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»

Смирнов Евгений Александрович

Тематическая сегментация диалогов контактного центра

03.04.01 —Прикладные математика и физика

Выпускная квалификационная работа
(магистерская диссертация)

Научный руководитель:
д. ф.-м. н. Воронцов Константин Вячеславович

Москва
2018

Содержание

Аннотация

Массовая продажа продукта по телефону обычно производится согласно предписанным инструкциям в виде скрипта. Скрипт содержит множество тем, которые могут быть затронуты в диалоге с абонентом и рекомендуемую последовательность их упоминания. Контроль за качеством работы каждого из сотрудников, а также выделение наиболее успешных сценариев диалога может быть осуществлено на основе коллекции диалогов переведенных из аудио в текстовый формат. Требуется метод тематической сегментации последовательного текста для разговорной речи. В данной работе задача тематической сегментации разбита на два этапа: построение полного пула тем и представление каждого диалога в виде последовательности монотематических сегментов. Определение полного пула тем осуществляется подходом ARTM для тематической модели коротких текстов WNTM. Сравниваются два подхода для сегментации текста — подход тематического моделирования с постобработкой и нейронная сеть с рекуррентными слоями.

Ключевые слова: *тематическая сегментация, ARTM, TopMine, контактный центр, BiLSTM, FastText.*

1 Введение

Качество работы контактного центра существенно зависит от методов обучения и контроля за качеством работы операторов. В случае продажи продукта по телефону, сотрудников обучают строить диалог с абонентом согласно установленному скрипту. Основным методом оценки качества работы операторов является прослушивание небольшой подвыборки разговоров, что позволяет сформировать лишь частичное представление о работе контактного центра.

Благодаря развитию области глубокого обучения, качество распознавания речи стало приемлемым для перехода от аудио к текстовому представлению диалогов. Текст отдельного диалога позволяет автоматизировать контроль за качеством работы в каждом звонке, а массив распознанных текстов может быть использован для формирования наиболее успешных сценариев и улучшения существующего скрипта.

Скрипт задает лишь структуру диалога и для каждого его этапа содержит рекомендуемые фразы, операторы могут изменять отдельные фразы и в зависимости от речи абонента их порядок. В силу разнообразия разговорной речи, нельзя осуществлять контроль качества и выявлять успешные сценарии, используя отдельные слова в качестве структурных элементов. Необходим подход, разделяющий текст на монотематические сегменты.

Задача разбивается на определение полного списка тем и представление каждого диалога в виде последовательности монотематических сегментов. Полный список тем может быть составлен либо экспертами предметной области, либо при помощи кластеризации текстов диалогов. Исходя из практического использования сегментации текста, задачу следует ставить в терминах обучения с учителем.

Подход вероятностного тематического моделирования полагает, что появление каждого слова в каждом из документов связано с множеством латентных тем. Осуществляя двухфакторное матричное разложение частот встречаемости каждого слова из словаря в каждом документ, тематическая модель определяет для каждой темы вероятностное распределение на словаре коллекции, а также для каждого документа вероятностное распределение на множестве тем.

В работе используются тексты диалогов представленные в виде последовательности реплик оператора и абонента, разделение речи говорящих происходит на этапе записи диалогов. Каждая реплика может содержать несколько тем, темы в репликах могут повторяться. В силу мелкой гранулярности тем, будем рассматривать отдельные реплики в качестве документа.

Word Network Topic Model [1] - подход тематического моделирования для работы с короткими текстовыми документами. Короткие документы содержат мало информации о связи слов внутри документов, поэтому для построения тематической модели они заменяются на псевдо-документы. Каждый псевдо-документ соответствует отдельному слову из словаря и содержит объединение его локальных контекстов по всей коллекции.

Определение тем и границ тематических сегментов сложно-формализуемая задача для коллекции реальных диалогов. Упростить задачу и повысить качество разметки можно за счет составление списка примеров для каждой из темы. Получить описание каждой темы можно в полуавтоматическом режиме, последовательно фильтруя множество уникальных реплик и используя экспертные знания. На первом этапе, отсекаются реплики имеющие низкую вероятность темы. Далее, для из-

бавления от похожих реплик, оставшиеся реплики кластеризуются по мешку слов и выбираются центры кластеров. На последнем этапе эксперт выделяет из реплик монотематические сегменты и формирует из них описания для каждой из тем.

Речь оператора содержит значительную долю шаблонных фраз, образующих плотное покрытие множества диалогов. Выделение шаблонных фраз сводится к выделению устойчивых словосочетаний, задача может быть выполнена последовательной фильтрацией или алгоритмом TopMine [2].

В работе производится сравнение подхода тематического моделирования с нейросетевой архитектурой. Тематическая модель восстанавливает распределение слов по темам, считая известным множество тем и имея для каждой темы описание в виде набора монотематических сегментов. Далее, распределения слов по темам внутри каждого документа сглаживаются, согласно лингвистической эвристике, о близости тематических профилей соседних слов в документе. В нейросетевом подходе слова переводятся в пространство эмбеддингов. Сегментация текста осуществляется последовательностью из рекуррентных слоев.

Во втором разделе ставится задача тематического моделирования. В третьем разделе описываются подходы для извлечения устойчивых словосочетаний. В четвертом разделе представлены два подхода для тематической сегментации. В пятом разделе описывается интерфейс инструмента для разметки обучающей выборки. В шестом разделе описывается вычислительный эксперимент, в котором сравниваются два подхода для тематической сегментации.

2 Вероятностная тематическая модель

Пусть D — коллекция текстовых документов, W — множество всех употребляемых в документах слов. Будем рассматривать каждый документ $d \in D$ как последовательность из n_d терминов $(w_1, w_2, \ldots, w_{n_d})$ из словаря W. Обозначим число вхождений термина w в документ d через n_{dw}.

Предполагается, что существует конечное множество тем T, и каждое употребление термина w в каждом документе d связано с некоторой темой $t \in T$. Коллекция документов рассматривается как случайная и независимая выборка троек $\{(w_i, d_i, t_i)\}_{i=1}^n$ из дискретного распределения $p(w, d, t)$ на конечном множестве $W \times D \times T$. Термины w и документы d являются наблюдаемыми переменными, темы $t \in T$ — *латентными*(скрытыми).

