineet Gupta. «Modeling information diffusion in social networks using latent topic information». В: *International Conference on Intelligent Computing*. Springer. 2014, с. 137—148.

[18] Konstantin Vorontsov и Anna Potapenko. «Tutorial on Probabilistic Topic Modeling: Additive Regularization for Stochastic Matrix Factorization». В: *Analysis of Images, Social Networks and Texts - Third International Conference, AIST 2014, Yekaterinburg, Russia, April 10-12, 2014, Revised Selected Papers*. Под ред. Dmitry I. Ignatov и др. Т. 436. Communications in Computer and Information Science. Springer, 2014, с. 29—46. isbn: 978-3-319-12579-4.

[19] Konstantin Vorontsov и Anna Potapenko. «Additive regularization of topic models». В: *Machine Learning* 101.1-3 (2015), с. 303—323.

[20] Konstantin Vorontsov и др. «BigARTM: Open Source Library for Regularized Multimodal Topic Modeling of Large Collections». В: *Analysis of Images, Social Networks and Texts - 4th International Conference, AIST 2015, Yekaterinburg, Russia, April 9-11, 2015, Revised Selected Papers*. Под ред. Mikhail Yu. Khachay и др. Т. 542. Communications in Computer and Information Science. Springer, 2015, с. 370—381. isbn: 978-3-319-26122-5.

[21] George Kingsley Zipf. *Human behavior and the principle of least effort: An introduction to human ecology*. Ravenio Books, 2016.

[22] Anastasia Ianina, Lev Golitsyn и Konstantin Vorontsov. «Multi-objective topic modeling for exploratory search in tech news». В: *Conference on Artificial Intelligence and Natural Language*. Springer. 2017, с. 181—193.

[23] Denis Kochedykov и др. «Fast and modular regularized topic modelling». В: *Open Innovations Association (FRUCT), 2017 21st Conference of*. IEEE. 2017, с. 182—193.

[24] Anna Potapenko, Artem Popov и Konstantin Vorontsov. «Interpretable probabilistic embeddings: bridging the gap between topic models and neural networks». В: *Conference on Artificial Intelligence and Natural Language*. Springer. 2017, с. 167—180.

Плохая тематическая модель (Φ_{bad})

тема 16 : язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить ”я это сделаю” или ”это будет”, а сказать ”это возможно” или ”я хочу это сделать”. Они говорят о будущем, но ”попадают” в будущее непрямым путём.

тема 12 : политика

И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных гражданств. В литературе последних десяти лет бытуют такие выражения, как гендерное гражданство и экономическое гражданство. Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC l_2	SC Cos	SC Var	TL	FC
5500	11000	-4.8	-3.1	-13	0.95	-37000	2.9	-140000
16000	**38000**	**-3.7**	**-2.7**	**-3.7**	0.70	**-8100**	**3.5**	**-54000**

Хорошая тематическая модель (Φ_{good})

тема 16 : язык

Категория будущего времени в большинстве языков Африки отсутствует. Есть много способов говорить о будущем, но это более сложные способы, касающиеся предположения, желания. Нормальный африканский грамматический приём — не говорить ”я это сделаю” или ”это будет”, а сказать ”это возможно” или ”я хочу это сделать”. Они говорят о будущем, но ”попадают” в будущее непрямым путём.

тема 12 : политика

И я посылаю деньги борцам за независимость Курдистана, участвую в акциях поддержки курдских повстанцев и так далее. Вот такое наложение друг на друга разных членств, разных гражданств. В литературе последних десяти лет бытуют такие выражения, как гендерное гражданство и экономическое гражданство. Первое указывает на членство в воображаемом сообществе женщин, приверженных идеям феминизма.

SQ (S)	SQ (H)	N	M	SC l_2	SC Cos	SC Var	TL	FC
5500	11000	-4.8	-3.1	-13	0.95	-37000	2.9	-140000
16000	**38000**	**-3.7**	**-2.7**	**-3.7**	0.70	**-8100**	**3.5**	**-54000**

Рис. 15: На картинке изображены два сегмента размера по 50 слов в каждом, по разным темам. Показано, как «плохая» и «хорошая» модели справляются с сегментацией такого текста. Сами сегменты были извлечены из одного из полусинтетических документов, в котором эти сегменты были смежными. Слова, которые никак не выделены, были отнесены моделями к темам, отличным от двух тем сегментов. Под сегментами показаны значения качеств сегментации и когерентностей. SQ (S) — Soft качество сегментации, SQ (H) — Hard качество сегментации, N — Newman, M — Mimno, SC — SemantiC, TL — TopLen, FC — FoCon. Значения, выделенные жирным, показывают увеличение значений функций для «хорошей» модели по сравнению с «плохой» (что хорошо: функции чувствуют улучшение модели).

