Министерство науки и высшего образования Российской Федерации

ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ АВТОНОМНОЕ ОБРАЗОВАТЕЛЬНОЕ УЧРЕЖДЕНИЕ ВЫСШЕГО ОБРАЗОВАНИЯ

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

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 ВКР: Ульянцев В.И., к.т.н. _____

Санкт-Петербург, 2020 г.

Обучающийся Савон Юлия Константиновна
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 «26» июня 2020 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

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

УТВЕРЖДАЮ

Руководитель ОП

проф., д.т.н. Парфенов В.Г. _____

«____» _____ 20___ г.

ЗАДАНИЕ
НА ВЫПУСКНУЮ КВАЛИФИКАЦИОННУЮ РАБОТУ

Обучающийся Савон Юлия Константиновна

Группа М3437 **Факультет** ИТиП

Квалификация: Бакалавр

Направление подготовки (специальность): 01.03.02 Прикладная математика и информатика

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

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

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

2 Срок сдачи студентом законченной работы до: «20» 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 голосовой диалоговой системы. 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 квалификационной 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

«НАЦИОНАЛЬНЫЙ ИССЛЕДОВАТЕЛЬСКИЙ УНИВЕРСИТЕТ ИТМО»

АННОТАЦИЯ
ВЫПУСКНОЙ КВАЛИФИКАЦИОННОЙ РАБОТЫ

Обучающийся: Савон Юлия Константиновна

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

Наименование организации, где 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: 11

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	2	1	2	2	4

6 Использование информационных ресурсов Internet: да, число ресурсов: 1

7 Использование современных пакетов компьютерных программ и технологий:

Пакеты компьютерных программ и технологий	**Раздел работы**
Интегрированная среда разработки `PyCharm`	Глава 2.2
Программное обеспечение для автоматизации развертывания и управления приложениями в средах с поддержкой контейнеризации `Docker`	
Распределённая система контроля версий `Git`	Глава 2.2

8 Краткая характеристика полученных результатов

9 Гранты, полученные при выполнении работы

10 Наличие публикаций и выступлений на конференциях по теме выпускной работы

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

Обучающийся Савон Ю.К. _____

Руководитель ВКР Ульянцев В.И. _____

«_____» _____ 20___ г.

СОДЕРЖАНИЕ

ВВЕДЕНИЕ

Эта работа посвящена исследованию в области диалоговых систем.

Диалоговая система – это алгоритм, который умеет принимать участие в диалоге на естественном языке и использует правила общения между людьми.

В качестве примера диалоговых систем можно привести:

— Чат-ботов
— Голосовых помощников
— Автоответчиков в колл-центрах

Такие диалоговые системы могут быть как довольно простыми (например, чат-бот отвечающий на заранее известный набор команд), так и сложными (бот, отвечающий на вопрос, сформулированный на естественном языке и возвращающий некоторую информацию из базы знаний).

В последнее время набрали популярность технологии распознования и генерации речи, которые позволили создавать диалоговые системы, ведущие голосовой разговор. Данная работа рассматривает алгоритмы именно для голосовых диалогов. Пример решения такой диалоговой системы можно рассмотреть в статье [10].

Такие звонки, с одной стороны, должны быть не отличимы от звонков человека, с другой – они должны придерживаться некоторого сценария.

Сценарий диалога с оператором – некоторый алгоритм, предоставленный человеку, который звонит по некоторому набору номеров (либо же принимает входящий звонок или общается посредством программного обеспечения для аудиосвязи). Целью сценария обычно является получение или донесение до клиента информации.

Несмотря на то, что под сценарием диалога понимается некоторый алгоритм, необходимо понимать что для человека и диалоговой системы это принципиально разные сущности. Между скриптом[1] и алгоритмом, с точки зрения набора действий для машины есть большая разница.

Для человека это скорее структурированный список вопросов которые он должен задать и некоторая дополнительная база знаний с ответами на нестандартные вопросы. Кроме того человек может помнить некоторые фак-

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

ты и выдавать их дополнительно в зависимости от контекста. Он сам умеет обрабатывать ситуации, такие как отвлечение от основных вопросов или переспрашивание.

Для скрипта же, реакцию на любую фразу человека надо прописывать, все возможные данные хранить и обновлять. Кроме того, есть требование поддерживать этот скрипт доступным для восприятия человеком (например лингвистом), поскольку возникает необходимость в ручном анализе и редактировании. В данном случае такой скрипт будет представлен в виде графа с дополнительной информацией.

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

На данный момент существуют графы для диалогов, которые создаются вручную. Но писать их долго, а продумывать все важные случаи реакций сложно и трудозатратно.

Помимо этого, хочется иметь возможность усложнять вариативность диалогов, делая их более похожими на процесс общения двух людей. В связи с этим, ставится глобальная задача по созданию графа из набора проведенных диалогов.

Достаточно часто возникает ситуация, когда некоторый общий вопрос может возникнуть в любом месте диалога (например, «А что это за компания?»). Таким образом, было решено отделить их в отдельные подграфы и сделать возможность переходить в них при некоторых условиях из каждой вершины. В дальнейшем мы будем называть такие случаи **отвлечениями**.

В работе будут рассмотрены различные алгоритмы автоматического поиска таких отвлечений.

ГЛАВА 1. ОБЗОР ДИАЛОГОВОЙ МОДЕЛИ

1.1. Разговорная диалоговая система

В данной работе будут рассматриваться интелектуальные диалоговые системы. Введем несколько определений.

Разговорная диалоговая система – система, позволяющая пользователю-человеку получать доступ к информации и сервисам доступным на компьютере или в сети Интернет, используя разговорный язык как средство взаимодействия, согласно [7].

Разговорные диалоговые системы используют распознавание речи для преобразования фраз человека в формат, удобный для использования диалоговым менеджером.

1.2. Диалоговый менеджер

Диалоговый менеджер – компонент диалоговой, системы ответственный за текущее состояние и ход диалога, согласно[9]. В основном, диалоговые менеджеры оперируют предобработанными текстами. У диалогового менеджера так же есть состояние, которое может хранить в себе например последний неотвеченный вопрос или историю диалога.

Диалоговый менеджер в том числе можно представить в виде автомата, пример которого можно увидеть на Рисунке 1:

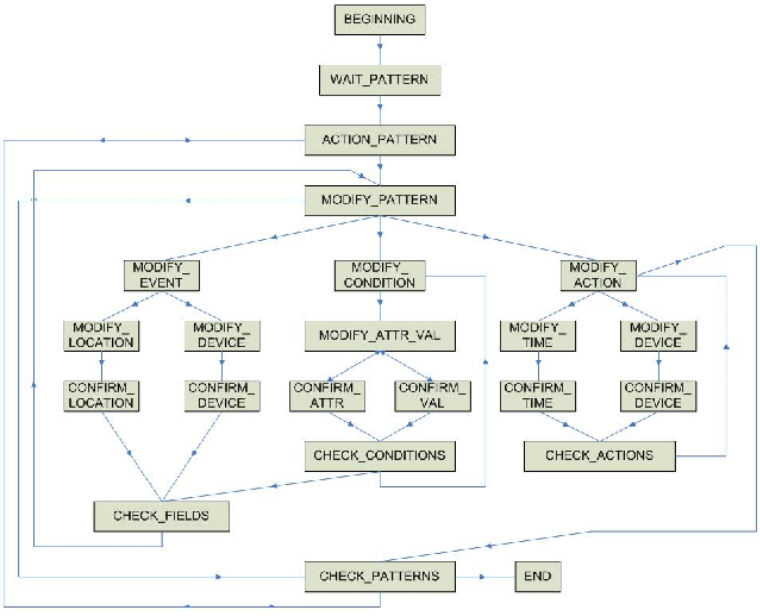

Рисунок 1 – Пример автомата для диалогового менеджера

В этом примере, в зависимости от текущего состояния и информации, предоставленной пользователем – система пытается принять решение о том, какое действие хочет выполнить пользователь. Эта система использует не только текущий ответ, а так же информацию о контексте. Кроме того, если системе не хватает информации, она запрашивает её у пользователя. Рисунок и описание взяты из [4].

Основные способы создания моделей диалоговых менеджеров:

— Составление вручную.
— Генерация моделии менеджера с помощью машинного обучения.

Плюсами первого подхода можно назвать прозрачность процесса перевода имеющегося скрипта в модель. Серьёзным минусом создания модели диалогового менеджера вручную является необходимость каждый раз полностью создавать структуру диалога, что очень ресурсозатратно.

Плюсом генерации модели менеджера с помощью машинного обучения является невысокая трудозатратность на каждую новую систему. Минусом такого подхода можно назвать необходимость в большом количестве данных для обучения и зависимости от качества исходных данных.

Пример модели диалогового менеджера, улучшенного с помощью методов машинного обучения представлен в статье [6]. Данное улучшение позволило организовать управление инвалидным креслом с помощью голоса, посредством инструкций более похожих на естественный язык.

На данном этапе, в компании используются диалоговые менеджеры, составленные вручную. Создание одного такого диалогового менеджера может требовать много ресурсов, которые при таком подходе нельзя значительно соптимизировать.

В этой работе рассматривается построение на основе графов.

Глобальная задача состоит в следующем: необходимо по набору диалогов восстановить диалоговую систему.

Диалоговая система в данном случае представлена в виде графа, но так же существуют и другие представления.

На данный момент, кроме восстанавливаемой нами модели – разработаны графовые модели, которые составляются вручную. В упрощенном виде, в этих моделях вершинами являются фразы диалоговой системы, а рёбра ассо-

циированы с группами возможных ответов человека. То есть, при различных ответах человека происходит переход в разные вершины.

Особенностью данной структуры, которую 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 – это вопрос, который может быть задан в любом месте диалога. Кроме того, к одному из подтипу отвлечений относятся случаи, когда оператор не услышал фразу или вопрос и вынужден переспросить. В зависимости от модели, это может выделяться в специальную сущность или являться отвлечением.