Полагая, что появление термина w в документе d зависит только от темы t и воспользовавшись формулой полной вероятности получим вероятностную модель порождения данных

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

Обычно число тем $|T|$ много меньше размера коллекции $|D|$ и словаря $|W|$, поэтому задача сводится к поиску приближённого представления заданной матрицы частот

$$F = (\hat{p}_{wd})_{W \times D}\,,\ \hat{p}_{wd} = \hat{p}(w|d) = \frac{n_{dw}}{n_d},$$

в виде произведения $F \approx \Phi\Theta$ двух матриц меньшего размера — *матрицы терминов тем* Φ и *матрицы частот документов* Θ:

$$\Phi = (\phi_{wt})_{W \times D}\,,\ \phi_{wt} = p(w|t),\ \phi_t = (\phi_{wt})_{w \in W}\,,$$

$$\Theta = (\theta_{td})_{T \times D}\,,\ \theta_{td} = p(t|d),\ \theta_d = (\theta_{td})_{t \in T}\,.$$

Матрицы Φ и Θ являются *стохастическими*, то есть имеют неотрицательные нормированные столбцы, представляющие дискретные распределения. Для нахождения параметров Φ и Θ максимизируется логарифм правдоподобия выборки при ограничениях нормировки и неотрицательности:

$$L(\Phi, \Theta) = \ln \prod_{d \in D} \prod_{w \in d} p\,(w|d)^{n_{dw}} = \sum_{d \in D} \sum_{w \in d} n_{dw} \sum_{t \in T} \phi_{wt}\theta_{td} \to \max_{\Phi, \Theta}; \tag{1}$$

$$\sum_{w \in W} \phi_{wt} = 1,\ \phi_{wt} \geq 0,\ \sum_{t \in T} \theta_{td} = 1,\ \theta_{td} \geq 0. \tag{2}$$

Существенной проблемой при решении этой задачи является неединственность и неустойчивость её решения. Правдоподобие (1) зависит только от произведения $\Phi\Theta$, которое определено с точностью до линейного преобразования: $\Phi\Theta = (\Phi S)(S^{-1}\Theta)$, при условии, что матрицы $\Phi' = \Phi S$ и $\Theta' = S^{-1}\Theta$ также стохастические. Требуется метод, позволяющий контролировать преобразование S.

2.1 Аддитивная регуляризация ARTM

Для решения проблемы неединственности и неустойчивости используется подход, основанный на многокритериальной регуляризации ARTM [3]. Он позволяет строить

модели, удовлетворяющие многим ограничениям одновременно. Каждое ограничение формализуется в виде регуляризатора — оптимизационного критерия $R_i(\Phi, \Theta)$, зависящего от параметров модели. Задача сводится к максимизации линейной комбинации критериев $L(\Phi, \Theta)$ и $R_i(\Phi, \Theta)$ с неотрицательными *коэффициентами регуляризации* τ_i:

$$L(\Phi, \Theta) + R(\Phi, \Theta) \to \max_{\Phi, \Theta}, \; R(\Phi, \Theta) = \sum_{i=1}^{r} \tau_i R_i(\Phi, \Theta). \tag{3}$$

Подход ARTM разделяет множество тем на предметные и фоновые $T = S \sqcup B$. Предполагается, что каждая предметные тема описываются небольшим числом терминов, фоновая же тема содержит слова общей лексики для представленной коллекции документов. Предположение формализуются в виде регуляризаторов, накладывающих ограничения на распределения $p(w|t)$ и $p(t|d)$.

2.2 Регуляризаторы

Схожесть двух дискретных $(p_i)_{i=1}^n$ и $(q_i)_{i=1}^n$ распределений предлагается оценить дивергенцией Кульбака–Лейблера:

$$KL(p||q) \equiv KL_i(p||q) = \sum_{i=1}^{n} p_i \ln \frac{p_i}{q_i}. \tag{4}$$

Минимизация KL-дивергенции эквивалентна максимизации правдоподобия модельного распределения q по эмпирическому распределению p.

Сглаживающий регуляризатор минимизирует различие между распределениями ϕ_t, θ_d и заданными $\beta = (\beta_w)_{w \in W}$, $\alpha = (\alpha_t)_{t \in T}$:

$$R(\Phi, \Theta) = \beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} + \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max. \tag{5}$$

Разреживающий регуляризатор базируется на предположении, что каждый документ и каждый термин связан с небольшим числом тем. Он максимизирует KL-дивергенцию между модельными распределениями ϕ_t, θ_d и заданными распределениями $\beta = (\beta_w)_{w \in W}$, $\alpha = (\alpha_t)_{t \in T}$ с большой энтропией, например, равномерным:

$$R(\Phi, \Theta) = -\beta_0 \sum_{t \in T} \sum_{w \in W} \beta_w \ln \phi_{wt} - \alpha_0 \sum_{d \in D} \sum_{t \in T} \alpha_t \ln \theta_{td} \to \max.$$

Декоррелирующий регуляризатор повышает различность тем, что улучшает интерпретируемость модели. Он минимизирует ковариации между вектор-столбцами ϕ_t, ϕ_s:

$$R(\Theta, \Phi) = -\gamma \sum_{t \in T} \sum_{s \in T \setminus t} \sum_{w \in W} \phi_{wt} \phi_{ws} \to \max.$$

Регуляризатор частичного обучение улучшает интерпретируемость тем, используя априорную информацию о распределениях p(w|t) и p(t|d) для подмножества слов W' и подмножества документов D'. В частном случае, когда тематические профили слова $p(w|t)$ и документов $p(t|d)$ вырождены:

$$R(\Phi, \Theta) = \sum_{t \in T} \sum_{w \in W'} \beta_{wt} \ln \phi_{wt} + \sum_{d \in D'} \sum_{t \in T} \alpha_{td} \ln \theta_{td} \to \max.$$

2.3 Мультимодальные Тематические Модели

Мультимодальная тематическая модель описывает документы, о которых известна мета-информация. В роли мета-информации для документа могут быть представлены: авторы, моменты времени, классы, жанры, графические изображения. Мета-данные позволяют более точно определять тематику документов.

