Московский государственный университет имени М.В. Ломоносова
Факультет вычислительной математики и кибернетики
Кафедра алгоритмических языков

Аникевич Юлия Вадимовна

Методы извлечения аспектов из мнений пользователей

Магистерская диссертация

Научный руководитель:
к.ф.–м.н.
Н.Э. Ефремова

Москва, 2023

Извлечение аспектов

Мнение – суждение автора по поводу некоторого объекта/услуги/продукта/товара/организации/...

Сущность – объект мнения (продукт, сервис, тема, человек, организация, ...)

Аспекты (аспектные категории) – составные части, свойства, характерные признаки и черты сущности

1. Выделение аспектных терминов

Официанты вели себя по-хамски

Стейк был очень вкусным

Интерьер на высоте

2. Формирование аспектных категорий

Стейк Чизкейк Салат Огуречный лимонад ...	Официант Персонал Сервис ...	Дизайн Атмосфера Освещение ...
ЕДА	ОБСЛУЖИВАНИЕ	ОКРУЖЕНИЕ

3. Маркировка предложений аспектами

Официанты вели себя по-хамски –> ОБСЛУЖИВАНИЕ

Стейк был очень вкусным –> ЕДА

Интерьер на высоте –> ОКРУЖЕНИЕ

Задача диссертации

В работе была рассмотрена задача извлечения аспектов из мнений пользователей. Для ее решения было необходимо:

1. Провести обзор современных методов решения задачи извлечения аспектов.

2. По результатам обзора выбрать методы извлечения аспектов из мнений на русском языке и программно их реализовать.

3. Подготовить наборы мнений о двух различных сущностях.

4. Провести экспериментальное исследование качества работы реализованных методов и проанализировать полученные результаты.

Выбранные методы

- **Метод на основе частоты встречаемости слов (FREQ)**

  аспектные термины - наиболее частотные существительные; аспекты - кластеры над аспектными терминами

- **Тематическая модель битермов –** пары слов в предложении **(BiTERM)**

  тематическая модель, основанная на совместном появлении слов

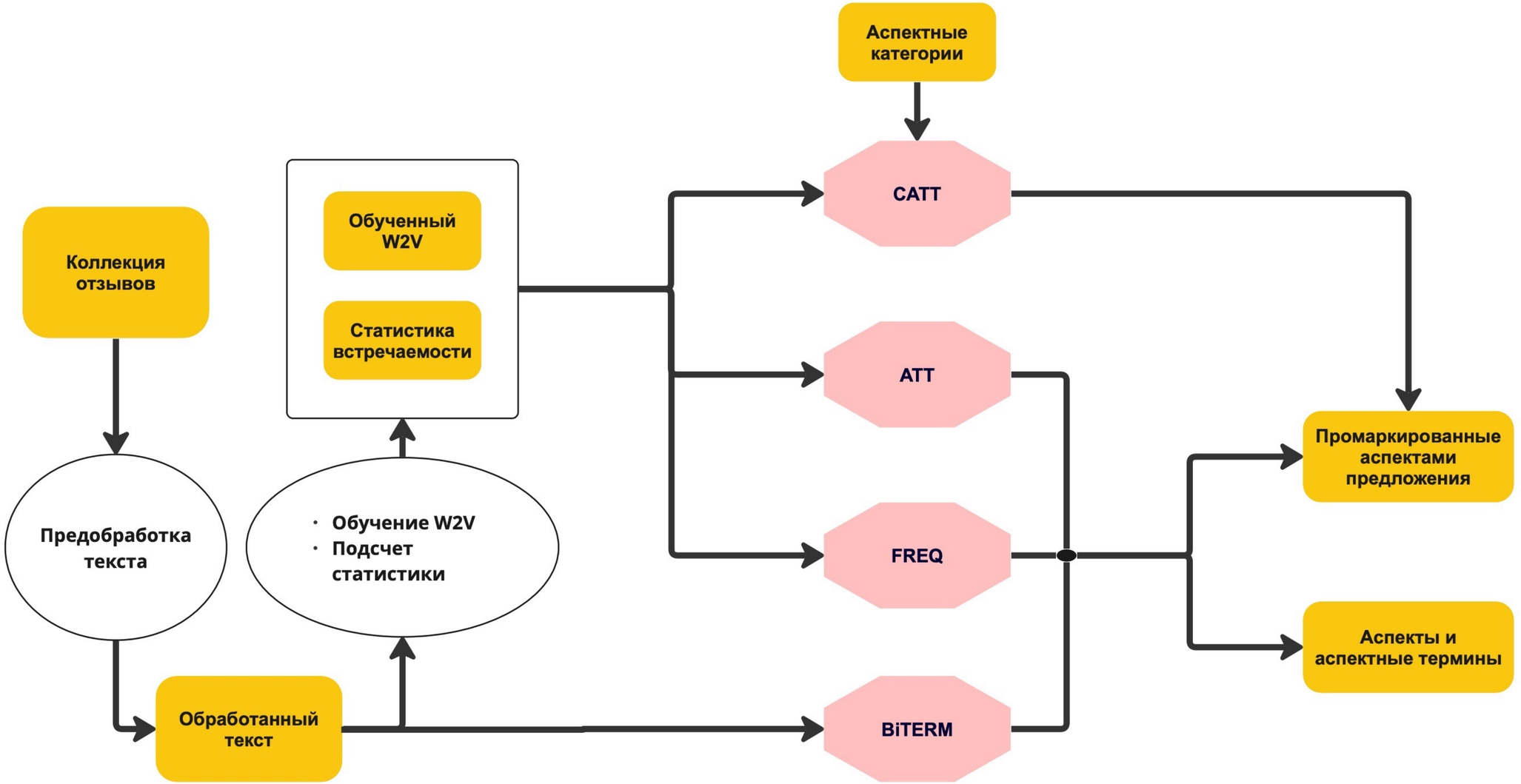
- **Метод на основе механизма внимания (ATT)**