В задаче восстанавления графа по набору диалогов используется видоизменённая модель. В ней вершинами являются, как кластеры фраз оператора, так и кластеры фраз человека. А ребро соответствует наличию перехода между соответсвующими кластерами в диалоге.

Кластеризация текстов (фраз) – автоматическая группировка текстовых документов, например веб-страниц, электронных писем, человеческих фраз, основанная на схожести содержимого. В качестве входных данных алгоритмы принимают набор фраз и количество желаемых кластеров и выдают сгруппированные в соответствующее количество – кластеры $K_1, K_2, ...$ – [8].

1.3. Анализ задачи выделения отвлечений

На вход подаётся набор диалогов, по которым нужно получить граф для диалоговой системы, которая бы могла 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 требование привести его в состояние, когда его можно изучать вручную.

То есть, так же как и в графе создаваемом вручную, появляется необходимость выделять отвлечения. В этом случае, его структура становится более

удобной для анализа. Алгоритм выделения, в частности, позволяет анализировать особенности графа непосредственно во время процесса выделения отвлечений.

Как следствие, особенности структуры 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 диалоговая система либо даст некорректный ответ, либо несвоевременно закончит работу.

Хорошая кластеризация текстов является важным предварительным шагом для алгоритма, поскольку любой алгоритм выделения отвлечений непосредственно опирается на информацию, полученную в результате кластеризации. Полезно, при этом учитывать имеющуюся информацию о контексте, то есть фразы до текущей и после.

1.4. Постановка задачи

Целью данной работы является реализация возможности поиска отвлечений в алгоритме восстановления графа по набору диалогов.

Отвлечения могут встретиться в любом месте, таким образом после выделения и перестроения графа, граф будет более структурирован, что позволит строить более сложные конструкции для диалогов, которые всё ещё будут поддаваться анализу и контролю вручную.

Если провести аналогию для человека-оператора, то в его скриптах нет необходимости расписывать вопросы-отвлечения. Человек сам прекрасно понимает когда ответ отклоняет его от скрипта. Но в отличие от оператора, диалоговой системе нужны однозначные инструкции, которые бы покрывали большинство возможных типов вопросов.

Необходимо выделить отвлечения и перестроить граф таким образом, чтобы отвлечения выделились в отдельный подграф, куда можно было бы перейти из любого места диалога. Такой граф будет покрывать большее количество диалогов.

1.5. Выводы по первой главе

Рассмотрена предметная область диалоговых систем, введены определения диалоговой системы, менеджера диалогов, отвлечения. Описаны подходы к построению менеджеров диалогов с соответствующими достоинствами и

недостатками. Так же описано предствление диалоговой системы в виде графа. Разобрана структура графа, используемого в продуктовом окружении, и структура графа для восстанавливаемой модели. Проведён анализ задачи восстановления графа. Поставлена задача выделения отвлечений.

ГЛАВА 2. ОПИСАНИЕ АЛГОРИТМОВ ПОИСКА ОТВЛЕЧЕНИЙ И РЕКЛАСТЕРИЗАЦИИ

2.1. Процесс разработки

Восстановление графа делается поэтапно, в этот процесс включены следующие этапы в соответсвующем порядке:

— Распознавание аудио.
— Преобразование аудио в тексты.
— Препроцессинг в виде преобразования и исправления текстов.
— Кластеризация и восстановление графа.
— Выделение отвлечений.
— Конвертация графа для возможности запуска в существующей системе.

Данная работа не будет затрагивать первые 2 этапа.

В процессе разработки были сделаны улучшения для третьего этапа и затронут четвёртый. Основная работа касается этапа выделения отвлечений, кроме того написаны алгоритмы для преобразования, что даёт возможность сравнить графы с существующими.

Важным будет отметить, что поиск отвлечений встроен в алгоритм построения графа и таким образом появляется возможность внести некоторые изменения в его структуру. Одним из таких изменений может быть выделение отвлечений в отдельный подграф, что позволит рассматривать части независимо.

2.2. Задача выделения отвлечений

Поскольку на этапе выделения отвлечений в графе его перестроение всё ещё возможно, а информация о произнесенных человеком фразах является ценной, так как содержит в себе контекст, то фразы человека оставлены 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 бьются 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 с некоторой вероятностью 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, эта идея используется в графе, который реализован для реального окружения. Там диалоговая система так же повторяет фразу, её сокращенную версию или её иную формулировку, которую произносил, перед тем, как перейти в отвлечение.

Поскольку подходы используют разные идеи их так же имеет смысл комбинировать и использовать данные полученные в обоих подходах.

2.3. Подход поиска отвлечений с большим количеством рёбер входящих в вершину

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.

В некоторых случаях отвлечения не ограничиваются одной вершиной. В связи с этим появляется необходимость определить, где заканчивается то или иное отвлечение, и где соответственно оператору необходимо вернуться в вершину с которой он в это отвлечение ушёл. Для этого предположим следующую гипотезу – если в вершину мы можем попасть только пройдя через отвлечение, то это значит, что фраза является частью соответствующего отвлечения.

Для того, чтобы найти соответствующий хвост для отвлечения мы используем следующий алгоритм: заблокируем вершину и рассмотрим все вер-

шины, достижимые из вершины старта. Важно так же помнить о возможности того, что в кластерах могут быть допущены небольшие ошибки, поэтому если почти весь кластер недостижим после блокировки, то он так же входит в отвлечение.

Ниже представлен псевдокод функции поиска таких хвостов для отвлечений:

function DIGRESSION_FINDER($graph, node, dialogs$)
  for $dialog \in dialogs$ **do**
    $flag \leftarrow True$
    for $msg \in dialog$ **do**
      $msgs[msg.cluster] \leftarrow msgs[msg.cluster] + 1$
      if $(msg.cluster == node.cluster) \wedge flag$ **then**
        $achieved.add(msg.cluster)$
        $achieved\_msgs[msg.cluster].add(msg)$
      else
        $flag \leftarrow False$
      end if
    end for
  end for
  for $cluster \in graph.clusters$ **do**
    if $clusternotinachieved$ **then**
      $digression\_clusters.add(cluster)$
    else
      if $\frac{achieved\_msgs[cluster]}{msgs[cluster]} > 0.9$ **then**
        $digression\_clusters.add(cluster)$
      end if
    end if
  end for
  return $digression\_clusters$
end function

Во время тестирования на реальных диалогов человека с человеком была выявлена важная особенность. На работу алгоритма очень сильно влияет качество кластеризации. Поскольку фразы в голосовых 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.

2.4. Подход поиска отвлечений с поиском циклов

Поскольку говорит на сторонние темы только человек, а оператор идёт по скрипту, то в подавляющем большинстве случаев после ответа на сторонний вопрос, оператор задаёт свой вопрос заново. В связи со спецификой человеческий диалогов можно выделить несколько полезных особенностей:

— Подавляющее большинство поддиалогов отвлечений достаточно короткие.
— Вопрос повторяет оператор, поэтому циклы можно искать с повторением только операторской вершины.
— Возможно появление отвлечения внутри отвлечения, поэтому необходимо найти оба.

Для решения особенности первого пункта было добавлено ограничение на расстояние между ними в диалоге. Оно должно быть не слишком велико, поскольку ясно что отвлечение бывает длинным очень редко, а иначе можно случайно выкинуть почти весь диалог.

Проблема последнего пункта решается тем, что мы удаляем циклы по мере нахождения. Таким образом внутренний цикл будет удалён раньше внешнего.

В этом случае в качестве потенциальных отвлечений мы выбираем вершины человеческих фраз которые первые следуют после начала цикла. После перевода графа в продуктовый режим у нас эти вершины станут рёбрами и таким образом в графе появятся рёбра-триггеры, которые будут перенаправлять ход диалога в вершину-отвлечение.

Ниже находится псевдокод для поиска одного цикла:

function REMOVE_CYCLES($dialog$)
    for $msg \in dialog$ **do**
        $msg\_num \leftarrow msg\_num + 1$
        if $(last\_msg[msg.cluster] - msg\_num) <= 5$ **then**
            $cycle\_end \leftarrow last\_msg[msg.cluster]$
            $dialog.remove\_cluster(cycle\_start, msg\_num)$
        end if

end for
end function

2.5. Выборка хороших кластеров

В качестве решения проблемы некачественной кластеризации было предложено 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 недостаточно 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 равный 0.5, то такая пара считалась хорошей.

Константа 0.5 была выведена эмпирически. Для большего значения в кластере начинало содержаться большое количество фраз разных по смыслу.

Проверка всех пар фраз имеет асимптотику по времени $O(n^2)$, где n – количество фраз. Так как в предыдущей части восстановления все операции имели ассимтотику не более чем $O(n \log n)$, то получилось так, что эта часть занимала значительно больше половины времени от всего восстановления графа.

Было решено для каждого кластера брать случайную выборку, равная утроенному размеру кластера, асимптотически это занимало уже $O(n)$. В силу достаточно больших размеров кластеров (размер их для основной массы данных составляет несколько сотен фраз), статистически, это показывало те же результаты, что и перебор всех пар.

Утверждение 1. При размере диалога от 40 фраз необходимое количество экспериментов для попадания в доверительный интервал $Q = 0.95$ и доверительной вероятности[1] $\varepsilon = 0.1$ достаточно $3n$ экспериментов.

Доказательство. Значение $p = 0{,}5$ – наихудшее, в том смысле, что для него вероятность порогового отклонения превысит выбранное значение ε, при наибольшем количестве экспериментов. Таким образом достаточно доказать утверждение для него. Для такого случая результат равен 96, причём вне зависимости от размера возможных исходов, т.е. вне зависимости от размера кла-

[1]Определение доверительного интервала можно изучить здесь.[3]

стера. Эксперименты рассчитанные для разных значений приведены в учебнике.[1]

Таким образом, для более маленьких кластеров можно рассчитать эти значения перебрав все пары полностью, а для больших кластеров такого количества экспериментов будет достаточно.