Каждый тип мета-данных образует отдельную модальность со своим словарем. Предполагается, что слова в каждой модальности приходят из разных вероятностных пространств. Модальности устойчивых словосочетаний и буквенных n-грамм образуют отдельные вероятностные пространства и могут быть рассмотрены в качестве отдельных модальностей.

Пусть M — множеством модальностей. Каждая модальность имеет свой словарь токенов W_m, $m \in M$. Эти множество попарно не пересекаются. Задача тематического моделирования для мультимодальных данных:

$$\sum_{m \in M} \tau_m \sum_{d \in D} \sum_{w \in W_m} n_{dw} \sum_{t \in T} \phi_{wt} \theta_{td} \to \max_{\Phi, \Theta}, \tag{6}$$

где τ_m — веса модальности m, позволяющие сбалансировать модальности по их важности.

2.4 WNTM

Подход тематического моделирования описывает скрытое множество тем благодаря информации о совстречаемостях групп слов в контексте документа. Матрица частот коллекции коротких документов имеет сильно разреженную структуру и не содержит статистическую информацию о совстречаемостях слов. Требуется пересмотреть понятие документа для коротких текстов.

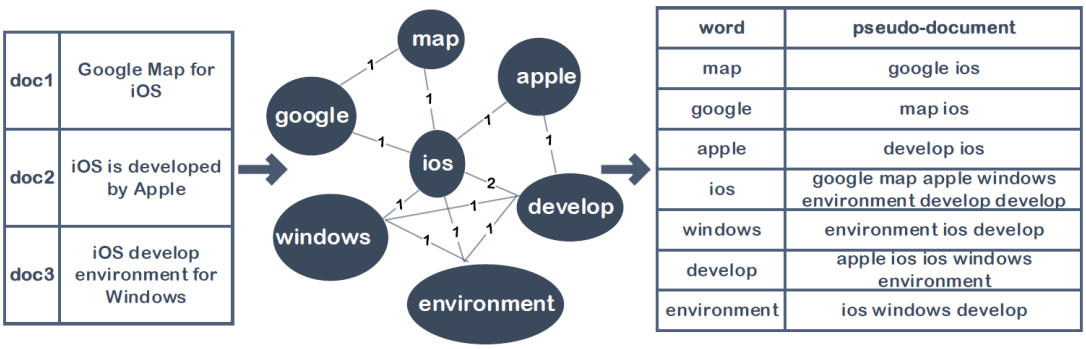


Рис. 1: Пример построения псевдо-документов. Заимствовано из [1]

Модель WNTM строит для каждого слова словаря w псевдо-документ d', содержащий объединения слов из локальных контекстов по всей коллекции документов. В локальным контекст для слова w_i в документе d в случае линейной структуры текста входит множество слов $\{w_k\}_{k=i-n}^{i+n}$ из окна длины $2n$. Далее, для коллекции документов D' и словаря W решается задача тематического моделирования в стандартной формулировке.

3 Извлечение устойчивых словосочетаний

Объединения слов в устойчивые словосочетания повышает интерпретируемость моделей машинного обучения, использующих гипотезу мешка слов. В данной работе устойчивые словосочетания используются при построении мульмодальных тематических моделей, кластеризации предложений по мешку слов и сегментации текста. В работе были исследованы два алгоритма — последовательная фильтрация и алгоритм TopMine.

3.1 Последовательная фильтрация

Устойчивым словосочетанием будем называть последовательность слов несущих законченную мысль и неслучайно часто встречающихся вместе. Оценить вероятность того, что слово v встречается неслучайно вместе со словом w в коллекции документов D можно при помощи частот совместной встречаемости $n(w, v)$ и частот слов $n(v)$ и $n(w)$:

$$pmi(v, w) = \log \frac{p(w, v)}{p(w)p(v)} = \log \frac{n(w, v)n^2}{n(w)n(v)N}, \tag{7}$$

где n — число слов в документах, N — число пар слов в документах. Дополнительно можно накладывать морфологические и синтаксические фильтры на получаемый список словосочетаний.

Таким образом, процесс выделения устойчивых словосочетаний может быть представлен в виде последовательности из частотного, pmi, морфологического и синтаксического фильтров. Точность и полноту устойчивых словосочетаний можно изменять варьируя пороги для каждого из фильтров. Дополнительно, на каждом этапе, для повышения качества фильтрации можно использовать экспертное мнение асессоров.

3.2 TopMine

TopMine — метод извлечения устойчивых словосочетаний коллекции документов без учителя. Алгоритм за два этапа преобразует последовательность слов в последовательность фраз для каждого документа. На первом этапе формируется словарь частых словосочетаний совместно с их частотами. На втором этапе слова агломерационно соединяются в фразы согласно метрике значимости.

Эффективность построения словаря частот достигается благодаря двум свойствам:

1. Если фраза P является редкой, тогда любая фраза, ее содержащая гарантированно будет редкой.

2. Если документ не содержит частых фраз длины n, тогда он гарантированно не содержит частых фраз длины $> n$

Предполагается в качестве нулевой гипотезы h_0, что коллекция документов получена в результате последовательности независимых испытаний Бернулли. Согласно этому предположению, появление фразы P может быть описано Биномиальной

случайной величиной, которая может быть аппроксимирована нормальной случайной величиной при больших объемах корпуса L:

$$h_0(f(P)) = \mathcal{N}(Lp(p), Lp(P)(1 - p(P))) \approx \mathcal{N}(Lp(P), Lp(P)), \tag{8}$$

где $p(P) = \frac{f(P)}{L}$ — вероятность успеха в испытании Бернули.

Рассмотрим фразу P, составленную из фраз P_1 и P_2. Математическое ожидание и дисперсия частоты объединенной фразы $f(P_1 \oplus P_2)$, при условии независимости появления фраз P_1 и P_2:

$$\mu_0(f(P_1 \oplus P_2)) = Lp(P_1)p(P_2),\ \sigma^2_{f(P_1 \oplus P_2)} \approx f(P_1 \oplus P_2). \tag{9}$$

Оценим различие частоты появления объединенной фразы $f(P_1 \oplus P_2)$ и ее ожидаемой частотой согласно нулевой гипотезе h_0:

$$sig(P_1, P_2) = \frac{f(P_1 \oplus P_2) - \mu_0(f(P_1 \oplus P_2))}{\sqrt{f(P_1 \oplus P_2)}}. \tag{10}$$

Далее, алгоритм агломерационно объединяет частые фразы, до тех пор, пока значение $sig(P_1, P_2)$ больше заданного порога.

