Правительство Российской Федерации Федеральное государственное
автономное образовательное учреждение высшего образования
Национальный исследовательский университет
«Высшая школа экономики»
Факультет гуманитарных наук

Образовательная программа
«Компьютерная лингвистика»

Магистерская дисертация
На тему «Извлечение сценариев диалогов из пьес: кластеризация, классификация и семантическая маркировка графов пьес»

Название темы на английском «Towards Mining Dialogue Scenarios from Plays: Clustering, Classification and Semantic Labeling of Play Graphs»

Студента 2 курса
группы МКЛ181
Власова Владимира Павловича

Научный руководитель:
Клышинский Эдуард Станиславович,
доцент, кандидат технических наук

Москва — 2020

Содержание

Введение

В современном мире существует множество моделей, которые используются для ведения диалога с пользователем. Это могут быть как простые чат-боты для технической поддержки первой линии на маленьких сайтах, так и полномасштабные голосовые помощники от крупных компаний, такие как Алиса или Siri.

При этом при появлении желания создать нового или развить имеющегося диалогового агента, встает вопрос поиска подходящего корпуса текстов. Даже для построения агента, который будет просто вести диалог с пользователем, важным является определение темы этого разговора. Это является достаточно большой проблемой. Если у компании, которая начала вести разработку своей диалоговой системы и есть корпус диалогов, то вполне вероятно, что он не размечен и потребуется вкладывать дополнительные средства в разметку.

В данной работе приводится исследование того, **как на неразмеченном корпусе построить модель кластеризации отдельных реплик и коротких диалогов по темам**, что может быть использовано в будущем для полуавтоматической разметки диалогов. Фактически, после того, как модель будет обучена необходимо будет лишь разметить по темам сами кластеры, что на порядки снижает сложность разметки.

При этом стоит заметить, что полученные кластеры должны быть, с одной стороны, достаточно компактными, то есть в них должны входить семантически, стилистически или синтаксически похожие фразы или диалоги (иначе ими нельзя будет пользоваться, так как они не будут решать задачи разметки). С другой стороны, полученных кластеров не должно быть слишком много (иначе нельзя будет получить полноценной выгоды от применения модели - легче сделать полноценную разметку и применять методы обучения с учителем для определения класса будущих фраз).

В качестве объекта исследования был избран корпус русских пьес (Russian Drama Corpus)[8]. Выбор данного корпуса обусловлен двумя причинами. Во-первых, диалоги представленные в пьесах очень разнообразны как тематически, так и стилистически. В нем присутствуют диалоги из различных эпох и покрывающие огромное количество ситуаций. Следовательно, при помощи него можно выработать метод, подходящий для диалогового корпуса в наиболее об-

щем своем виде. Во-вторых, изучение корпуса русских пьес помогает выработать метод, который может быть полезен для проведения гуманитарных исследований, где наличие размеченного корпуса еще более маловероятно, чем в коммерческой среде.

В рамках работы был проведен анализ источников и не было найдено моделей, которые решают данную задачу для диалогов столь разных направлений, что представлены в корпусе русских пьес

Обзор литературы

В современном мире построение диалоговых систем является достаточно распространенной задачей. В то же время для обучения любой диалоговой системы, как ориентированной на решение задач, так и на ”разговор” с пользователем (что сейчас является одной из неотъемлемых функций голосовых ассистентов, таких как Алиса или Siri) требуется большой корпус данных.

При этом качественно размеченные данные обходятся дорого. Ученые пытаются решить эту проблему тремя способами:

- создавая модели для разметки данных;
- выделяя диалоговую структуру для эффективного использования неразмеченных данных;
- оптимизируя политику сбора данных для эффективного получения высококачественных данных. [1]

Данная работа ориентирована в первую очередь на первый из трех способов решения этой задачи.

[2]

[3]

[4]

[5]

Vladimir Vlasov, [1 июня 2020 г., 22:20:23]: Тогда я расскажу про две-три найденные работы, которые очень хорошо связаны с моей задачей. Плюс опишу word2vec, BERT, тензорфлоу эмбедер.

И про рекуррентные нейронки,...