Ниже приведён псевдокод для функции, считающей метрику для вершин с большим количеством фраз:

function IS_GOOD_CLUSTER($msgs$)
    $num\_of\_comparings \leftarrow 3 \cdot msgs.size()$
    for $(msg1, msg2) \in msgs.getPairs(num\_of\_comparings)$ **do**
        $good\_pairs+ = msg\_compare(msg1, msg2)$
    end for
    $cluster\_threshold \leftarrow 0.5$
    return $\frac{good\_pairs}{num\_of\_comparings} >= cluster\_threshold$
end function

Функция $msg\_compare(msg1, msg2)$ отвечает за сравнение двух сообщений.

2.6. Функция сравнения сообщений

Для алгоритма выше необходима функция сравнения двух сообщений. Для неё должны быть выполнены следующие условия:

— Функция должна быть бинарной. $True$ – в случае схожести сообщений, $False$ – иначе.
— Функция должна работать за $O(1)$, при условии, что длину сообщений мы так же принимаем за $O(1)$.

Второе условие необходимо, поскольку в противном случае при среднем размере групп сообщений от нескольких сотен разница во времени будет в разы увеличиваться.

В качестве алгоритма для сравнения сообщений был выбран следующий: берутся наборы слов в виде множеств и пересекаются с учётом весов слов. Это значение записываем в числитель. В качестве знаменателя берём объединение наборов слов с теми же коэффициентами слов и соответсвующими количественными коэффициентами и записываем в знаменатель.

Полученная дробь должна превышать некоторый порог, который ищется вручную.

В качестве альтернативных вариантов функций могут быть использованы следующие:

— Сравнение фраз на основе семантической близости. Об этом алгоритме можно прочитать в следующей статье.[2]
— Сравнение фраз с помощью эмбеддингов и на основе косинусного расстояния. Об этом подходе можно почитать здесь.[5]

Все перечисленные и описанные функции подходят под перечень изначальных условий, поскольку функцию возвращающую нецелые значния в диапазоне от 0 до 1 можно превратить в ступенчатую, сделав её бинарной. Все они работают за асимптотически необходимое время.

2.7. Слияние кластеров операторских сообщений

Изначальный алгоритм, который создавал вершины операторов предполагал точный выбор количества кластеров. Новое решение предполагает избыточное разбиение на кластеры. Поскольку количество кластеров на которые разделяются операторские сообщения можно контролировать вручную, это регулируется достаточно просто.

После такого разбиения выделятся кластеры, фразы в которых максимально похожи между собой, в таком разбиении будет меньше ошибок вида – в одном кластере содержательно разные сообщения.

Далее после выбора хороших человеческих вершин, для каждой операторской вершины рассматриваются все другие операторские вершины. Далее было несколько версий метрик для сравнения соседних вершин.

В первой версии рассматривались все соседские вершины. Они не делились группы по положению до или после и соответсвенно это было недостаточно хорошим решением.

Во второй версии были выбирались все вершины до, исходя из предположения, что они будут больше коррелировать с содержанием операторских вершин и не будет проблемы с тем, что не учтён порядок.

2.8. Оценка для сравнения графов

В качестве верной модели использовалась модель написанная вручную. Модель хранится в формате `JSON` и в ней фразы человека соответственно являются условиями перехода.

Для того, чтобы можно было сравнивать текущую модель и изначальную, необходимо было перевести вторую модель к формату первой.

Здесь стоит напомнить, что в восстанавливаемой модели кластеры человеческих фраз так же являлись вершинами. Таким образом для того, чтобы была возможность перевести их в рёбра и соответственно сопоставить изначальному графу, для каждой вершины с человеческими кластерами должны были быть верны следующие условия:

— У вершины должна быть ровно одна предыдущая, причём эта вершина должна быть вершиной операторских фраз.
— Последующая вершина должна быть вершиной обязана так же быть операторской и единственной для данной.
— Ни одна из набора фраз данного кластера не должна совпадать с фразами из кластеров, в которые ведут рёбра из предыдущей операторской вершины.

Реализация была сделана при помощи классов графа, новых вершин, рёбер. После проверки на то, что нет коллизий описанных выше можно провести инициализацию с помощью псевдокода приведенного ниже:

```
function CREATE_NODES(dialogs, digression)
    for dialog ∈ dialogs do
        for msg ∈ dialog do
            if msg.is_incoming then
                if msg.next then)
                    transactions[msg.prev.cluster].add(vertex[msg.next.cluster])
                else
                    end_vertexes.add(msg.prev.cluster)
                end if
            else
                vertex_msgs[msg.cluster].add(msg)
                vertex_clusters.add(msg.cluster)
            end if
        end for
    end for
    for clusterinvertex_clusters do
```

$vertexes.add(Vertex(transactions[cluster], vertex\_msgs[cluster], cluster \in$

end for

return $vertexes$

end function

Проверив, что все необходимые условия соблюдены, можно преобразовывать граф, выделяя отвлечения следующим образом:

— Каждая вершина новой модели сопостовлялась одной изначальной. Сопоставление проводилось путём сравнения произносимой фразы.

— Вершины фраз человека были выделены в группы, каждая группа относилась к одному ребру и являлась будущем набором исходящих рёбер.

— Технические вершины в изначальном графе игнорировались, но учитывалось их место в последовательности модели.

— Сравнивались наборы вершин отвлечения в изначальном графе и в полученном.

— При создании метрики для графа была использована статья [11]

ГЛАВА 3. РЕЗУЛЬТАТЫ ЭКСПЕРИМЕНТОВ

3.1. Используемые данные

В качестве данных использовалось два принципиально разных типа наборов диалогов. Первые проводились уже с существующим скриптом. Второй тип, это данные диалогов человека с человеком.

1.Особенность первого заключается в том, что там легко кластеризовать фразы оператора, так как они произносятся всегда одинаково, за исключением вариаций для коротких вариантов.

Но необходимо понимать, что случаи повторения фразы имеют другие вершины соседей поскольку расчитаны на специфичные случаи. Такие объединения были полезны для алгоритма поиска циклов, но находить их было сложно в силу того, что они явным образом не попадали под условия. Так же можно было выявить сликом заниженные коэффициенты, при слиянии фраз, которые не являлись одинаковыми.

Важным вариантом использования таких данных являлась возможность перевести полученный граф в тот же формат и сравнить полученный результат с оригиналом.

2.Особенности второго соответсвенно следующие: во первых там говорят разные операторы и у них разный стиль подачи одних и тех же данных. Во вторых диалоги более сложные и зачастую отходят от скрипта. В третьих люди решают более сложные вопросы и умеют давать ответы на не предусмотренные скриптом вопросы. На этот вариант данных стоит ориентироваться, но в связи с отсутствием оригинального скрипта в удобном формате напрямую сравнить полученный результат представляется возможным только вручную.

В обоих случаях есть сложности с переводом речи людей в текст, поэтому иногда даже рассматривая текст диалога вручную нельзя понять что человек имел ввиду.

Для всех датасетов первого типа соответсвенно существуют файлы, содержащие в себе оригинальный граф, который принимается за верный. Все данные сериализованы в формате `JSON`.

На Рисунке 2 представлен восстановленный граф, получившийся после одного из запусков. С этой моделью происходила непосредственная работа:

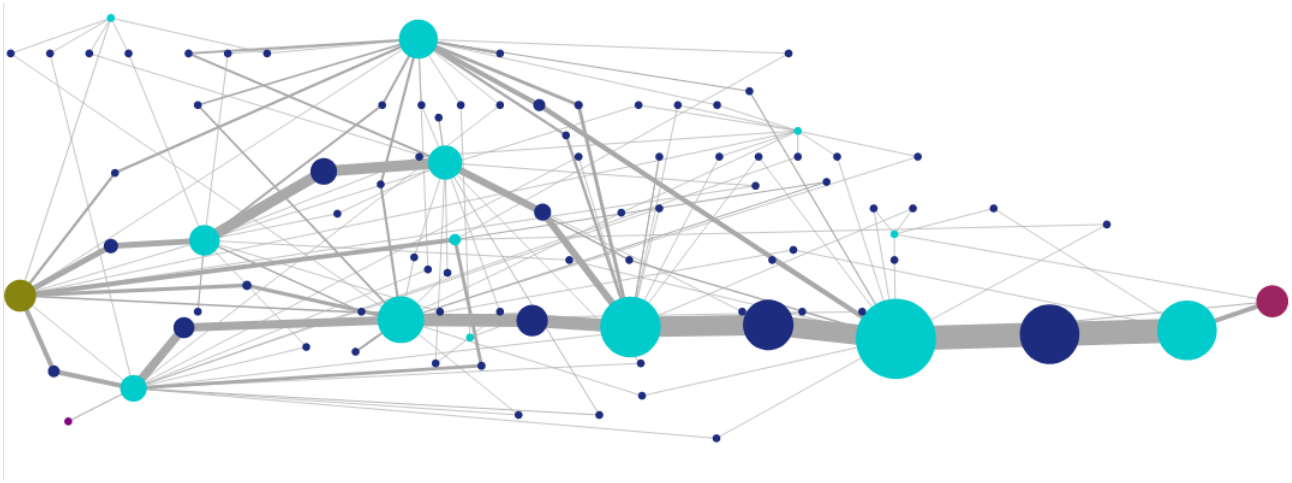

Рисунок 2 – Восстановленный граф

В этой модели голубым цветом обозначены операторские вершины, синим соответсвенно человеческие. Крого того, для удобства анализа выделены две фиктивные вершины: стартовая и вершина окочания диалогов. Вдальнейшем для приведённых данных они в расчёт не брались.

