УДК 004.89

МЕТОДЫ ОПРЕДЕЛЕНИЯ СТРУКТУРЫ И ОСНОВНЫХ ТЕМАТИК ПОЛЬЗОВАТЕЛЬСКИХ ДИАЛОГОВ

Фельдина Е.А.
(Университет ИТМО)
Научный руководитель – к. т. н. Махныткина О.В.
(Университет ИТМО)

Данная работа посвящена теме определения структуры и основных тематик пользовательских диалогов. Рассмотрены этапы предварительной обработки, алгоритмы и методы решения задачи кластеризации и определения тем кластеров из диалогов пользователей с операторами контакт-центров на русском языке.

Исследования выполнены за счет финансирования университета ИТМО в рамках НИР № 619423 «Разработка виртуального диалогового помощника для поддержки проведения дистанционного экзамена на основе аргументационного подхода и глубокого машинного обучения».

Одним из требований для узкоспециализированных диалоговых систем является высокая точность в определении тематики. Для создания автоматизированного консультанта необходимо определить тематики, по которым будет отвечать виртуальный консультант. В большинстве случаев выделение тематик производится на основе ручной разметки, что занимает много времени, ресурсов и требует формирование методики проставления меток и обучение группы, осуществляющей разметку. Для решения описанных проблем необходимо создать инструмент, позволяющий сформировать структуру тематик на основе пользовательских диалогов. Основные требования к такому инструменту - независимость от предметной области, время выполнения задачи и легкость в использовании.

Существуют различные методы моделирования темы в зависимости от целей исследования, уникальности семантики рассматриваемых источников данных, наличия размеченных данных. Часто сообщения из социальных сетей, новостей и веб-страниц используются в качестве источника данных. Эта информация доступна и позволяет протестировать предложенные методы моделирования темы. Большинство подходов тематического моделирования включают в себя предварительную обработку текста, которая создает многомерное векторное пространство. Например, используя предварительно обученные вектора слов, нечеткий пакет слов (FBoW), который отображает каждый документ в нечеткий вектор фиксированной длины базовых терминов, динамическую модель встроенных тем (D-ETM).

После этапа предварительной обработки текста для определения структуры используются методы кластеризации. Существуют различные алгоритмы кластеризации:

- агломеративная кластеризация;
- DBSCAN;
- k-means.

Для каждого кластера в сформированной структуре необходимо определить основную тему. Среди подходов к решению проблемы тематического моделирования выделяются следующие методы:

- байесовская непараметрическая тематическая модель;
- вероятностный метод (LDA);
- логистический LDA (logistic LDA);
- сравнительное скрытое распределение Дирихле (CompareLDA);
- тематическая модель для иерархических документов (hdLDA) для захвата иерархической структуры текстов;
- стохастическая блочная модель (SBM) с непараметрическими априорами.

В данной работе были рассмотрены методы предварительной обработки текста, алгоритмы кластеризации для определения структуры тем и методы тематического моделирования

для выделения смысловых аннотаций кластеров. В дальнейшей работе планируется разработка алгоритма поиска оптимального количества кластеров на одном уровне и оптимального количества уровней.