  автокодировщик с механизмом внимания

- **Метод на основе контрастивного внимания (CATT)** – *только для решения подзадачи маркировки предложений аспектами*

  использование контрастивного внимания и ядро радиальной базисной функции (RBF)

Архитектура программной реализации



Наборы данных

Рестораны

- Мнения собраны с Google Карт с помощью инструмента Google Maps Reviews Crawler.
- Размеченные предложения взяты из набора, предоставленного в рамках соревнования на конференции SentiRuEval в 2015 году
- Набор содержит 419345 мнений и 404 размеченных аспектами предложения
- Аспекты: Кухня, Интерьер, Сервис, Цена

Мобильные телефоны

- Набор отзывов был взят с Kaggle.
- Разметка была произведена на платформе Яндекс.Толока.
- Набор данных про мобильные телефоны содержит 458433 отзывов и 1300 размеченных аспектами предложения
- Аспекты: Камера, Аккумулятор, Динамики, Цена, Память, Дисплей, Производительность

Оценка работы методов

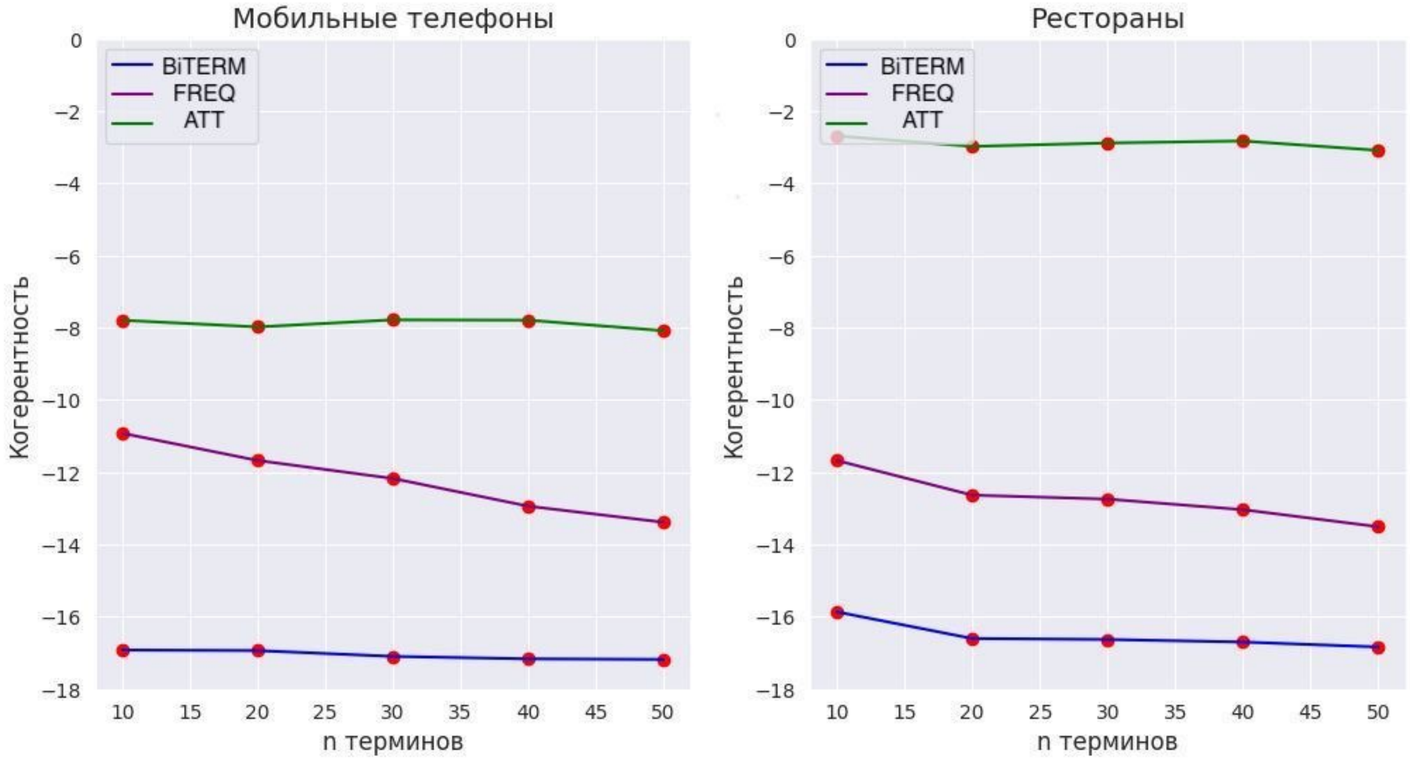
Когерентность аспектов – это мера, используемая для оценки связности полученных аспектов. Чем выше значение, тем более интерпретируемым считается аспект