На Рисунке 3 представлена модель адаптированная для текста работы. На ней будут показаны результаты работы алгоритмов:

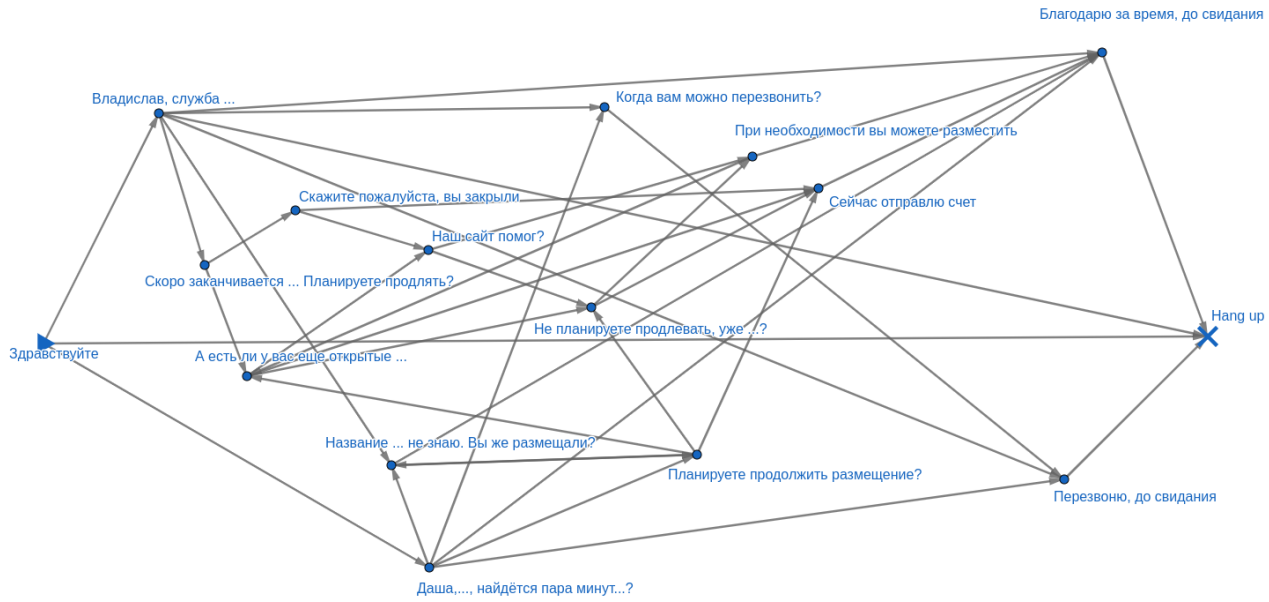

Рисунок 3 – Восстановленный граф

NB: В данном примере вершины человеческих кластеров и вершины операторских кластеров с фразами «Алло», «Я говорю» и «Повторите пожалуйста» были намеренно убраны, поскольку при их наличии восприятие вос-

приятие графа становится затруднительным. Про эти вершины будут написаны отдельные заметки везде, где граф будет использоваться в качестве примера.

По тем же причинам были удалены большинство рёбер в конец – человек может бросить трубку в любой момент, на алгоритмы эти рёбра не влияют.

Изначально при создании графа вершины человеческих фраз просто группируются по тому, какие 2 кластера находятся до и после, поэтому становится возможным их на рёбра, не потеряв ценной информации.

3.2. Результаты улучшения кластеризации

Алгоритм вносит небольшие изменения и способен объединять некоторые одинаковые кластеры.

Основная сложность состояла в анализе, так как для данных из разговоров с искуственным интелектом кластеризация фраз оператора изначально была очень хорошей в силу того, что фразы повторялись. Для случаев разговоров человека с человеком анализ результатов приходилось проводить вручную в силу отсутствия разметки данных.

Были проведены исследования по влиянию параметров алгоритма на результат. В результате чего были выявлены следующие закономерности:

— Пороговое значение схожести сообщений для каждого кластера должно варьироваться от 0.2 до 0.5. В этом промежутке в качестве пар схожих сообщений выбираются схожие пары сообщений. Для меньшего значения соответственно практически полностью одинаковые. Для большего у них появляется вариативность.
— Пороговое значение количества хороших пар в кластере должно варьироваться так от 0.25 до 0.5. Для больших значений в одном кластере начинают появляться наборы фраз несущие разный смысл.

Ниже, на Рисунке 4 представлены примеры различных значений промежутков порога для количества хороших пар в кластере:

нет
нет нет
нет
нет александр александрович

(a) 0.3

до свидания
спасибо до свидания
спасибо всего доброго до свидания
всё спасибо вам за напоминание всего доброго

(b) 0.5

але
на какую тему подскажите
даша
сколько минут
интернет
я

(c) 0.6

Рисунок 4 – Различные значения пороговой функции количества хороших пар

Для примера графа выше будет происходить следующее слияние отмеченное на Рисунке 5:

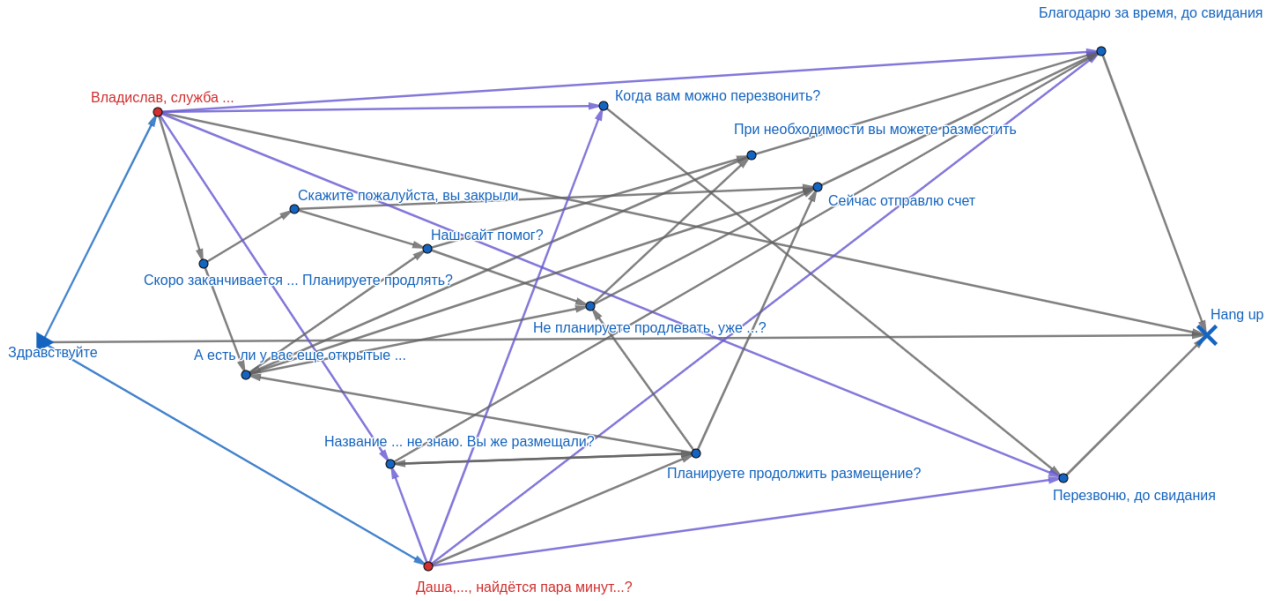

Рисунок 5 – Восстановленный граф

Среди данных вершин будут найдены две вершины с приветствием из разных запусков. Важно отметить, что при уменьшении количества кластеров, на которые будут поделены вершины будут сливаться такие пары как «Скоро заканчивается ..., планируете продлевать?» и «Не планируете продлевать, уже ...», в связи с тем, что семантика этих фраз похожа, а приветсвия сформулированы по-разному. Таким образом уменьшение количества кластеров проблему не решает.

Пары соответсвующий рёбер ведущих к одинаковым вершинам (рёбра помечены фиолетовым и синим) имеют схожие фразы, которые объединяются при рекластеризации (кластеры : («слушаю, да»), («нет вакансий»), («здрав-

ствуйте») и т.д.). Таким образом у вершин появляются множества кластеров до и после с большим пересечением и они объединяются.

Просто пересечь группы кластеров-оператора до и после не работает, т.к. в таком случае фразы «Алло» и «Повторите пожалуйста» будут объединяться в силу того, что в них идут рёбра и из них идут рёбра практически во все вершины.

3.3. Результаты работы алгоритма поиска отвлечений по рёбрам

Для данного алгоритма критично качество кластеризации и наличие небольшого количества выбросов.

В случае с данными из реальных диалогов зачастую не было видно правильно выделеных отвлечений из-за смешения кластеров и операторских и человеческих вершин.

Для данных из диалогов с существующей диалоговой системой алгоритм находил все отвлечения и так же отмечал некоторое количество дополнительных вершины, которые отвлечениями не являются. Основная причина столь большого различия в том, что операторские фразы во втором случае кластеризуются практически идеально.

На данном этапе этот алгоритм корректно использовать в качестве отсечения вершин, которые не подходят в качестве отвлечений, но были ошибочно выбраны другими алгоритмами. Так же алгоритм хорошо подсвечивает цепочки фраз, которые часто произносятся друг за другом, такие как например «Когда вам перезвонить» + «Перезвоню, до свидания», среди цепочек, которые содержат первую фразу и не обрываются после неё – **74.8%**.

3.4. Результаты алгоритма поиск циклов

После того, как мы нашли циклы мы их удаляем, таким образом основной граф становится менее громоздким и его легче анализировать. Из особенностей реализации важно отметить, что удаляются не все циклы, а только размера не более чем 5. Рассмотрим пример следующего диалога:

Оператор: Здравствуйте

Человек: Ало

*Оператор: Меня зовут Даша, вы ***[1]?*

[1]Здесь и далее *** будут использоваться для обезличивания частей диалога, которые не содержат важных для алгоритма данных, но несут в себе персональную информацию или информацию о компаниях.

*Человек: *шуршание**[2]

Оператор: Извините, я вас не расслышал, повторите пожалуйста

Человек: Я говорю давайте задавайте

*Оператор: Вы сейчас ***?*

Человек: сейчас нет

*Оператор: Вы нашли ***?*

Человек: да

*Оператор: Убрать *** в архив?*

*Человек: *неразборчивый текст**

Оператор: Извините, я вас не расслышал, повторите пожалуйста

*Человек: ненадо пока убирать ****

Оператор: Извините за беспокойство, всего доброго, до свидания

Если мы удалим все фразы между двумя просьбами повторить,то потеряем состояния, которые не являются отвлечениями.