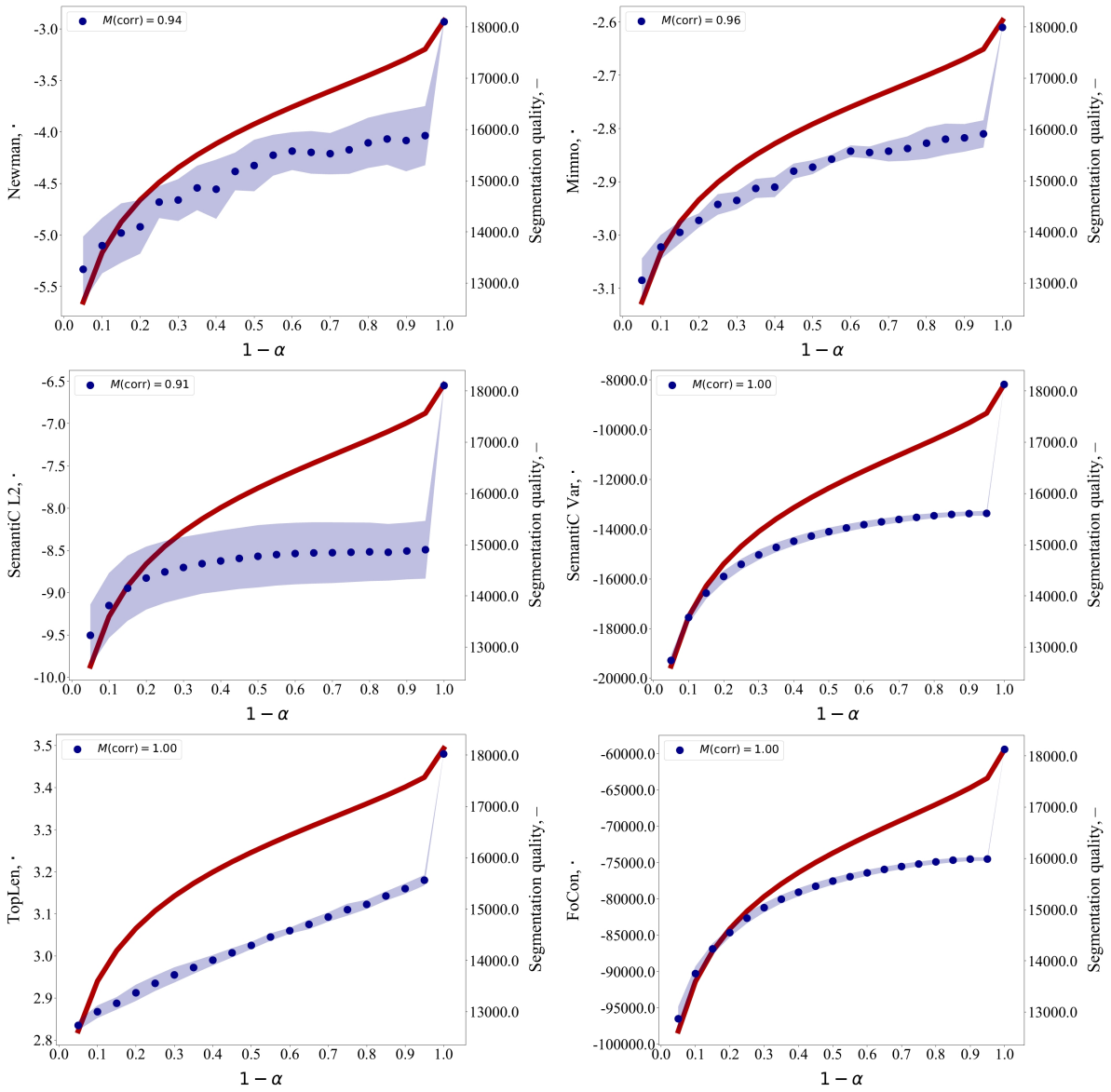

Рис. 16: Зависимость когерентностей и качества сегментации от качества тематической модели ($1 - \alpha$ — доля «хорошей» матрицы Φ_{good}). Точки, изображающие значения когерентностей, есть результат усреднения от четырёх серий экспериментов с разными «плохими» матрицами Φ_{bad}.