$$C(z; S^z) = \sum_{n=2}^{N} \sum_{l=1}^{n-1} log \frac{D_2(w_n^z, w_l^z) + 1}{D_1(w_l^z)}$$

$S^z = \{w_1^z, ..., w_N^z\}$ – набор из *N* более близких слов аспекта *z*

$D_1(w)$ – частота использования слова *w* в предложениях

$D_2(w_1, w_2)$ – частота совместного использования слов w_1 и w_2 в предложениях



Оценка работы методов. Экспертная оценка

Постановка задачи:

Даны наборы терминов. Необходимо предложить, о каком аспекте предложенной сущности идет речь и разметить термины, которые к предложенному аспекту не относятся. Термины упорядочены по близости к аспекту.

- 4 эксперта
- 10 аспектов
- 50 терминов на каждый аспект

Результаты:

Количество верно извлеченных аспектов (более половины экспертов смогли определить аспект):

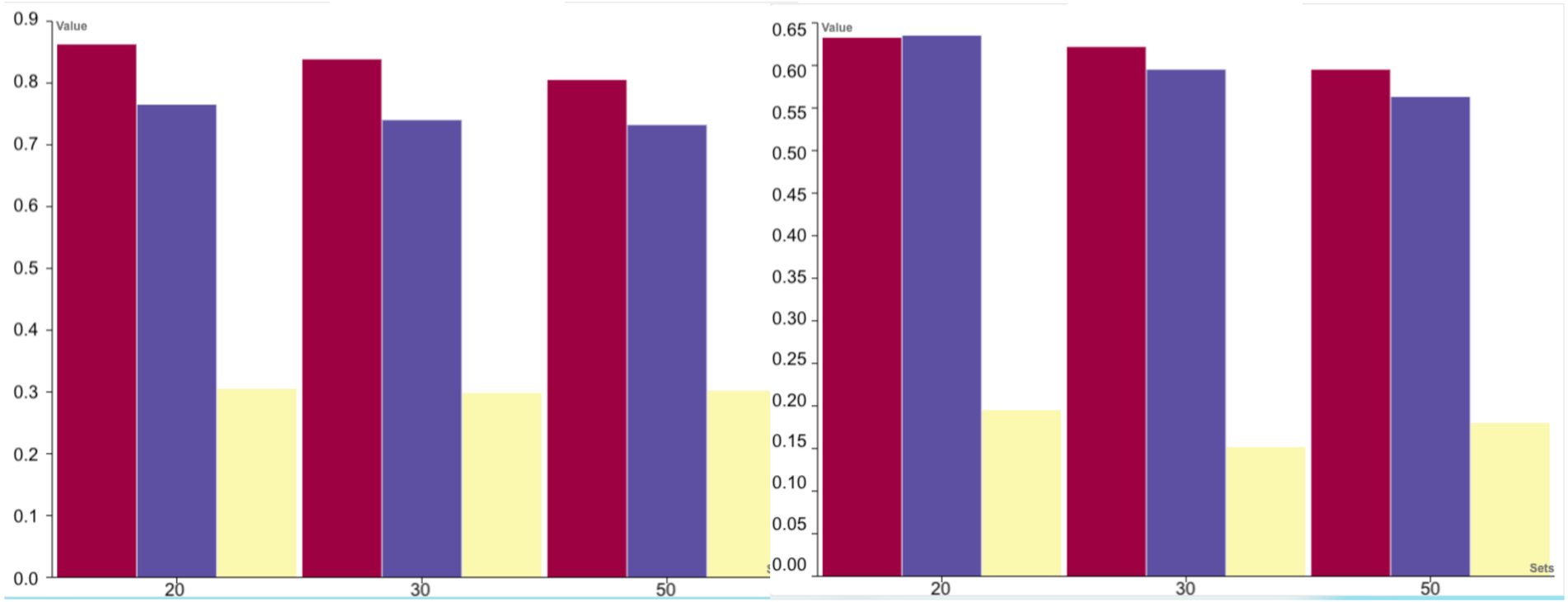
	Рестораны	Телефоны
FREQ	7	**8**
ATT	**9**	7
BiTERM	5	5