Ниже, на Рисунке 6 приведен график, в котором сравниваются количества вершин, до и после применения алгоритма удаления циклов. Для каждого графа так же указано, датасет какого размера был использован при создании:

[2]Здесь и далее *описание* для пропуска не содержательных частей с коротким описанием.

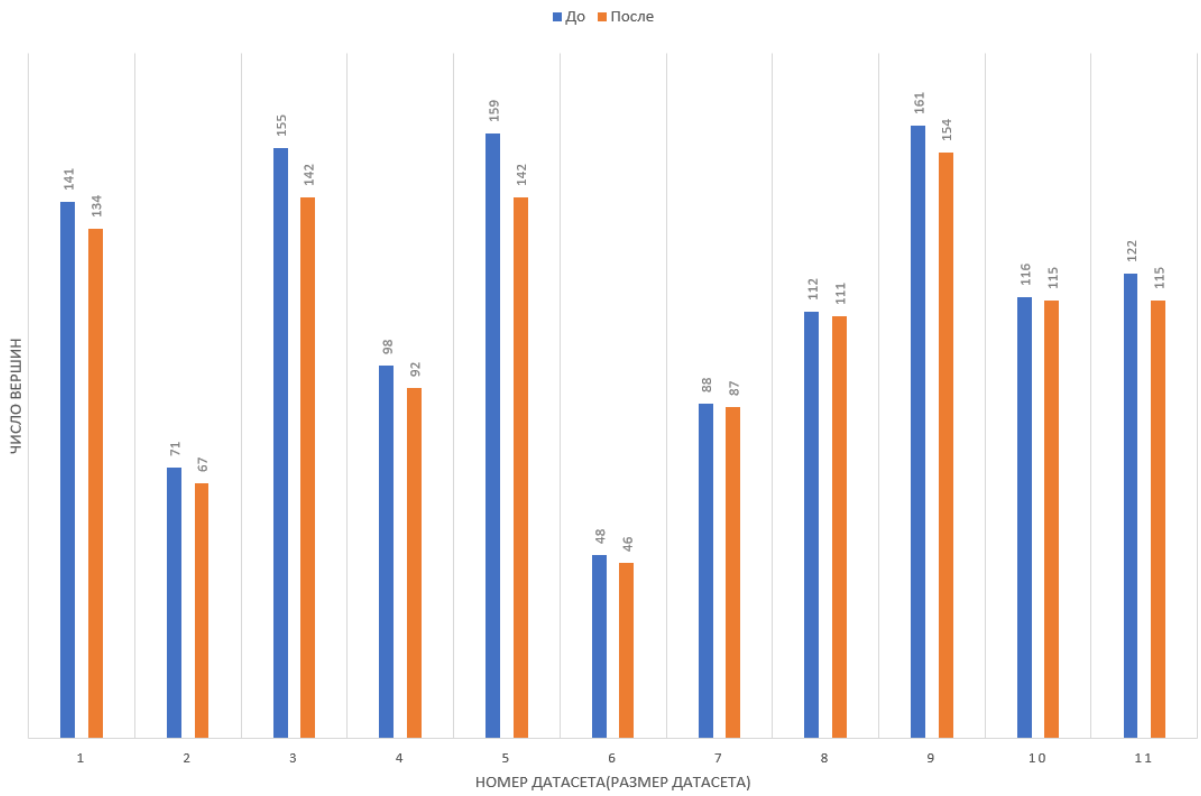

Рисунок 6 – Изменение количества вершин

Ниже, на Рисунке 7 приведен график, в котором сравниваются количества рёбер, до и после применения алгоритма удаления циклов.

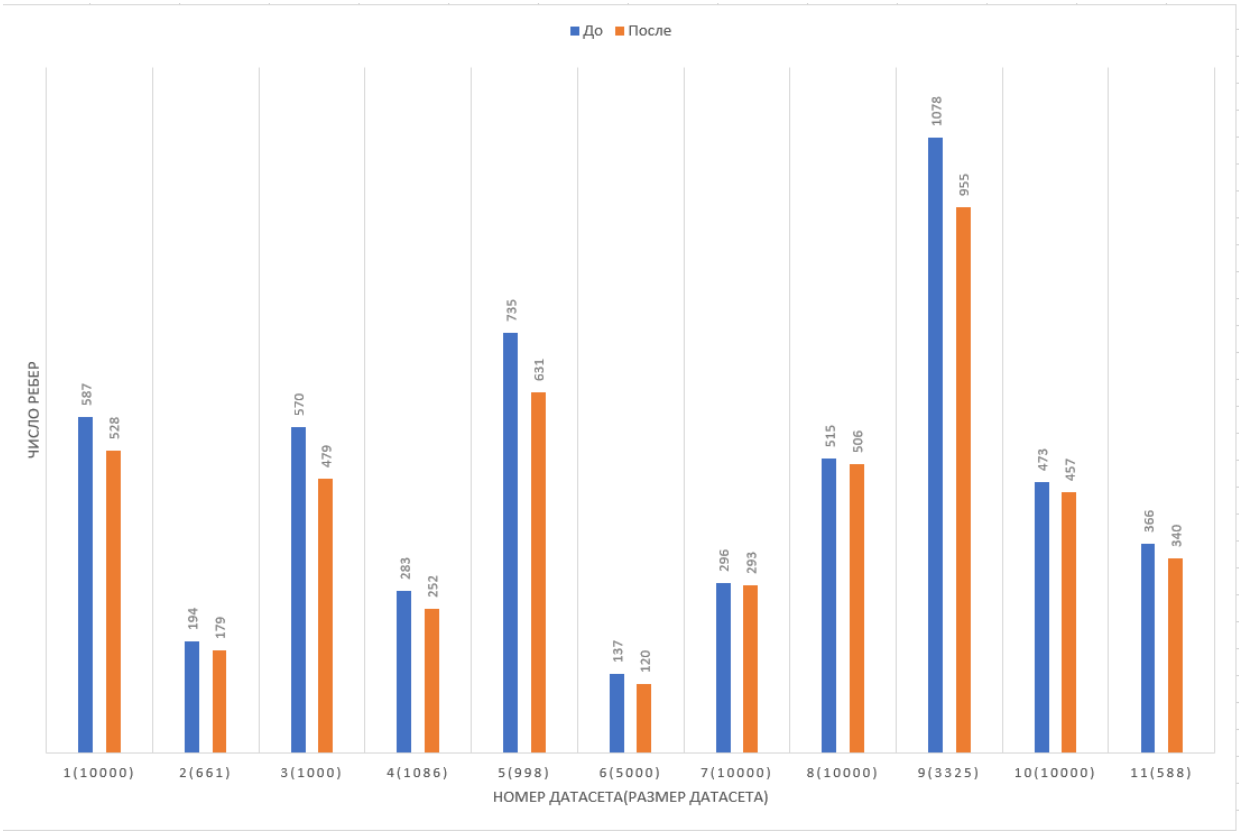

Рисунок 7 – Изменение количества вершин

Среди этих датасетов есть 2 принципиально разных типа:

Первый тип – **оператор + человек**, эти датасеты нам предоставили компании, и в них человек придерживается некоторого скрипта, но заметно от него отклоняется, поэтому при меньших размерах порядок вершин в них такой же. К этому типу относятся соответственно датасеты 2, 4, 5, 9, 11.

Второй тип – **искуственный интелект + человек**, эти датасеты были собраны соответственно компанией при звонках со скриптами написанными вручную. К этому типу соответственно относятся 1, 3, 6, 7, 8, 10.

В обоих типах при увеличении размера датасета увеличивается количество вершин. Важно отметить, что большинство из вершин, это ответы человека, количество же операторских вершин варьируется от 12 до 25. Для случая диалогов человека с человеком размеры графа могут быть изначально больше, это объясняется тем, что в нём большее разнообразие фраз.

Вырезанные ребра и вершины соответственно отделяются и становится проще анализировать граф.

Если рассмотреть какие вершины из операторских попадают в циклы, то хорошо видно тенденцию того, что есть несколько вершин у которых большое количество ответов попадает в циклы. Некоторые такие вершины, как можно

заметить исчезают вообще. Для остальных можно подобрать пороговое значение при котором можно будет считать, что вершина является триггером отвлечения.

Для примера графа указанного выше отметятся следующие вершины, содержащие большое количество фраз в циклах (см. Рисунок 8:

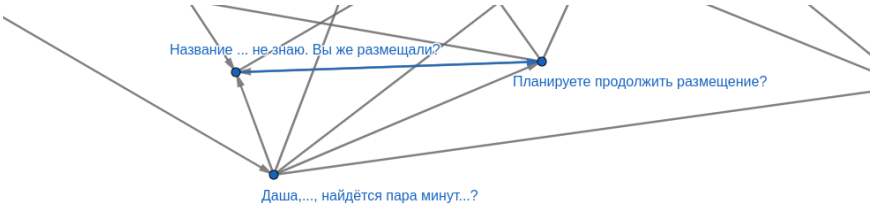

Рисунок 8 – Отмеченный цикл

Интересным будет заметить, что из двух вершин именно «Название ... не знаю, вы же размещали.» попадёт в отвлечения, т.к. внутри цикла находятся именно её сообщения, а сообщения «Планируете продолжить размещение?» находятся на границах цикла.

Более того, самый большой процент (84.7%) попавших в циклы вершин будет иметь фраза «Я говорю», которая была исключена, т.к. сильно усложняла граф. Она является примером **общего** для всех графов отвлечения (ответом на вопрос «повторите» и т.п.), так что поведение алгоритма выделившего её является корректным.

3.5. Программная реализация

Предложенные алгоритмы были реализованы на языке `Python` с использованием интегрированной среды разработки `Pycharm`. Кроме того использовался текстовый редактор `Visual Studio Code` с внутренним плагином для визуализации графов описанных в формате `JSON`.