4 Тематическая сегментация

4.1 Существующие датасеты для тематической сегментации

Качество работы практически всех существующих алгоритмов сегментации текста ранее оценивалось либо на синтетических корпусах, либо корпусах небольшого объема. Самый известный датасет для сравнения методов сегментации текста был представлен в работе [4]. Каждый документ коллекции был составлен конкатенацией 10 случайных предложений из корпуса Брауна. В работе [5] были вручную отсегментированных пять документов, содержащие тексты манифестов. В работе [6] качество сегментации оценивалось на двух корпусах собранных из текстов статей википедии про популярные сайты и химические элементы, каждый объемом порядка 100 документов.

В работе [7] был представлен датасет wiki-727k, содержащий 727 тысяч статей английской википедии. Каждый документ корпуса имеет иерархическую структуру сегментации, определяемую блоком содержания соответствующей статьи википедии.

4.2 Обзор алгоритмов сегментации текста

В работах [6], [8] используется байесовский подход для построение тематическая модели LDA [9]. Далее, для каждого предложения вычисляется вероятностное распределение в пространстве тем. Предложения объединяются в сегменты на основании меры когерентности рядом стоящих предложений.

В работе [7] сегментация текста рассматривается как задача обучения с учителем. Авторы используют две иерархически-связанных нейронных сети, базирующихся на архитектуре LSTM [10]. Первая сеть учит скрытое представление для каждого предложения, вторая сеть для последовательности предложений определяет является ли предложение граничным для сегмента.

4.3 Постановка задачи

Описанные выше алгоритмы сегментации текста рассматривают предложение в качестве минимального структурного элемента, предполагая что все предложения являются монотематичными. Более того, существующие методы не решают задачу определения темы для каждого сегмента. Текст диалогов, используемый в данной работе, не содержит разделения на монотематические фрагменты. В силу гранулярности разбиения на множество тем, восстановление правильной пунктуации также не разобьет текст на последовательность монотематических сегментов. Задача сводится к определению темы для каждого слова в каждой реплике.

Пусть каждая реплика $d \in D$ определяется последовательностью слов $w_{d1}, \dots w_{dn_d}$, где n_d — длина реплики d. Пусть дополнительно определено полное множество тем T для коллекции D. Задача тематической сегментации — определить для каждого слова $w \in d$ в каждой реклике $d \in D$ метку темы $t \in T$. В результате, каждая реплика d может быть представлена в виде последовательности монотематических сегментов $s_{t_1}, \dots s_{t_{n_s}}$, где n_s — число монотематических сегментов в реплике s.

В данной работе сравниваются два подхода для решения задачи тематической сегментации — подход тематического моделирование и нейросетевая архитектура с

рекуррентными слоями над эмбеддингами. Первый подход является интерпретируемым на каждом этапе и требует меньше данных для обучения, но является менее робастным и зависит от качества предобработки данных. Нейросетевой подход решает проблему OOV, но требует больше размеченных данных для обучения.

4.4 Тематическая модель для сегментации

Тематическую сегментацию текста будем производить в два этапа. Сначала построим вектора, описывающие тематические профили $p(t|w)$ для каждого слова $w \in W$. Затем для каждой реплики составим матрицу из последовательности векторов ее слов и определим границы преобладающих тем для каждого сегмента.

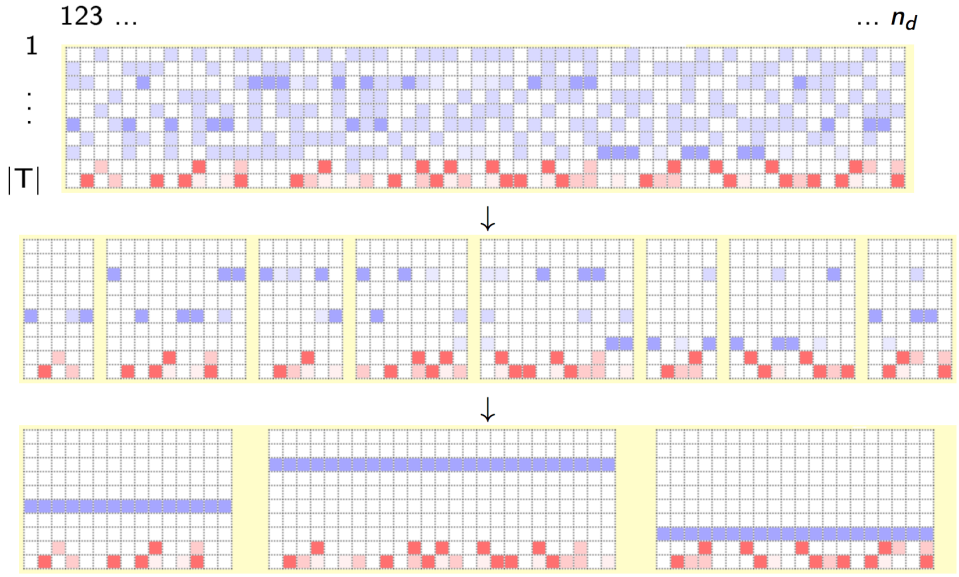


Рис. 2: Механизм сглаживание тематических профилей последовательности слов в реплике.

Будем считать, что для каждой темы $t \in T$ задано множество документов D_t, содержащих тему t и быть может фоновую $b \in S$. Восстановим распределение $p(w|t)$, построив тематическую модель с регуляризатором частичного обучения. Вычислим $p(t|w)$ для каждого слова $w \in w$ и каждой темы $t \in T$ по формуле Байеса:

$$p(t|w) = \frac{p(w|t)}{\sum\limits_{t \in T} p(w|t)}. \tag{11}$$

Темы в последовательном тексте плавно перетекают из одной в другую, а темы отдельных слов зависят от их контекста. Положим, что тематический профиль слова w_{di_j} влияет на профили слов w_{di_k}, $k \in [j-l, j+l]$ внутри окна ширины l.