	FREQ	ATT	BiTERM
1	Расположение – Локация –	Интерьер – Пространство Интерьер	Интерьер Обстановка Пространство Интерьер
2	Профессионализм Конфликт Сотрудник Стафф	Обслуживание Ожидание Обслуживание Обслуживание	– – – –
3	– – – –	Меню Меню Еда Меню	– Впечатление – Впечатление
4	– Обстановка Пространство –	Празднование Фуршет Праздник Праздничное мероприятие	Обслуживание Ожидание Обслуживание Обслуживание
5	Еда Еда Блюдо Меню	Сотрудники Персонал Администрация Резерв	– – – –

Оценка работы методов. Экспертная оценка

Используется оценка ***precision@n*** на основе экспертной разметки в зависимости от общего числа терминов n и размеченных среди них нетерминов m :

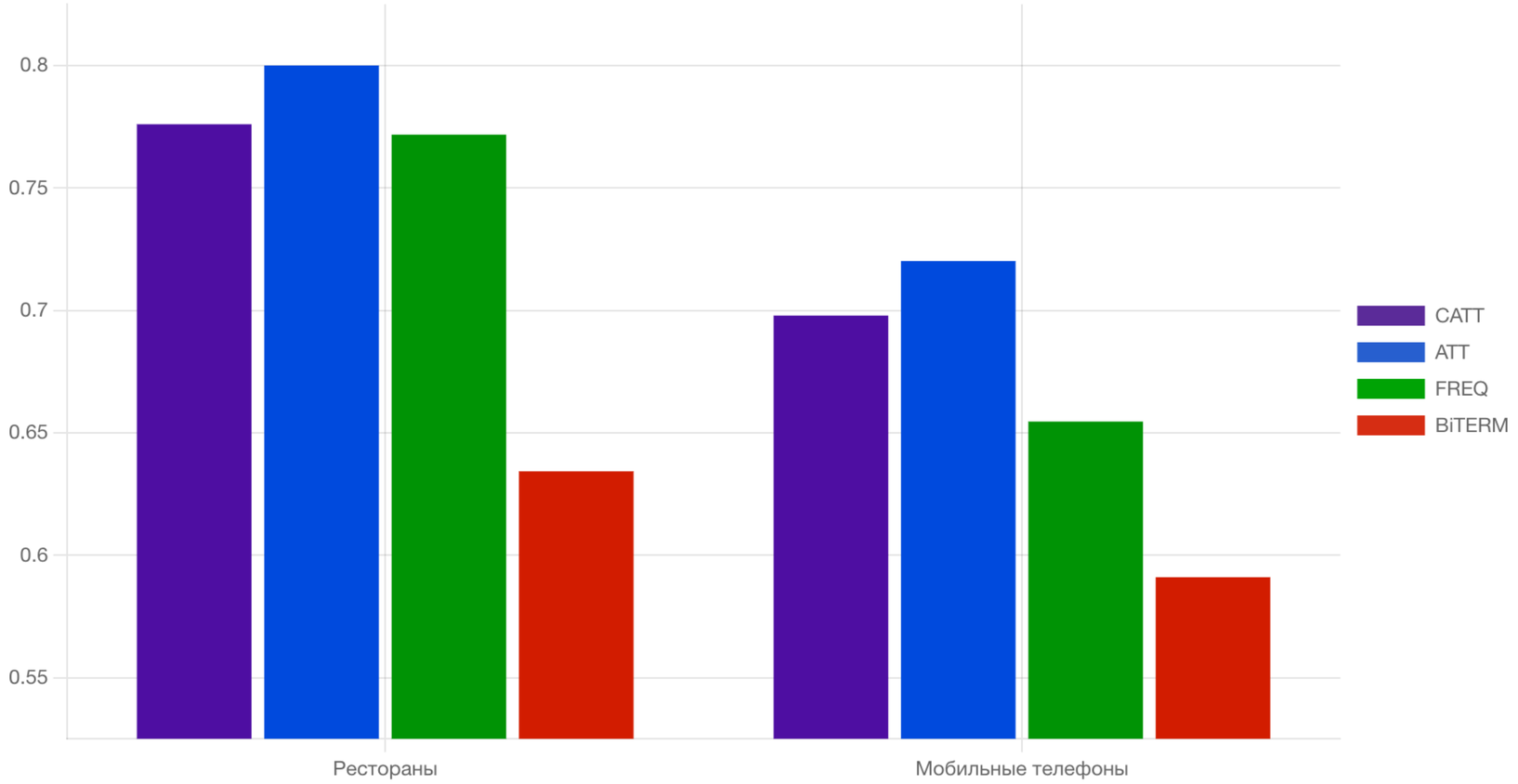
$$precision@n = \frac{\boldsymbol{m}}{\boldsymbol{n}}$$