Компания использует приватный `Git`-репозиторий, находящийся на хостинге `Gitlab`. Для удобства локальной разработки приложение запускалось в `Docker`-контейнере.

ЗАКЛЮЧЕНИЕ

В ходе выполнения работы была рассмотрена задача выделения отвлечений в графовой модели для голосовой диалоговой системы. Для её решения были разработаны и написаны алгоритмы основанные на особенностях разговоров оператора с человеком. Были учтены особенности существующей модели графовых диалогов и данные полученные при её использовании.

Были написаны алгоритмы выделения циклов и поиска отвлечений по рёбрам. Эти алгоритмы позволяют выделить значительное количество отвлечений и уменьшают размер основного графа. Кроме того в ходе разработки стало ясно, что алгоритмы очень чувствительны к кластеризации и возникла необходимость улучшить её качество. Для чего был разработан алгоритм на основе данных о соседних фразах.

Поскольку работа является частью большего проекта. То интеграция в реальное окружение запланирована по завершению всех частей проекта. Разработка должна будет автоматизировать часть рабочих процессов.

Одно из возможных направлений развития работы, использовать информацию о пользователе в диалоге, сделав его таким образом более естественным. Так же существует возможность выделить некоторые общие отвлечения для всех диалогов, получая таким образом возможность предсказывать отвлечения в виде подсказок.

СПИСОК ИСПОЛЬЗОВАННЫХ ИСТОЧНИКОВ

1 *Мухин О.* Моделирование систем: учебник [Электронный ресурс]. — 2014. — URL: http://stratum.ac.ru/education/textbooks/modelir/lection34.html (дата обр. 30.04.2020).

2 *Нгок Н. Б.*, *Тузовский А. Ф.* Классификация текстов на основе оценки семантической близости терминов // Izvestiya Tomskogo Politekhnicheskogo Universiteta Inziniring Georesursov. — 2012. — Т. 320, № 5. — С. 43–48.

3 *Чернова Н.* Теория вероятностей: Учеб. пособие/Новосиб. гос. ун-т // Новосибирск. 2007, 160 с. — 2007.

4 Dialogue-based Management of user Feedback in an Autonomous Preference Learning System. / J. Lucas-Cuesta [et al.] //. Vol. 1. — 09/2010. — P. 330–336.

5 *Dönmez İ.*, *Pashaei E.*, *Pashaei E.* Word Vector Space for Text Classification and Prediction According to Author. —.

6 *Doshi F.*, *Roy N.* Efficient model learning for dialog management // 2007 2nd ACM/IEEE International Conference on Human-Robot Interaction (HRI). — 2007. — P. 65–72.

7 *Jokinen K.*, *McTear M.* Spoken Dialogue Systems. — Morgan & Claypool Publishers, 2010. — (Synthesis lectures on human language technologies). — ISBN 9781598295993. — URL: https://books.google.ru/books?id=ualwulnD020C.

8 *Li H.* Text Clustering // Encyclopedia of Database Systems / ed. by L. LIU, M. T. ÖZSU. — Boston, MA : Springer US, 2009. — P. 3044–3046. — ISBN 978-0-387-39940-9. — DOI: 10.1007/978-0-387-39940-9_415. — URL: https://doi.org/10.1007/978-0-387-39940-9_415.

9 *Wikipedia contributors*. Dialog manager — Wikipedia, The Free Encyclopedia. — 2020. — URL: https://en.wikipedia.org/w/index.php?title=Dialog_manager&oldid=941891414 (visited on 04/18/2020).

10 *Williams J. D.*, *Young S.* Partially observable Markov decision processes for spoken dialog systems // Computer Speech & Language. — 2007. — Vol. 21, no. 2. — P. 393–422.

11 *Wills P.*, *Meyer F. G.* Metrics for graph comparison: A practitioner's guide // Plos one. — 2020. — Vol. 15, no. 2. — e0228728.

ПРИЛОЖЕНИЕ А. ИСХОДНЫЙ КОД НА ЯЗЫКЕ PYTHON

```python
import json
import logging
import os

import numpy as np
import requests
from flask import Flask, Response, jsonify, request
from flask_caching import Cache
from sklearn.cluster import AgglomerativeClustering, KMeans

from reconstruction import algorithm_descriptions, reconstruct_graph
from reconstruction.data_structure import DatasetStorage
from reconstruction.data_structure.graph import save_graph
from reconstruction.data_structure.message import messages_to_json
from reconstruction.embedding import EmbeddingStorage

@app.route('/graph', methods=['POST'], strict_slashes=False)
@cache.cached(key_prefix=make_cache_key)
def build_graph():
    dataset_name, dialogs, alg_name, hparams = parse_request()
    graph = reconstruct_graph(dialogs, alg_name, hparams, embedding_storage)
    save_graph(graph, dataset_name)
    return jsonify(graph)
```

```python
from reconstruction.reconstruct_graph import reconstruct_graph, algorithm_descriptions
```

```python
from sklearn.cluster import KMeans

import reconstruction.algorithm as alg
from reconstruction.embedding.mean_vectorizer import MeanEmbeddingVectorizer

algorithms = {
    "cluster_all_messages": {
    },
    "cluser_incoming_outgoing": {
    },
    "cluster_incoming_outgoing_on_each_step": {
    },

    "cluster_outgoing_and_incoming_after": {
    },
    "cluster_prev_and_next": {
    },
    "cluster_prev_and_next_recluster_incoming": {
        "reconstruction": alg.cluster_prev_and_next_recluster_incoming,
```

```
        "clustering": {
            "emb_name": "tenth.norm-sz500-w7-cb0-it5-min5.w2v",
            "vectorizer": MeanEmbeddingVectorizer,
            "alg": KMeans
        }
    }
}
```

```
from collections import Counter, defaultdict

import numpy as np
from sklearn.metrics.pairwise import cosine_similarity

from reconstruction.data_structure.graph import Node
from reconstruction.converter.converter import Graph
from reconstruction.digressions.digressions import cycle_finder


def calc_centroid(messages):


def create_nodes(messages, cluster_labels, remove_msgs=[]):
        #secsitive code - create nodes using information about msgs from cycle, remove


def add_start_node(graph, dialogs, skip):
    graph["nodes"].append({
        "id": -1,
        "group": -1,
        "centroid": 0,
        "texts": [],
        "seq_num": -1,
        "seq_num_std": 0,
        "incoming": False,
        "size_coef": 0.4,
        "size": len(dialogs),
        "x": 0,
        "y": 0
    })
    start_cluster = -1
    graph["start"] = start_cluster
    weights = defaultdict(int)
    for dialog in dialogs:
        if (len(dialog) > 0):
            msg = dialog.log[0]
            cluster = msg.cluster
            if cluster and cluster not in skip:
                weights[cluster] += 1
    for cluster in weights:
```

```
        graph["links"].append({
            "source": int(start_cluster),
            "target": int(cluster),
            "weight": int(weights[cluster])
        })


def add_end_node(graph, dialogs, skip):


def nodes_to_graph(nodes, dialogs, node_size_limit=0):
    #secsitive code - create nodes, count size, linkes with weights, work with digressi


def num_of_vertexes(msgs, remove_msgs=[]):
    vertexs = defaultdict(set)
    none_counter = 0
    for msg in msgs:
        if msg in remove_msgs:
            continue
        if msg.next_msg is not None and msg.next_msg not in remove_msgs:
            vertexs[msg.cluster].add(msg.next_msg.cluster)
    edges_num = 0
    for _, set_list in vertexs.items():
        edges_num += len(set_list)
    print("vertexs:", len(vertexs), " edges: ",
          edges_num, " nones: ", none_counter)


def create_graph(messages, cluster_labels, dialogs, node_size_limit=0, call_remove_cycl
    #sensitive code - create nodes, graph, call cycles finder, count statistics
```

```
from reconstruction.config import algorithms
from reconstruction.clustering import MessageClusterizer
from reconstruction.create_graph import create_graph
from reconstruction.coords import graphviz_coords
import logging


def get_vectorizer(embedding_storage, clustering_config):


def get_clustertering_alg(clustering_config, **kwargs):


def preprocess_dialogs(vectorizer, dialogs):


def reconstruct_graph(dialogs, alg_name, hparams, embedding_storage):
```

```
    alg_config = algorithms[alg_name]
    reconstruction_alg = alg_config["reconstruction"]

    clustertering_alg = get_clustertering_alg(alg_config["clustering"],
                                              **{"n_clusters": int(hparams["n_clusters"

    vectorizer = get_vectorizer(embedding_storage, alg_config["clustering"])
    preprocess_dialogs(vectorizer, dialogs)

    clustered_messages, cluster_labels = reconstruction_alg(
        dialogs, clustertering_alg, **hparams)

    min_node_size = int(hparams["min_node_size"])

    remove_cycles = hparams["remove_cycles"]
    graph = create_graph(clustered_messages, cluster_labels, dialogs,
                         node_size_limit=min_node_size, call_remove_cycles=remove_cycle

    digressions_by_different_income(graph, dialogs)

    graphviz_coords(graph)
    return graph
```

```
from reconstruction.algorithm.cluser_incoming_outgoing import \
    cluser_incoming_outgoing
from reconstruction.algorithm.cluster_all_messages import cluster_all_messages
from reconstruction.algorithm.cluster_incoming_outgoing_on_each_step import \
    cluster_incoming_outgoing_on_each_step
from reconstruction.algorithm.cluster_outgoing_and_incoming_after import \
    cluster_outgoing_and_incoming_after
from reconstruction.algorithm.cluster_prev_and_next import \
    cluster_prev_and_next
from reconstruction.algorithm.cluster_prev_and_next_recluster_incoming import \
    cluster_prev_and_next_recluster_incoming
```

```
import itertools
import logging
import multiprocessing
import random
import sys
import time
from collections import Counter, defaultdict
from functools import partial
from typing import Dict, List

import numpy as np
from sklearn.cluster import KMeans