Будем итеративно обновлять тематические профили слов для предметных тем. Тематический профиль слова w_{dj}, находящегося внутри окна ширины l для слова w_{dk} на итерации i, будем изменять по формуле экспоненциального сглаживания:

$$p(t|w_{dj})_{i+1} = p(t|w_{dj})_i + \frac{p(b|w_{dk})p_i(t|w_{dk})}{(j-k)^2}. \tag{12}$$

В заключении, тему для каждого слова w_{dj} определим как $t = \underset{t \in T}{argmax}\ p(t|w_{dj})$. Ширина окна l и число итераций сглаживания i_n — гиперпараметры алгоритма.

4.5 Нейронная сеть для сегментации

Нейронная сеть может выполнить аппроксимацию любой функции суперпозицией нелиненых функций [11]. В подходе тематического моделирования мы использовали эвристический метод, основанный на скользящем среднем для обновления тематических профилей соседних слов. Остается открытым вопрос, возможно ли подобрать более сложную функцию для построения тематических сегментов. Будем считать, что простой линейной модели достаточно, если ее качество окажется сравнимым с нейросетевой аппроксимацией.

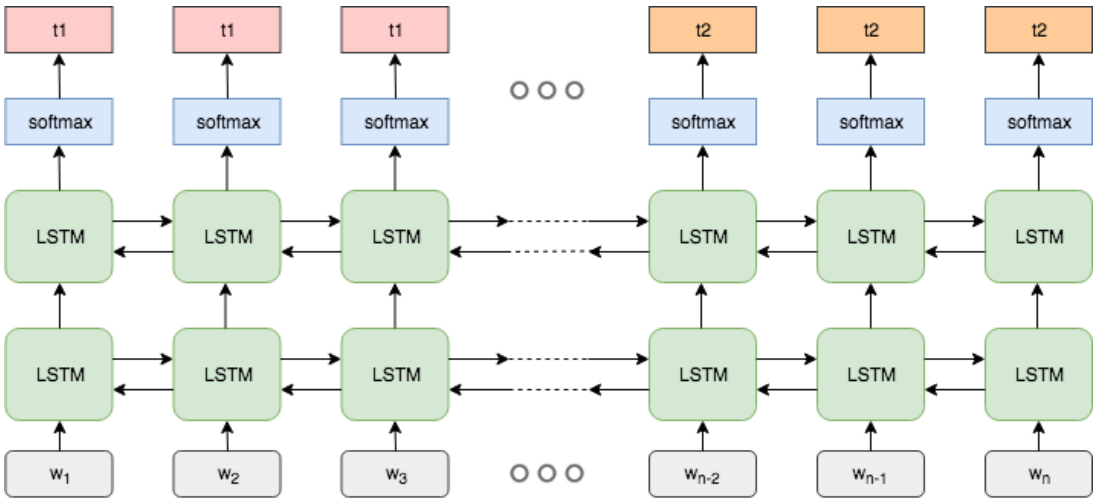


Рис. 3: Архитектура нейронной сети для тематической сегментации.

В данной работе используются эмбеддинги FastText [12] учитывающие буквенные n-граммы, чтобы нивелировать ошибки при переводе речи в текст. Связи между словами в тексте моделируются двунаправленной рекуррентная сетью, учитывающей левый и правый контексты каждого слова. Итоговая вероятность каждой из тем определяется softmax преобразованием.

5 Инструмент для разметки обучающей выборки

Асессорам требуется разделить текст на последовательность тематических сегментов. Задачу можно декомпозировать на выделение монотематичных сегментов и определение их тем.

Диалоги содержат конечное число интерпретируемых тем, так как проходят по предопределенному скрипту. Описание каждой темы можно получить, выделив монотематические сегменты из их суммаризаций [13]. Далее, описания тем будут использоваться на этапе обучения асессоров и в процессе разметки в качестве справочной информации.

Диалоги происходят по однотипным сценариям, строго регламентированным в начале разговора и допускающим вариации на более поздних стадиях диалога. Будем выбирать репрезентативное подмножество реплик для увеличения эффективности разметки диалогов. Дополнительно для каждой реплики будем отображать контекст в виде трех предыдущих и трех последующих реплик в целях улучшения их интерпретируемости.

Рис. 4: Интерфейс ассессорского интрумента для тематической сегментации.

Инструмент реализован в рамках проекта VisARTM [14]. Реплика для разметки(Блок 3) вместе с его контекстом(Блок 2) отображается в интерфейсе приложения. Асессор последовательно выделяет монотематические сегменты и назначает им темы. Каждая тема получает свой уникальный цвет при выборе из списка неактивных тем(Блок 6) и перемещается в список активных тем(Блок 5). При щелчке по активной теме появляется ее описание(Блок 7). Дополнительно асессору отображает номер и идентификатор задания(Блок 1).

6 Вычислительный эксперимент

В работе использовались тексты диалогов операторов с абонентами по продажам продуктов банка [15], [16]. Текст каждого диалога разбит на последовательность реплик с указанием метки говорящего. В репликах отсутствует разделение на предложения и любая другая пунктуация. Ошибка распознавание речи измеряется долей неправильно распознанных слов WER и составляет 0.13 — для речи оператора и 0.25 — для речи абонента. Реплики абонентов — короткие, монотематичные и распознаются с худшим качеством, поэтому в работе использовались только реплики операторов. На момент написания работы доступно порядка 100 гигабайт распознанных текстов.

Вычислительные эксперименты производились на двух серверах — ubuntu, 24 CPU, 64гб RAM и ubuntu, 40 CPU, 256гб RAM, GPU Tesla M40 и ноутбуке — OS X, 8 CPU, 16гб RAM. В работе использовались библиотеки BigARTM [17] для построения тематических моделей, keras [18] для обучение нейронной сети. Дополнительно был создан docker-контейнер для синтаксического разбора предложений на русском языке с моделью syntaxnet с поддержкой GPU [19].

6.1 Предобработка данных

Текст проходит несколько стадий предобработки. Сначала весь текст лемматизируется — слова приводятся в нормальную форму, дополнительно выделяются именованные сущности: имена, города, время, цифры. Параллельно осуществляется синтаксический разбор предложений Рис. 5.