Рестораны **Мобильные телефоны**

Оценка работы методов. Маркировка предложений

На диаграмме представлены значения метрики макро-F1 (усредненная метрика F1 по всем аспектам) для каждого метода:



Результаты

В данной работе получены следующие результаты:

1. Проведен обзор актуальных подходов к решению задачи извлечения аспектов.

2. По результатам обзора выбраны и программно реализованы методы для применения к мнениям на русском языке

3. Подготовлены наборы мнений о двух различных сущностях: рестораны и мобильные телефоны

4. Проведено экспериментальное исследование качества работы реализованных методов и проанализированы полученные результаты.

Опубликована статья в сборнике "Вестник научных конференций”. Реализованные методы и собранные наборы данных выложены в открытый доступ.

Спасибо за внимание

Экспертная оценка – Рестораны

	FREQ	ATT	BiTERM
1	Расположение – Локация –	Интерьер – Пространство Интерьер	Интерьер Обстановка Пространство Интерьер
2	Профессионализм Конфликт Сотрудник Стафф	Обслуживание Ожидание Обслуживание Обслуживание	– – – –
3	– – – –	Меню Меню Еда Меню	– Впечатление – Впечатление
4	– Обстановка Пространство –	Празднование Фуршет Праздник Праздничное мероприятие	Обслуживание Ожидание Обслуживание Обслуживание
5	Еда Еда Блюдо Меню	Сотрудники Персонал Администрация Резерв	– – – –
6	Персонал Персонал Сотрудники Персонал	Общие впечатления Стиль Атмосфера Атмосфера	– – – –
7	Оценка Оценка качества Оценка Оценка	Паттерн посещения Любимое место Локация –	Стоимость Меню Цена Стоимость
8	Напитки Напитки Напиток Напитки	Рецепт Рецепт Готовка Рецепт	Еда Еда Еда Кухня
9	Еда Горячее Основное блюдо Меню	Стилистика Интерьер Стиль Стиль	Общие впечатления Оценка Качества Впечатления
10	Состав Ингредиенты Ингредиенты Ингредиенты	– – Событие Семейное посещение	– – – –

Экспертная оценка – мобильные телефоны

	K-means	ABAE	BTM
1	– Покупка Покупка Покупка	– Замена – Замена	– – – –
2	Камера Камера Камера Камера	Форм-фактор Дизайн Стиль Внешний вид	Батарея Батарея Зарядка Батарея
3	Настройка Настройки – Перепрошивка	Использование Функционал Мобильный интернет Приложения	– – Сервис –
4	Повреждения Прочность Падение Краш-тест	Состояние Качество покрытия Состояние Корпус	– – – –
5	Операционные системы Модель – Модели	Повреждения Повреждение Поломка Краш-тест	– Покупка Стоимость Стоимость
6	Модели Модели Модели Бренды	Камера Камера Камера Камера	– – – –
7	– – – –	– Глюки Звонок Звонок	Покупка – Покупка Покупка
8	Использование Социальные сети Социальные сети Приложения	Модели Модель Модель Бренды	Повреждения – Поломка Повреждения
9	Производительность Производительность Глюк Глюки	– Приложение – Прошивки	– – Использование Приложения
10	– Магазин Продажа Магазин	– – – –	Модели Модель Модель Бренды