```

```
14 from reconstruction.data_structure import Dialog
15 from reconstruction.data_structure.dialog import get_dialog_messages
16 from reconstruction.clustering import MessageClusterizer
17 from reconstruction.algorithm import cluster_prev_and_next
18 from reconstruction.algorithm.utils import split_incoming_outgoing
19
20
21 def vocab_distribution(words):
22     total = len(words)
23     vocab_dist = dict(Counter(words))
24     for word in vocab_dist.keys():
25         vocab_dist[word] /= total
26     return vocab_dist
27
28
29 def distribution_text_similarity(msg1, msg2, vocab_dist):
30     words1 = msg1.text.split()
31     words2 = msg2.text.split()
32     counter1 = dict(Counter(words1))
33     counter2 = dict(Counter(words2))
34     unique_words1 = set(words1)
35     unique_words2 = set(words2)
36
37     cross = sum(
38         min(counter1[word], counter2[word]) * vocab_dist[word]
39         for word in unique_words1 & unique_words2
40     )
41
42     union = sum(
43         max(counter1.get(word, 0), counter2.get(
44             word, 0)) * vocab_dist[word]
45         for word in unique_words1 | unique_words2
46     )
47
48     similarity = cross / union
49     return similarity
50
51
52 def remove_unvalid_clusters(msgs, clusters, similarity_threshold=0.4, cluster_threshold
53     start = time.time()
54     words = [w for t in msgs for w in t.text.split()]
55     vocab_dist = vocab_distribution(words)
56     print("vocab_distribution:", time.time() - start, file=sys.stderr)
57
58     start = time.time()
59     cluster_messages = defaultdict(list)
60     for cl_n, msg in zip(clusters, msgs):
61         cluster_messages[cl_n].append(msg)
62
63     print("cluster_messages:", time.time() - start, file=sys.stderr)
```

```
64
65      start = time.time()
66      msgs_in_cluster = dict(Counter(clusters))
67      good_connects = defaultdict(int)
68      msgs_in_stat = defaultdict(int)
69      pool = multiprocessing.Pool(5)
70
71      for cl_n, msg_l in cluster_messages.items():
72          if cl_n in msgs_in_cluster:
73              for i, msg in enumerate(msg_l):
74                  for j in range(i):
75                      # check ~3 * cl_num random pairs
76                      if (random.randint(0, msgs_in_cluster[cl_n] // 3) == 0):
77                          msgs_in_stat[cl_n] += 1
78                          sim = distribution_text_similarity(
79                              msg, msg_l[j], vocab_dist)
80                          if sim > similarity_threshold:
81                              good_connects[cl_n] += 1
82      bad_clusters = {
83          cl_n for cl_n, correct in good_connects.items()
84          if (correct / msgs_in_stat[cl_n] < cluster_threshold)
85      }
86
87      print(msgs_in_cluster)
88      print("msg list size: ", len(cluster_messages),
89            " keys: ", cluster_messages.keys)
90      print([
91          cl_n for cl_n, correct in good_connects.items()
92          if (correct / msgs_in_stat[cl_n] > cluster_threshold)
93      ], " good one")
94      print([
95          cl_n for cl_n, correct in good_connects.items()
96          if (correct / msgs_in_stat[cl_n] < cluster_threshold)
97      ], "bad one")
98
99      start = time.time()
100     new_msgs, new_clusters = [], []
101     for msg, cl_n in zip(msgs, clusters):
102         if cl_n not in bad_clusters:
103             new_msgs.append(msg)
104             new_clusters.append(cl_n)
105     print("filter:", time.time() - start, file=sys.stderr)
106
107     return new_msgs, new_clusters
108
109
110 def recluster_incoming(clustered_messages, cluster_labels, dialogs, cluster, **hparams)
111     """Делит входящие сообщения иначе, с учётом того, как они кластеризовались независи
112     1. Кластеризуем входящие сообщения независимо
113     2. Разделяем ноды из входящих, если там много сообщений каких-то типов
```

```
    3. Объединяем ноды с полностью одинаковыми кластерами
    """
    print(hparams, file=sys.stderr)

    proportion_threshold = float(hparams.get("proportion_threshold", 0.25))
    cluster_threshold = float(hparams.get("cluster_threshold", 0.5))
    similarity_threshold = float(hparams.get("similarity_threshold", 0.5))

    start = time.time()

    messages = get_dialog_messages(dialogs)
    incoming, _ = split_incoming_outgoing(messages)
    print(" get_dialog_messages split_incoming_outgoing:",
          time.time() - start, file=sys.stderr)
    start = time.time()
    incoming_cluster_labels = cluster.fit_predict(incoming)
    print("fit_predict:", time.time() - start, file=sys.stderr)

    start = time.time()
    incoming, incoming_cluster_labels = remove_unvalid_clusters(
        incoming, incoming_cluster_labels,
        similarity_threshold=similarity_threshold,
        cluster_threshold=cluster_threshold)

    print("remove_unvalid_clusters:", time.time() - start, file=sys.stderr)

    start = time.time()
    pair_clusters = defaultdict(int)  # new clusters, incoming only
    for msg, cluster in zip(incoming, incoming_cluster_labels):
        if msg in clustered_messages:  # cluster_messages doesn't include starts of dia
            pair_clusters[msg] = cluster

    node_clusters = defaultdict(lambda: defaultdict(list))
    new_clustering = defaultdict(int)

    for msg, cluster in zip(clustered_messages, cluster_labels):
        if msg in pair_clusters:
            node_clusters[cluster][pair_clusters[msg]].append(msg)

    new_cluster_num = cluster_labels.max()
    alloc_clusters = defaultdict()
    changing_msgs = 0

    for _, clusters in node_clusters.items():
        node_size = sum(map(len, clusters.values()))
        small_part = []
        for income_cluster_n, msgs in clusters.items():
            cluster_proportion = (len(msgs) / node_size)
            if (cluster_proportion > proportion_threshold):
                if income_cluster_n not in alloc_clusters:
```

```
164                     new_cluster_num += 1
165                     alloc_clusters[income_cluster_n] = new_cluster_num
166
167                 cluster_num = alloc_clusters[income_cluster_n]
168                 for msg in msgs:
169                     new_clustering[msg] = cluster_num  # check
170                     changing_msgs += 1
171             else:
172                 small_part.append(income_cluster_n)
173
174         if len(small_part) == 1:
175             alone_cluster = small_part[0]
176             if alone_cluster not in alloc_clusters:
177                 new_cluster_num += 1
178                 alloc_clusters[alone_cluster] = new_cluster_num
179             cluster_num = alloc_clusters[alone_cluster]
180             for msg in clusters[alone_cluster]:
181                 new_clustering[msg] = cluster_num  # check
182                 changing_msgs += 1
183     print("split clusters:", time.time() - start, file=sys.stderr)
184
185     start = time.time()
186     list_of_changes = []
187     for msg in new_clustering:
188         i = clustered_messages.index(msg)
189         clustered_messages[i].cluster = new_clustering[msg]
190         cluster_labels[i] = new_clustering[msg]
191         list_of_changes.append(i)
192     print("rename clusters:", time.time() - start, file=sys.stderr)
193     return clustered_messages, cluster_labels
194
195
196 def cluster_prev_and_next_recluster_incoming(dialogs: List[Dialog], cluster: MessageClu
197     """Кластер входящих определяется как единый кластер, если следующий исходящий класт
198
199     1. Строим кластеры для всех исходящий сообщений
200     2. Для каждого кластера находим все последующие входящие сообщения до следующего ис
201     3. Эти входящие сообщения разделяются на кластеры в зависимости от того какой исход
202     4. Номер кластера задается на основе следующего исходящего кластера.
203
204     Таким образом переход от кластера входящих сообщений к кластеру исходящих становитс
205
206     Входящие сообщения разделяются, с учётом того, как они кластеризовались независимо
207         1. Кластеризуем входящие сообщения независимо
208         2. Разделяем ноды из входящих, если там много сообщений каких-то типов
209         3. Объединяем ноды с полностью одинаковыми кластерами
210
211     Arguments:
212         dialogs {List[Dialog]} -- Список диалогов
213         cluster {MessageClusterizer} -- Алгоритм кластеризации исходящих сообщений.
```

```
214
215     Returns:
216         messages {List[Message]} -- Список сообщений
217         cluster_labels {np.array} -- Список меток кластера для сообщений
218     """
219
220     start = time.time()
221     clustered_messages, cluster_labels = cluster_prev_and_next(
222         dialogs, cluster)
223     print("Init clustering:", time.time() - start, file=sys.stderr)
224     start = time.time()
225     clustered_messages, cluster_labels = recluster_incoming(
226         clustered_messages, cluster_labels, dialogs, cluster, **hparams)
227     print("Overall reclustering:", time.time() - start, file=sys.stderr)
228     return clustered_messages, cluster_labels
```

```
1  import json
2  from collections import defaultdict
3
4  class Graph:
5      def __init__(self, messages, digressions=[]):
6          msgs_dict = defaultdict(set)
7          self.vertexes = []
8          for msg in messages:
9              msgs_dict[msg.cluster].add(msg)
10         for cluster, msgs in msgs_dict.items():
11             tr_dict = defaultdict(set)
12             transactions = []
13             for msg in msgs:
14                 if msg.next_msg:
15                     tr_dict[msg.next_msg.cluster].add(msg)
16             for next_cluster, msgs_tr in tr_dict.items():
17                 transactions.append(Transaction(msgs_tr, int(cluster), int(next_cluster
18             self.vertexes.append(Vertex(int(cluster), transactions, cluster in digressi
19         self.digressions = digressions
20
21     def __str__(self):
22         str_vert = []
23         for vertex in self.vertexes:
24             str_vert.append(str(vertex))
25         return '{"nodes": {' + ", ".join(str_vert) + '},' + str(self.get_digr()) + '}'
26
27     def get_digr(self):
28         digr_tr = None
29         return '"digressionNodes": '  + json.dumps(
30         {
31             "dig:root": {
32                 "Id": "dig:root",
33                 "OnEnter": {"Type": "None"},
```

```
                "Transitions": [digr_tr]
            }
        })

    def str_transactions_msgs(self):
        res = ''
        for vertex in self.vertexes:
            for transition in vertex.transitions:
                res += transition.str_msgs()
        return res


class Vertex:
    def __init__(self, cluster_n, transitions, is_digressions=False):
        self.cluster_n = cluster_n
        self.transitions = transitions
        self.is_digressions = is_digressions

    def __str__(self):
        str_trans = []
        for transition in self.transitions:
            str_trans.append(str(transition))
        return '"{}":'.format(self.cluster_n) + json.dumps({
            "Id": str(self.cluster_n),
            "OnEnter": {
                "Type": "Chain",
                "InnerReactions": [
                {
                    "Type": "Simple",
                    "Reactions": [
                    {
                        "msgId": "RawTextChannelMessage",
                        "text": 'text{}'.format(self.cluster_n)
                    }
                    ]
                }
                ],
            "Transitions": [
                [json.loads(json_tr) for json_tr in str_trans]
            ]
            }
        })

class Transaction:
    def __init__(self, msgs_set, from_cluster, to_cluster):
        self.msgs = msgs_set
        self.from_cluster = from_cluster
        self.to_cluster = to_cluster
        self.tr_name = 'fr_' + str(from_cluster) + '_to_' + str(to_cluster)