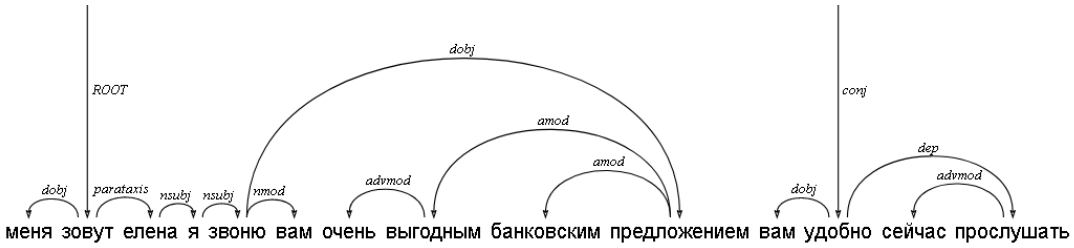


Рис. 5: Пример синтаксического разбора реплики.

смотрите мы предлагаем с первоначальным кредитным лимитом до трёхсот тысяч рублей можно указать желаемую беспроцентный период на покупки по карте до пятидесяти пяти дней это значит что когда вы расплачиваетесь картой потраченную сумму возвращаете беспроцентный период что вы ничего не переплачиваете ни каких процентов также по карте есть бонусная программа работает тогда расплачиваетесь картой получаете за то бонусные баллы можете экономить на другие и при этом можно не личные средства хранить также расплачиваясь получать баллы от покупок оформление для вас совершенно бесплатно и дома первой расходной операции ничего абсолютно никаких средств вас не взимается то есть вы можете получить у вас ознакомительное предложение оно что вы активировали

(a)

смотреть мы предлагать с первоначальный кредитный_ лимит до number_тысяча_рубль можно указывать_желать беспроцентный_период на покупка по карта до number date_time это_значить что когда вы расплачиваться_карта потратить_сумма возвращать беспроцентный_период что вы ничто не переплачивать ни какой процент также по карта быть бонусный_программа работать тогда расплачиваться карта_получать за то бонусный_балл мочь _экономить на другой и при это можно не личный_средство хранить также расплачиваться получать_балл от покупка оформление для вы совершенно_бесплатно и дома number расходный_операция ничто абсолютно никакой средство вы не взиматься то быть вы мочь получать у вы ознакомительный_предложение оно что вы активировать

(b)

Рис. 6: Пример предобработки реплики из исходного формата(a) в лемматизированный с выделением именованных сущностей и устойчивых словосочетаний(b). Синим цветом выделены именованные сущности, красным - устойчивые словосочетания.

Далее, формируется множество устойчивых словосочетаний, слова которых объединяются в один токен нижним подчеркиванием. В данной работе было выделено 19 тысяч словосочетаний.

Примеры устойчивых словосочетаний
стационарный_телефон, годовой_обслуживание, день_добрый, банк_другой, перевод_баланс, денежный_средство, план_тарифный, беспроцентный_период, сотрудник_являться, процентный_ставка, вопрос_какой, воспользоваться_карта, наличие_паспорт, номер_мобильный, время_удобный, не_слышать, номер_телефон, представитель_банк, салон_связь, бонусный_программа, тинькофф_точка, день_календарный, операция_расходный, оформление_заявка, мобильный_приложение, высокий_статус, кредитный_лимит, работать_круглосуточно, абсолютно_бесплатно, карта_кредитный, индивидуальный_предложение, погасить_задолженность, экономить_время, сумма_какой, кабинет_личный, начисляться_процент, действующий_кредит

На Рис. 6 представлена реплика в исходном и предобработанном форматах.

6.2 Подвыборка для разметки

Постановка задачи для асессоров начинается с определения множества тем и составления их описаний для двух продуктов банка - кредитная карта и обслуживание среднего и малого бизнеса SME. Сначала строится тематическая модель на коротких текстах с модальностью устойчивых словосочетаний, сглаживанием фоновых и декорреляцией предметных тем. Далее, для каждой темы составляется ее суммаризация. Из ранжированного списка суммаризации экспертно выбираются монотематические части реплик, таким образом определяется множетсво тем и их описание. В результате, было выделено 50 тем для каждого продукта и составлены описания для каждой из них.

Описание темы «интернет банк»:

- *смотрите если будете клиентом банка у вас будет кабинет вот свой личный у нас на сайте тинькофф точка ру сможете легко оплачивать телефон интернет штрафы многие другие услуги при том же без комиссии*
- *тинькофф банк признан одним из самых лучших интернет банков*
- *а как вы оплачиваете ежемесячные обязательные платежи штрафы интернет телефонию коммунальные услуги если не секрет*
- *в интернет банке вы также сможете даже активировать карту*
- *в интернет банке сможете узнать о различных акциях банка*

Множество реплик для разметки выбирается так, чтобы были покрыты все устойчивые словосочетания. Сначала для каждой реплики определяется множество

входящих в нее словосочетаний. Предполагается, что реплика содержащее большее число словосочетаний, содержит большее число тем и потому более ценна для попадания в обучающую выборку. Реплики сортируются по убыванию числа устойчивых словосочетаний. Будем называть словосочетание активным, если оно встретилось в обучающей выборке менее 5 раз. Затем реплики, содержащие активные словосочетания, последовательно добавляются в обучающую выборку, пока множество активных словосочетаний не станет пустым.