Пример экспертной разметки

[1] Интерьер

кабинка телевизор зона лестница проход оборудовать экран
беседка столиками вдоль коридор туалет вентиляция площадка
вешалка дуть кровать сквозняк зеркало дверь обувь здание лифт
балкон парковка игрушка потолок коляска пыль сидение лампа
раковина пешеходный кондиционер кресло расположен стул мультик
пыльный диван стена висеть комната повесить расположена
пространство туалетный помещение курение стоянка

[2] Обслуживание

ждали нести прождать ждать остыть мина через холодный минута
полчаса просидеть принести минут неполный спустя заказали
дождаться сорок полупустом напоминание наполовину горячее
долго горячий двадцать час яичница вареник напитки около
душный подкачать пустой готовиться манты подать еду грязный
третий остывать сырник пол освободиться вынести медленно
безвкусный заветренный полтора напоминать

Метод FREQ

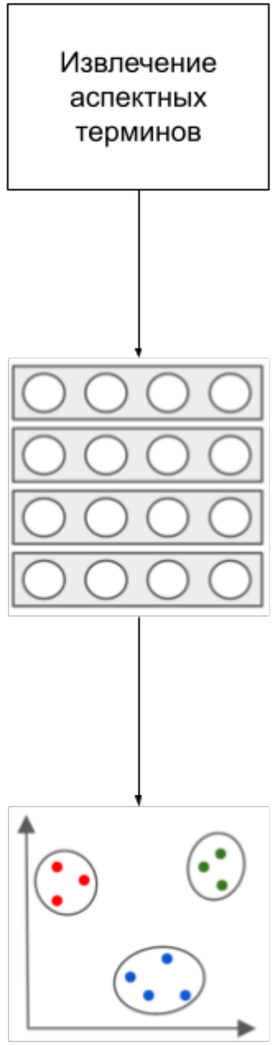
Идея: аспектные термины - наиболее частотные существительные; аспекты - кластеры над аспектными терминами

Вход: порог отсечения по частоте, количество кластеров.

Основные шаги:

1. Извлечение аспектных терминов. Извлекаются существительные, частота которых выше заданного порога отсечения.

2. Получение векторного представления терминов.
3. Получение аспектов. Кластеризация аспектных терминов в кластеры. (А мы их как-то именуем?) Имя аспекта - …
4. Разметка предложений. …



Метод BiTERM

Идея: вероятность появления аспекта в предложении равна ожидаемой вероятности того, что к этому аспекту относятся аспектные термины (битермы) из этого предложения.

Вход: количество аспектов категорий, параметры распределения Дирихле, количество итераций

Основные шаги:

1. Извлечение аспектные терминов (битермов). Рассматриваются …

2. Инициализация матрицы, отражающей количество вхождений (нужно обсудить!) каждого слова (? Или битерма) в каждый аспект.

3. Определение аспектов для каждого битерма в каждом документе (? Предложении?) с помощью вероятностных распределений.

4. Кластеризация битермов на аспекты.
5. Разметка предложений. …

Тематическая модель битермов

$$P(z|d) = \sum_b P(z|b)P(b|d)$$

где

$$P(z|b) = \frac{P(z)P(w_i|z)P(w_j|z)}{\sum_z P(z)P(w_i|z)P(w_j|z)}$$

$$P(b|d) = \frac{n_d(b)}{\sum_b n_d(b)}$$

Модель на основе механизма внимания (…)