```

```
    def __str__(self):
        return json.dumps({
          "Condition": {
            "Type": "Fact",
            "FactName": "input_info",
            "InnerCondition": {
                "Type": "Field",
                "FieldName": "type",
                "InnerCondition": { "Type": "ItemEquals", "Item": self.tr_name }
            },
            "Id": str(self.to_cluster),
            "Priority": 0
          }
        })

    def compare(self, tr):
        self.msgs.intersection(tr.msgs)

    def str_msgs(self):
        msgs = []
        for msg in self.msgs:
            msgs.append('{ "label":' + str(msg.labels or '{}') + ', "msg":' + '"{}"'.fo
        return '{ "name":' + '"{}"'.format(self.tr_name) + ', "msgs": [' + ", ".join(ms
```

```
from collections import defaultdict
from reconstruction.data_structure.dialog import Dialog

import numpy as np


def bfs_digressions_finder_0(graph, node, set_of_outgoing_clusters, set_of_incoming_clu
    """Trying to get tail from expected root, and cut vertices
    which can be potential continue after digressions. As a potential and
    we mark vertices befor root and next vertexes.

    Have troubles because of cycles, which does not exists in real dialogs"""
    first_step = set_of_outgoing_clusters[node]
    queue = set()
    marked = set()
    for node in first_step:
        queue = queue | set_of_outgoing_clusters[node]
        potential_end = set_of_incoming_clusters[node]
    for source in set_of_incoming_clusters[node]:
        potential_end = potential_end | set_of_outgoing_clusters[source]
    from_q = queue.copy()
    while queue:
        cur_node = queue.pop()
        for next_node in set_of_outgoing_clusters[cur_node]:
            if (next_node not in potential_end and next_node not in from_q):
```

```
                from_q.add(next_node)
                queue.add(next_node)
    marked = from_q.copy()
    # coming to digr part only from dirg + add to marked + add starts
    # queue = {graph[”start”]}
    # from_q = queue
    # while queue:
    #     cur_node = queue.pop()
    #     print(”Queue: ” + len(queue))
    #     for next_node in set_of_outgoing_clusters[cur_node]:
    #         if (next_node != node and (next_node not in from_q)):
    #             from_q.add(next_node)
    #             queue.add(next_node)
    # marked = marked.difference(from_q)
    if len(marked) > 0:
        graph['digressions'].append({
            ”root”: node,
            ”nodes”: list(marked)
        })


def potential_digr_ver(unachive, dialogs_num, dialogs_out_of_dig):
    possible_list = list()  # list with possible errors
    for cluster in unachive:
        if (len(dialogs_out_of_dig[cluster]) / dialogs_num[cluster] > 0.8):
            possible_list.append({
                ”cluster”: cluster,
                ”ex_list”: dialogs_out_of_dig[cluster]
            })
    return possible_list


def digressions_finder(graph, node, dialogs):
    cluster_labels = set()
    achieve = set()
    dialogs_num = defaultdict(int)
    dialogs_out_of_dig = defaultdict(list)
    for dialog in dialogs:
        stop = True
        for msg in dialog.log:
            dialogs_num[msg.cluster] += 1
            if (msg.cluster != node) and stop:
                achieve.add(msg.cluster)
                dialogs_out_of_dig[msg.cluster].append(msg.text)
            else:
                stop = False
            cluster_labels.add(msg.cluster)
    unachive = cluster_labels - achieve
    if len(cluster_labels) / 2 > len(unachive):
        graph['digressions'].append({
```

```
            "root": node,
            "nodes": list(unachive),
            "possible": potential_digr_ver(unachive, dialogs_num, dialogs_out_of_dig)
        })


def digressions_by_different_income(graph, dialogs):
    """Find digressions with finding potential starts and cut with some heuristic"""
    set_of_incoming_clusters = defaultdict(set)
    for link in graph["links"]:
        set_of_incoming_clusters[link["target"]].add(link["source"])
    num_of_incoming_clusters = dict()
    for node in graph["nodes"]:
        cluster = node["id"]
        num_of_incoming_clusters[cluster] = len(
            set_of_incoming_clusters[cluster])
        if (num_of_incoming_clusters[cluster] > 3):  # random const
            if (not node["incoming"]):
                digressions_finder(graph, cluster, dialogs)


def find_similar_nodes(nodes, income_shared):
    nodes_sets = defaultdict(set)
    potential_merge = [[]]
    for node in nodes:
        for transaction in node.transitions.keys():
            if transaction in income_shared:
                nodes_sets[node.cluster_n].add(transaction)
    for cluster1, set1 in nodes_sets.items():
        if (len(set1) > 0):
            nodes = []
            for cluster2, set2 in nodes_sets.items():
                if (len(set1 & set2)) / (len(set1 | set2)) > 0.7 and len(set1) * len(se
                    nodes.append(cluster2)
            if (len(nodes) > 0):
                nodes.append(cluster1)
            if (len(nodes) > 1):
                potential_merge.append(nodes)
    return potential_merge


def cut_cycle(dialog, start, end, messages, remove_marks):
    log = dialog.log
    if len(log) > end + 1:
        log[start].next_msg = log[end + 1]
        log[end + 1].prev_msg = log[start]
    else:
        log[start].next_msg = None
    new_log = log[: start + 1] + log[end + 1:]
    counter = start
```

```
126     for msg in log[start + 1: end]:
127         counter += 1
128         if msg in messages:
129             index = messages.index(msg)
130             remove_marks.append(index)
131     trigger_msg = log[start + 1]
132     end_msg = log[end]
133     while trigger_msg != end_msg and not trigger_msg.incoming:
134         if trigger_msg.next_msg:
135             trigger_msg = trigger_msg.next_msg
136         else:
137             trigger_msg = None
138     return Dialog(dialog._id, dialog.task, new_log), trigger_msg, log[start + 1: end]
139
140
141 def cycle_finder(dialogs, messages):
142     """Find digressions with finding cycles"""
143     remove_marks = []
144     new_dialogs = []
145     trigger_list = []
146     remove_part = []
147     for dialog in dialogs:
148         new_dialog = dialog
149         have_cycle_flag = True
150         cycles_counter = 0
151         while have_cycle_flag:
152             cycles_counter += 1
153             have_cycle_flag = False
154             nodes_map = {}
155             counter = 0
156             for msg in new_dialog.log:
157                 cluster = msg.cluster
158                 if cluster in nodes_map and counter - nodes_map[cluster] < 5:
159                     #print("Cycle №", cycles_counter)
160                     have_cycle_flag = True
161                     new_dialog, trigger, remove_part_local = cut_cycle(
162                         new_dialog, nodes_map[cluster], counter, messages, remove_marks
163                     remove_part += remove_part_local
164                     if trigger:
165                         trigger_list.append(trigger)
166                     break
167                 if not msg.incoming:
168                     nodes_map[cluster] = counter
169                 counter += 1
170         new_dialogs.append(new_dialog)
171     trigger_clusters = defaultdict(int)
172     for trigger in trigger_list:
173         trigger_clusters[trigger.cluster] += 1
174     sorted_dict = {k: v for k, v in sorted(
175         trigger_clusters.items(), key=lambda item: item[1])}
```

```
    print("TRIGGERS STAT:")
    for cluster, num in sorted_dict.items():
        print("        cluster: ", cluster, " num: ", num, " in ", sum(map(lambda x: x.
    return new_dialogs, remove_marks, remove_part
```

```
from reconstruction.embedding.mean_vectorizer import MeanEmbeddingVectorizer
from reconstruction.embedding.embedding_storage import EmbeddingStorage
```

```
import re

from reconstruction.embedding.int_to_str import convert_int


def norm_text(text):
    text = re.sub(r'ё', 'е', text).lower()
    text = re.sub(r'"|!|\?|,|\.|%', '', text)
    text = re.sub(r'-', ' ', text)
    text = re.sub(r':', ' ', text)
    text = re.sub(r'::', ' ', text)
    replace_dict = {
        'ненадо': 'не надо',
        "ололо": 'ало',
        "алло": 'ало',
        "але": 'ало',
        "ален": 'ало',
        "дадада": "да да да",
        "дадад": "да да да",
        "додо": "да да",
        "дада": "да да",
        "\bпасибо\b": "спасибо",
        "\bпасиб\b": "спасибо",
        "пасиба":  "спасибо"
    }
    for from_w, to_w in replace_dict.items():
        text = re.sub(from_w, to_w, text)

    text_arr = []
    for word in text.split():
        try:
            converted_text = convert_int(word)
        except ValueError:
            converted_text = word
        text_arr.append(converted_text)

    return " ".join(text_arr)
```

ПРИЛОЖЕНИЕ Б. ДИПЛОМ КОНГРЕССА МОЛОДЫХ УЧЁНЫХ