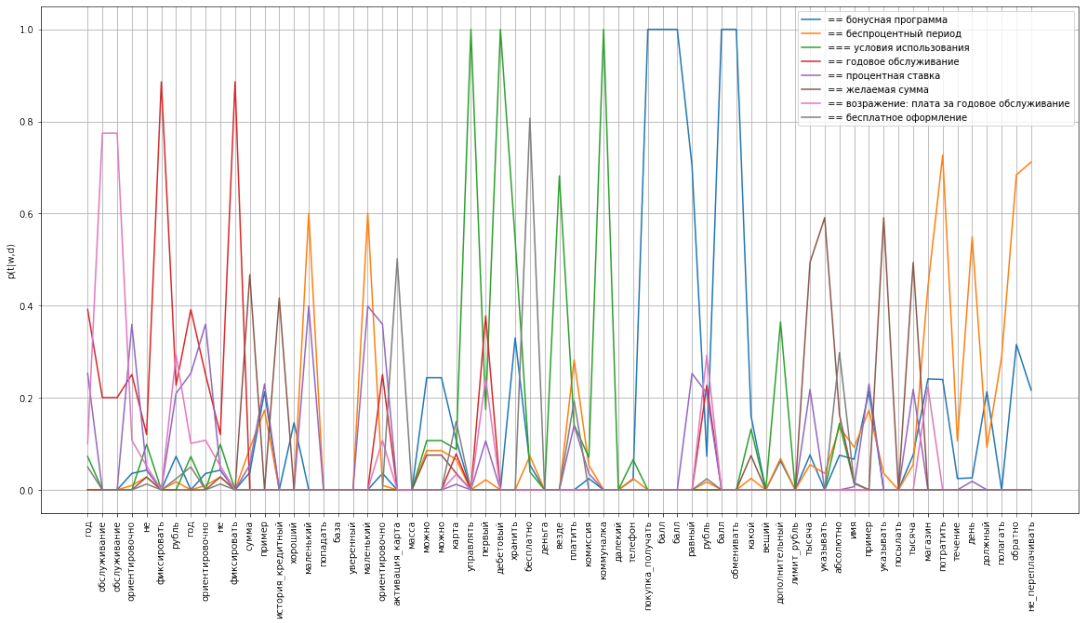
В результате, для продукта кредитная карта была получена обучающая выборка из 120 тысяч реплик, содержащих 700 тысяч монотематичных сегментов, для продукта SME — 33 тысячи реплик и 200 тысяч монотематичных сегментов.

6.3 Тематическое моделирование для тематичской сегментации

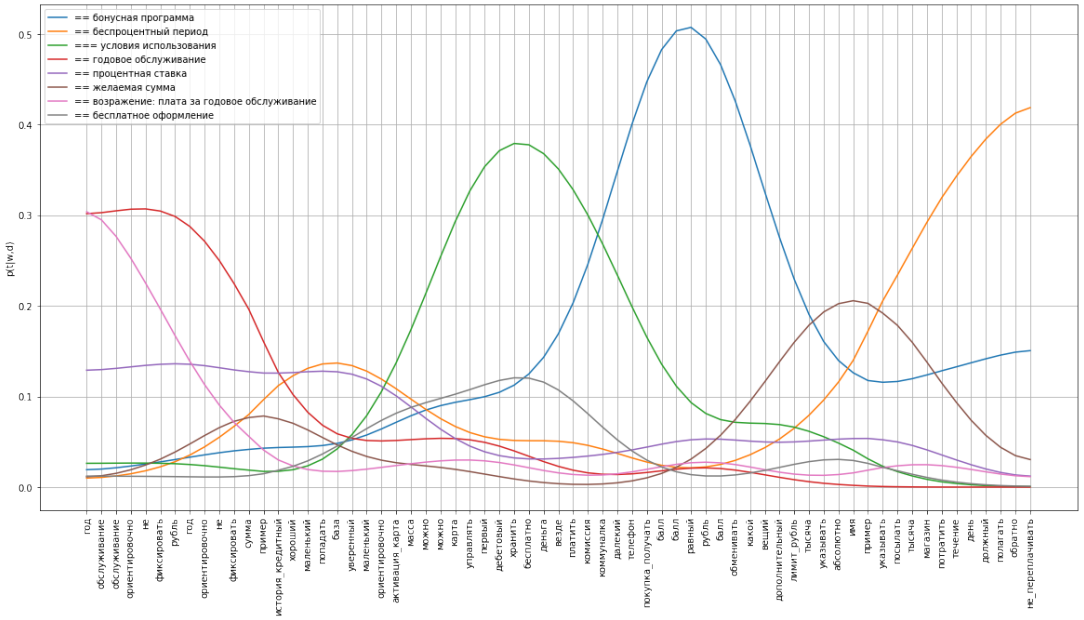
Монотематические документы для каждой темы были сгруппированы в документы. Далее, монотематичные документы были использованы для построения тематических профилей каждого из слов словаря при помощи регуляризатора частичного обучения.

Топ-токены темы «данные абонента»:
телефон индекс адрес номер заканчиваться указывать фамилия работать совпадать улица место мобильный исполняться стационарный_телефон адрес_рабочий дата_рождение номер_телефон место_проживание почта_электронный доход_персональный доход_месяц код_подразделение постоянный_регистрация номер_серия номер_паспорт кодовый_слово доход_составлять адрес_регистрация адрес_проживать индекс_почтовый

Далее, было произведено экспоненциальное сглаживание тематические профили для каждой реплики Рис. 7. Пример тематической сегментации текста представлен на 8.



(a)



(b)

Рис. 7: Пример сглаживания тематических профилей для последовательности слов реплики. Рисунок(a) — до сглаживания, рисунок(b) – после сглаживания.

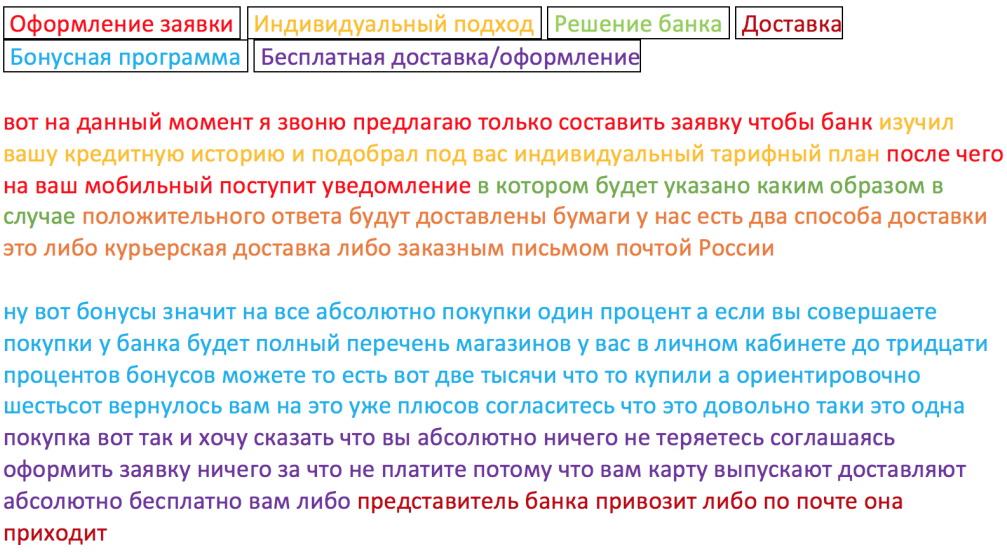


Рис. 8: Пример тематической сегментации реплики оператора.