Идея: использование автоэнкодер с механизмом внимания

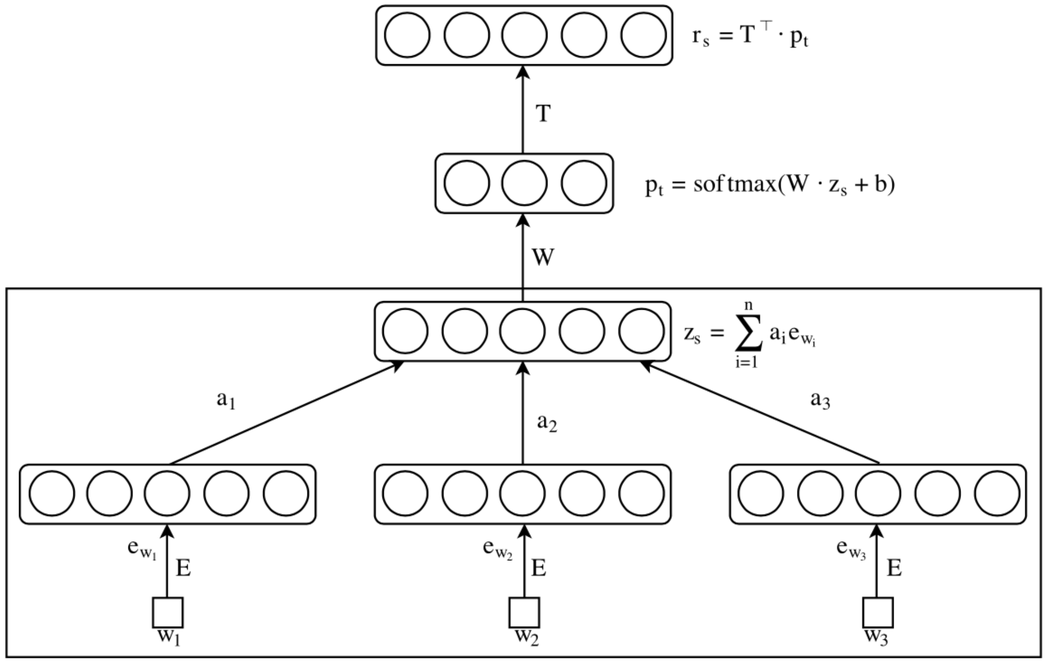
Вход: количество аспектных категорий, инициализация матрицы аспектов

Основные шаги:

1. Получение вектора предложения как взвешенную сумму векторов слов

2. Уменьшение размерности полученного предложения

3. Реконструкция вектора предложения как линейную комбинацию векторов из матрицы представлений аспектов, где коэффициенты интерпретируются как вероятность отнесения к аспекту



Модель на основе механизма внимания

1. Получение вектора предложения как взвешенную сумму векторов слов

$$z_s = \sum_{i=1}^{n} a_i e_{w_i} \qquad y_s = \frac{1}{n} \sum_{i=1}^{n} e_{w_i}$$

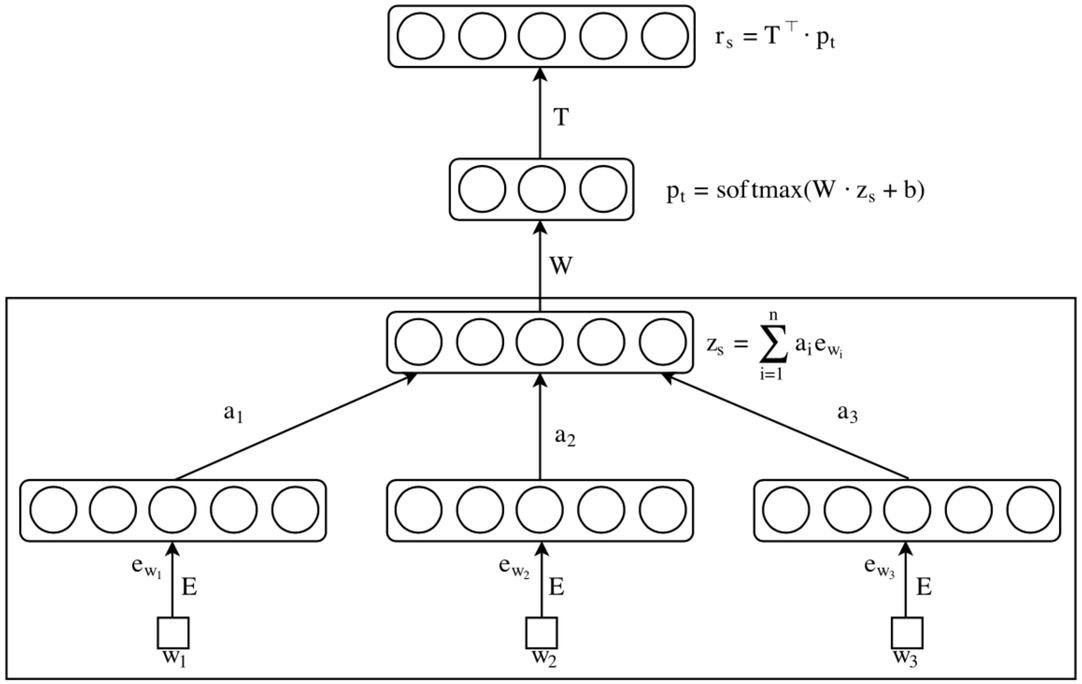
$$d_i = e_{wi}^T \cdot M \cdot y_s \qquad a_i = \frac{exp(d_i)}{\sum_{j=1}^{n} exp(d_j)}$$

