Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

ВЫПУСКНАЯ КВАЛИФИКАЦИОННАЯ РАБОТА

РАЗРАБОТКА АЛГОРИТМОВ РАБОТЫ С ФОРМАЛЬНОЙ МОДЕЛЬЮ ДИАЛОГОВ, ПРЕДСТАВЛЕННЫХ В ВИДЕ ГРАФОВ

Автор: Савон Юлия Константиновна _____

Направление подготовки: 01.03.02 Прикладная математика и информатика

Квалификация: Бакалавр

Руководитель: Ульянцев В.И., доцент, к.т.н. _____

К защите допустить

Руководитель ОП Парфенов В.Г., проф., д.т.н. _____

«____» _____ 20___ г.

Санкт-Петербург, 2020 г.

Студент Савон Ю.К. Группа М3437 Факультет ИТиП

Направленность (профиль), специализация
Математические модели и алгоритмы в разработке программного обеспечения

Консультанты:

а) Ступаков И.М., канд. тех. наук, доцент _____

ВКР принята «____» _____ 20__ г.

Оригинальность ВКР ____%

ВКР выполнена с оценкой _____

Дата защиты «____» _____ 20__ г.

Секретарь ГЭК Павлова О.Н. _____

Листов хранения _____

Демонстрационных материалов/Чертежей хранения _____

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

УТВЕРЖДАЮ

Руководитель ОП

проф., д.т.н. Парфенов В.Г. _____

«____» _____ 20__ г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Студент Савон Ю.К. **Группа** М3437 **Факультет** ИТиП

Руководитель Ульянцев В.И., доцент, к.т.н., доцент факультета информационных технологий и программирования Университета ИТМО

1 Наименование темы: Разработка алгоритмов работы с формальной моделью диалогов, представленных в виде графов

Направление подготовки (специальность): 01.03.02 Прикладная математика и информатика

Направленность (профиль): Математические модели и алгоритмы в разработке программного обеспечения

Квалификация: Бакалавр

2 Срок сдачи студентом законченной работы: «10» июня 2020 г.

3 Техническое задание и исходные данные к работе

Требуется провести исследование и разработать набор алгоритмов для выявления отвлечений в графовой модели для робота-оператора. Алгоритм принимает набор диалогов, в размере нескольких тысяч. Предварительно выстроенный граф и кластеризацию для фраз оператора.

На выходе ожидается получить набор отвлечений и перестроенный граф. В качестве метрики качества будет использоваться сравнение с уже существующими графами, которые создавались вручную.

4 Содержание выпускной работы (перечень подлежащих разработке вопросов)

Пояснительная записка должна описывать предметную область диалогов представленных в виде графов. Так же формулировать цель и задачу выделения отвлечений, содержать описание алгоритмов их поиска. Должны быть описаны сложности и методы их разрешения, если они возникали. Кроме того должны быть приведены примеры работы алгоритмов и сравнение с существующими решениями. Кроме того

пояснительная записка должна содержать описания задач из смежных областей и их то, как эти задачи связны с задачей решаемой в работе.

5 Перечень графического материала (с указанием обязательного материала)

Графические материалы и чертежи работой не предусмотрены

6 Исходные материалы и пособия

а) Среда разработки Visial Studio Code;

б) ГОСТ 7.32–2001 «Система стандартов по информации, библиотечному и издательскому делу. Отчет о научно-исследовательской работе. Структура и правила оформления».

7 Дата выдачи задания «01» сентября 2019 г.

Руководитель ВКР _____

Задание принял к исполнению _____

«01» сентября 2019 г.

Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«САНКТ-ПЕТЕРБУРГСКИЙ НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИНФОРМАЦИОННЫХ ТЕХНОЛОГИЙ, МЕХАНИКИ И ОПТИКИ»

АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Студент: Савон Юлия Константиновна

Наименование темы ВКР: Разработка алгоритмов работы с формальной моделью диалогов, представленных в виде графов

Наименование организации, в которой выполнена ВКР: Университет ИТМО

ХАРАКТЕРИСТИКА ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

1 Цель исследования: Разработа алгоритма выделяющего отвлечения в диалоге, представленном в виде графа.

2 Задачи, решаемые в ВКР:

а) Разработать алгоритмы выделения отвлечий;

б) Реализовать описанные алгоритмы;

в) Перестроить граф в соответствии с используемой моделью в компании;

г) Проанализировать результаты работы алгоритмов;

д) Интегрировать разработки в инфраструктуру компании.