6.4 Нейронная сеть для тематичской сегментации

Нейронной сети на вход подавались вектора FastText, обученные на корпусе из 900 миллионов лемматизированных реплик. При обучении нейронной сети, использовалась нормализация весов [20] и метод прореживания [21] рекуррентных слоев для регулирования переобучения сети. Нейронная сеть обучалась методом $AdamOptimezer$ с параметром $learning\_rate = 0.001$ и размером батча 128.

6.5 Сравнение моделей

Качество сегментации оценивалось по доле правильно угаданных тем у слов, с исключением из оценки неразмеченных асессорами слов:

$$Accuracy = \frac{1}{|D|} \sum_{w \in D} [t_i = t_i^*] \,. \tag{13}$$

Модель/Датасет	Кредитная карта	SME
Тематическая модель	0.725	0.687
Нейронная сеть	0.749	0.715

Таблица 1: Сравнение подходов для тематической сегментации. Метрика качества Accuracy.

Качество сегментации с использованием подхода тематического моделирования оказалось сравнимым с качеством сегментации нейронной сетью. Нейронная сеть является более робастной за счет использования буквенных n-грамм при получении векторов слов. Подход тематического моделирования более интерпретируемый и требует меньше вычислительных ресурсов.

7 Заключение

В работе был представлен метод определения полного пула тем коллекции диалогов с фиксированным скриптом и составления их интерпретируемого описания. Предложен метод выделения репрезентативной выборки и разработан асессорский инструмент для ее разметки. Представлены два подхода для тематической сегментации текста. Первый подход интерпретируем на каждом этапе, основанный на подходе тематического моделирования. Сначала строились тематические профили для каждого слова реплики, затем производилось их экспоненциальное сглаживание для каждой темы независимо. Второй подход — нейронная сеть, состоящая из последовательности рекуррентных слоев и эмбеддинги FastText, за счет чего является более робастным. Результаты работы были использованы для выделения наиболее успешных сценариев диалогов и будут использованы для контроля качества работы операторов. Результат работы тематической сегментации текста может быть использован в качестве интерпретируемых признаков.

Список литературы

[1] Yuan Zuo, Jichang Zhao, and Ke Xu. Word network topic model: a simple but general solution for short and imbalanced texts. *Knowledge and Information Systems*, 48(2):379–398, 2016.

[2] Ahmed El-Kishky, Yanglei Song, Chi Wang, Clare R Voss, and Jiawei Han. Scalable topical phrase mining from text corpora. *Proceedings of the VLDB Endowment*, 8(3):305–316, 2014.

[3] Константин Вячеславович Воронцов and АА Потапенко. Аддитивная регуляризация тематических моделей. In *Доклады Академии наук*, volume 456, pages 268–271, 2014.

[4] Freddy YY Choi. Advances in domain independent linear text segmentation. In *Proceedings of the 1st North American chapter of the Association for Computational Linguistics conference*, pages 26–33. Association for Computational Linguistics, 2000.

[5] Goran Glavaš, Federico Nanni, and Simone Paolo Ponzetto. Unsupervised text segmentation using semantic relatedness graphs. Association for Computational Linguistics, 2016.

[6] Harr Chen, SRK Branavan, Regina Barzilay, and David R Karger. Global models of document structure using latent permutations. In *Proceedings of Human Language Technologies: The 2009 Annual Conference of the North American Chapter of the Association for Computational Linguistics*, pages 371–379. Association for Computational Linguistics, 2009.

[7] Omri Koshorek, Adir Cohen, Noam Mor, Michael Rotman, and Jonathan Berant. Text segmentation as a supervised learning task. *arXiv preprint arXiv:1803.09337*, 2018.

[8] Martin Riedl and Chris Biemann. Topictiling: a text segmentation algorithm based on lda. In *Proceedings of ACL 2012 Student Research Workshop*, pages 37–42. Association for Computational Linguistics, 2012.

[9] David M Blei, Andrew Y Ng, and Michael I Jordan. Latent dirichlet allocation. *Journal of machine Learning research*, 3(Jan):993–1022, 2003.

[10] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[11] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[12] Piotr Bojanowski, Edouard Grave, Armand Joulin, and Tomas Mikolov. Enriching word vectors with subword information. *arXiv preprint arXiv:1607.04606*, 2016.

[13] Евгений Александович Смирнов and КВ Воронцов. Суммаризация тем в вероятностных тематических моделях. *machinelearning.ru*, 2016.

[14] Visartm - инструмент для визуализации тематических моделей. *GitHub*.

[15] Andrew Stepanov. Character level recurrent neural network language models for morphologically complex languages in speech recognition, 7 2017. In Russian.

[16] Anastasiia Torunova. Syllable-level acoustic modeling with convolutional neural networks, 7 2017. In Russian.

[17] Konstantin Vorontsov, Oleksandr Frei, Murat Apishev, Peter Romov, and Marina Dudarenko. Bigartm: Open source library for regularized multimodal topic modeling of large collections. In *International Conference on Analysis of Images, Social Networks and Texts*, pages 370–381. Springer, 2015.

[18] François Chollet et al. Keras, 2015.

[19] Смирнов Евгений. Docker syntaxnet gpu для русского языка. `https://hub.docker.com/r/evgenysmirnov/syntaxnet`, 2017.

[20] Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. *arXiv preprint arXiv:1502.03167*, 2015.

[21] Vincent Tinto. Dropout from higher education: A theoretical synthesis of recent research. *Review of educational research*, 45(1):89–125, 1975.