2. Уменьшение размерности с применением линейного слоя + softmax

$$p_t = softmax(W \cdot z_s + b)$$

3. Реконструкция вектора предложения как линейную комбинацию векторов из матрицы представлений аспектов

$$r_s = T^T \cdot p_t$$



Контрастивное внимание

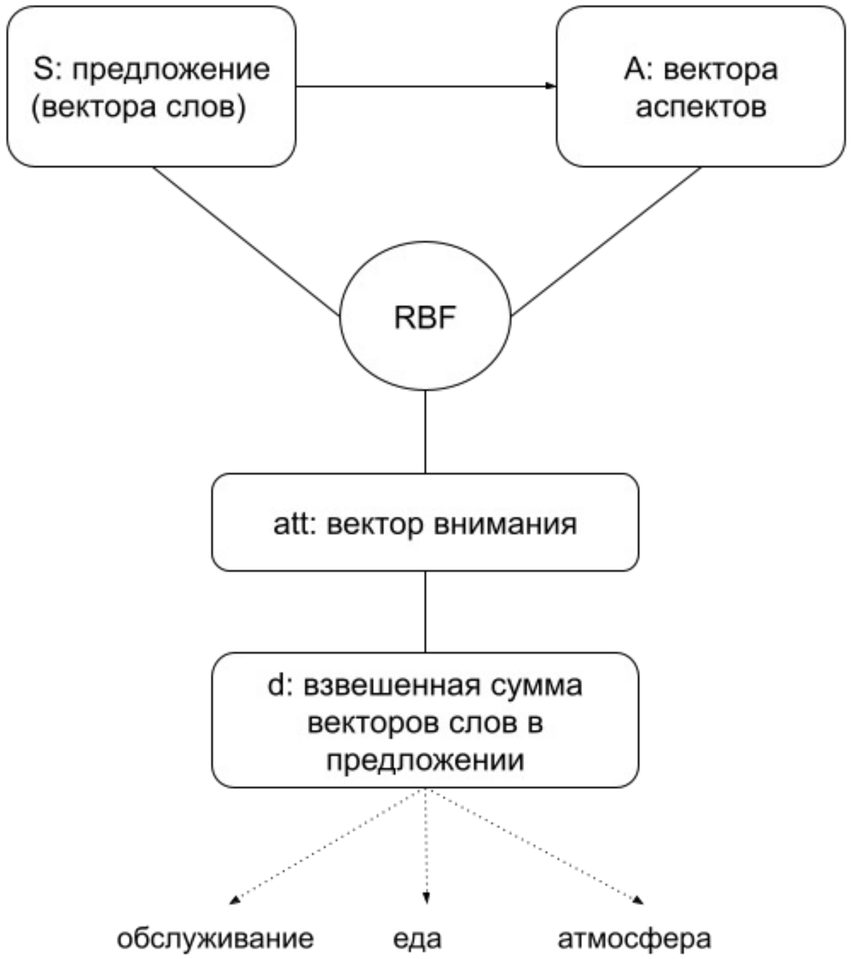
Идея: использование контрастивного внимания и ядро радиальной базисной функции (RBF)

Вход: аспектные категории, порог отсечения кандидатов в аспектные термины, коэффициент масштабирования

Основные шаги:

1. Обучение векторной модели на больших коллекциях мнений заданной сущности
2. Извлечение аспектных терминов как наиболее часто встречаемых существительных
3. Использование контрастивного внимания для получения взвешенной суммы слов в предложении

4. Маркировка предложения аспектом. Вычисляется косинусное сходство между вектором аспекта и вектором предложения



Контрастивное внимание

Модель на основе контрастивного внимания дает единое распределение внимания, которое вычисляется следующим образом:

$$att = \frac{\sum_{a \in A} rbf(w, a, \gamma)}{\sum_{w \in S} \sum_{a \in A} rbf(w, a, \gamma)}$$

Ядро радиальной базисной функции (RBF) вычисляет расстояние (близость) между векторами **x** и **y**:

$$rbf(x, y, z) = exp(-\gamma ||x - y||_2^2)$$

Связь аспекта и предложения определяется на основе косинусного сходства:

$$\hat{y} = \underset{c \in C}{\operatorname{argmin}}(cos(d, \vec{c}))$$