3 Число источников, использованных при составлении обзора: 0

4 Полное число источников, использованных в работе: 0

5 В том числе источников по годам:

Отечественных			**Иностранных**		
Последние 5 лет	От 5 до 10 лет	Более 10 лет	Последние 5 лет	От 5 до 10 лет	Более 10 лет
0	0	0	0	0	0

6 Использование информационных ресурсов Internet: нет

7 Использование современных пакетов компьютерных программ и технологий:

Пакеты компьютерных программ и технологий	**Раздел работы**
Пакет `tabularx` для чуть более продвинутых таблиц	**??**, Приложения **??**, **??**
Пакет `biblatex` и программное средство `biber`	Список использованных источников

8 Краткая характеристика полученных результатов

9 Гранты, полученные при выполнении работы

10 Наличие публикаций и выступлений на конференциях по теме работы

По теме этой работы был сделан доклад на Конгрессе Молодых Ученых.

Студент Савон Ю.К. _____

Руководитель Ульянцев В.И. _____

«____» _____ 20__ г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

На текущий момент времени всё больший процент работы телефонных операторов автоматизируется. В связи с этим возникает потребность в создании роботов, которые бы не просто начитывали текст, а взаимодействовали с человеком, могли отвечать на его вопросы, записывали бы полученную информацию и в целом, были бы неотличимы от человека.

Поскольку оператор обычно звонит с заданной целью, у него есть некоторый сценарий.

Сценарий диалога с оператором — некоторый алгоритм, предоставленный человеку, который звонит по заданному набору телефонных номеров. Целью сценария обычно является получить или донести до клиента информацию.

Между скриптом[1], который дан оператору, и алгоритмом действий для машины есть большая разница. Человеку достаточно задать набор вопросов и дать информацию о возможных продуктах и услугах; он сам умеет обрабатывать ситуации такие как отвлечение от основных вопросов или переспрашивание. Для графа же, любую реакцию надо прописывать.

Робот-оператор - программа, которая используя сценарий умеет проводить диалог с клиентом, интерпретировать и записывать информацию полученную от клиента, а так же состояния завершенного разговора. Кроме того робот умеет отвечать на заранее прописанный в скрипте набор вопросов и возвращаться обратно к диалогу.

На данный момент существуют графы для диалогов, но писать их долго, продумать все случаи реакций практически невозможно, поэтому ставится глобальная задача превратить набор диалогов в граф.

Поскольку количество возможных реакций из каждой вершины может быть бесконечным, то было решено отделить такие подграфы и сделать возможность переходить в них при некоторых условиях из каждой вершины.

В работе будут рассмотрены различные алгоритмы поиска таких отвлечений. А так же рассмотренны варианты улучшения для кластеризации, использующие данные о контексте.

---

[1]Здесь и далее в тексте **скрипт** и **сценарий** будут использоваться как синонимы

ГЛАВА 1. ОБЗОР

1.1. Описание предметной области

Изначальная общая задача стоит следующая: есть диалоги, необходимо восстановить граф и найти отвлечения.

На данный момент кроме восстанавливаемой нами модели есть граф который составляется вручную. В этом графе вершинами являются кластеры фраз робота-оператора, а рёбрами фразы человека. То есть при ответах человека происходит переход в разные новые вершины.

Особенностями данной структуры, которые важно отметить, является то, что помимо обычных переходов существуют так же скрытые переходы, которые по умолчанию могут встретиться в любом месте. В качестве примера можно привести вопрос: "А какую компанию вы представляете". При звонке люди могут задать этот вопрос не сразу.

Для того, чтобы не рассматривать каждый такой случай при переходе из каждой вершины, выделяются **отвлечения**.

Отвлечение - вопрос, который может быть задан в любом месте диалога.

В задаче восстанавления будет использоваться несколько другой граф. В нём вершинами будут являться кластеры и фраз оператора, и фраз человека. А ребрами последовательные пары фраз из кластеров в диалоге.

1.2. Анализ

На вход подаётся набор диалогов по которым нужно получить граф для робота, который бы мог проводить аналогичные диалоги.

Граф, если получать его путём обычной кластеризации операторских фраз, получается очень громоздким. В нём плохо видно структуру, его сложно анализировать.

Поскольку конечной целью является построить автомат аналогичный тому, что работает в продуктовой части команды, то появляется необходимость уметь изучать его руками.

То есть так же как и в графе создаваемом вручную, появляется необходимость выделить отвлечения. В этом случае его структура становится более удобной для изучения. Его можно использовать как вспомогательный инструмент.

Такие выделения отвлечений позволяют грамотно обрабатывать сценарии, которых не было в изначальном наборе диалогов. Если такие отвлечения не выделить, то в случае вопроса, не предусмотренного в этом месте робот либо ответит невпопад, либо зависнет.

Необходима хорошая кластеризация текстов, поскольку любой алгоритм выделения отвлечений так или иначе будет опираться на эту кластеризацию. При таких кластеризациях полезно учитывать имеющуюся информацию, то есть фразы до текущей и после.

1.3. Постановка задачи

В связи с описаной выше проблемой громоздкости графа, появляется задача выделить отвлечения, которые могли бы встретиться в любом месте.

В случае, когда диалог проводит оператор такие вопросы нет необходимости расписывать, но роботу нужен чёткий скрипт.

Необходимо выделить отвлечения и перестроить граф таким образом, чтобы он покрывал большее количество диалогов и его было проще анализировать.

1.4. Выводы по первой главе

Рассмотрена предметная область графовой структуры диалогов. Разобрана структура продуктового графа и структура графа для восстанавливаемой модели. Проведён анализ задачи восстановления графа. Поставлена задача выделения отвлечений.

ГЛАВА 2. ТЕОРЕТИЧЕСКАЯ ЧАСТЬ

2.1. Модель предлагаемого решения

Поскольку граф на данном этапе всё ещё перестраивается и информация о произнесенных человеком фразах является полезной, так как содержит в себе контекст. То мы оставляем фразы человека в качестве вершин.

Нужно понимать, что изначально кластеризация проводилась только по фразам оператора. Фразы вершин же были разделены на группы, где для каждой группы совпадала предыдущая и последующая вершины оператора. И уже внутри этих групп бились на некоторые подгруппы.

Рассмотрим два подхода в решении задачи выделения отвлечений:

— Первый подход заключается в том, чтобы выделить вершины, в которые идёт много рёбер. Порог считается функцией от количества кластеров на которые бъются фразы оператора.

  Мы предполагаем, что поскольку отвлечение встречается в разных местах, то и рёбра будут идти в него из множества различных вершин. Такая гипотеза является хорошей, поскольку в обычном графе в вершину обычно приходит одна или две ветки, в случае же отвлечения их должно быть много, или же оно встречается крайне редко.

— В качестве другого подхода можно выделить циклы и сказать, что вершина следующая в диалоге за вершиной повторения возможно будет являться началом отвечения.

  Здесь мы пользуемся наблюдением, что после отвлечения на стронний вопрос, оператор зачастую повторяет ту же или схожую фразу для возвращения в сценарий. Более того, в графе, который используется в реальном окружении робот-оператор так же повторает аналог той фразы, которую произносил, перед тем, как уйти в отвлечение.

2.2. Соответствие поставленным требованиям

Такие алгоритмы позволят найти для диалогов которые были проведены с роботом отвлечения, сопоставить их с выделенными лингвистами и сопоставить результат.

ГЛАВА 3. ПОЛУЧЕННЫЕ РЕЗУЛЬТАТЫ

3.1. Реализация

Выбираются вершины, в которые входит много ребер. Для значений размеров кластеров в интервале от двенадцати до пятнадцати был выбран параметр три. При увеличении количества кластеров соответственно должен увеличиваться и порог.

Во время тестирования на реальных диалогов человека с человеком была выявлена важная особенность. На работу алгоритма очень сильно влияет качество кластеризации. Поскольку фразы в телефонных диалогах не всегда верно переводятся в текст, сами фразы сравнительно короткие. А в случае людей-операторов ещё и очень вариативные, то кластеризация оказалась очень некачественной.

3.2. Улучение кластеризации

В качестве решения проблемы описанной выше было решено использовать данные из фраз человека. До этого для всех фраз человека между двумя фразами оператора, они кластеризировались и никак не использовались.

Было решено кластеризировать тексты пользователей. Но поскольку как описывалось выше, кластеризация не достаточно хорошая, то необходимо было отсеять плохие кластеры во избежание каскадных ошибок.

Для этого использовалось попарное сравнение фраз внутри каждого кластера. Для этого сравнивались наборы слов внутри фразы с весами. Если точность превышала порог 0.5, то пара считалась хорошей.

Поскольку проверять все пары оказалось очень долго, то для каждого кластера бралась случайная выборка, равная утроенному размеру кластера. Это, в силу достаточно больших размеров кластеров, статистически показывало те же результаты, что и полная выборка.