Описание использованных данных

[8]

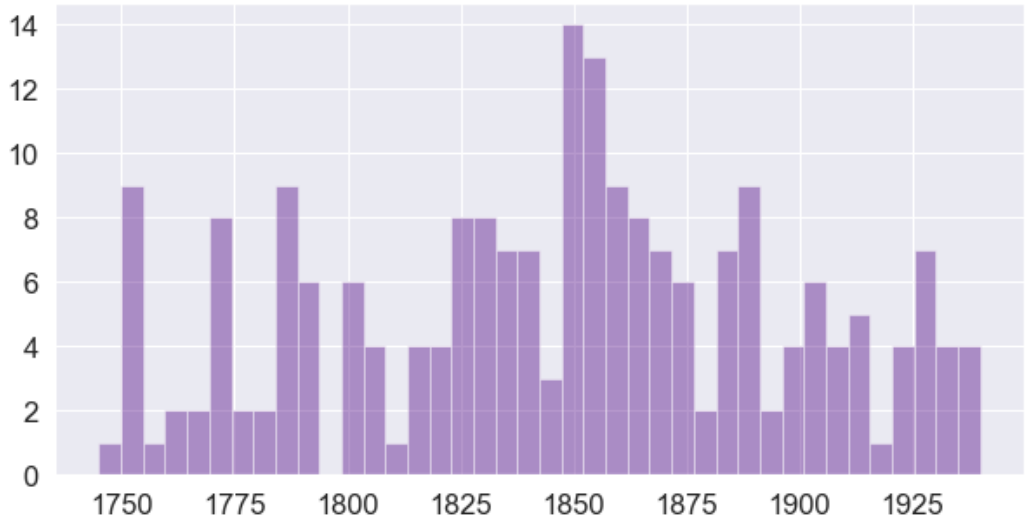

Рисунок . Распределение пьес в Russian Drama Corpus по году написания.

Метод кластеризации отдельных реплик диалогов

Задачей первого этапа работы являлась кластеризация отдельных реплик диалога. Для этого сначала каждая фраза представлялась в качестве вектора, далее получившееся множество векторов разбивалось на кластеры.

За время экспериментов было проверено несколько подходов как к векторному представлению фраз (эмбеддингов), так и к кластеризации. В качестве эмбеддингов были использованы:

- BERT для отдельных слов, без дообучения с использованием текущего датасета [9];
- BERT для отдельных слов, с дообучением с использованием текущего датасета [9];
- Word2vec модель, построенная алгоритмом fastText CBOW на корпусе Araneum [10];
- Convolutional Neural Network (CNN) encoder-decoder модель, обученная на корпусе диалогов из русских пьес;
- кросс-языковой Transformer/CNN-encoder для предложений, разработанный и обученный в Google [11].

В качестве алгоритмов кластеризации использовались KMeans, Affinity Propagation и HDBSCAN. При этом второй и третий оказались хуже и не давали качественных кластеров, поэтому было решено проводить все исследования с помощью метода KMeans.

В приложении 1 можно увидеть примеры случайных кластеров, которые были получены при помощи KMeans для разных векторных представлений.

Из 5 методов хуже всего показали себя **недообученный BERT и дообученный BERT**. По результатам ручной разметки на основе получившихся векторов не удалось выделить кластеры, в которых были бы похожие предложения. Например, в кластере, который находится в приложении, предложения из междометий соседствуют с короткими вопросами и повествовательными предложениями. При этом они не объединены какой-либо общей темой. Из-за этого, а

также других технических ограничений (например, на длину текстов, поступающих в модель) было решено перейти к другому семейству методов.

Гораздо лучше себя показала **модель Word2Vec обученная на корпусе Araneum** (который содержит в том числе и художественные тексты, что делает его в некоторой мере схожим с исследуемыми диалогами). В примере видно, что выделился кластер с предложениями, где есть слово ”значит”. К нему можно отнести различные уточняющие вопросы и ответы. Однако более плотная ручная разметка показала, что некоторые из кластеров нельзя описать или объединить из-за слишком разнородных предложений.

Следующий подход основан на **CNN encoder-decoder модели**. Вектора отдельных слов предложения (полученные из Word2Vec модели, которая использовалась ранее) подавались в сверточную нейронную сеть, которая должна была ”сжать” их до некоторого вектора, а потом получить из него первоначальные вектора. В первую очередь в данной модели для использования были интересны именно эти ”сжатые” вектора, которые должны представлять собой репрезентацию всего предложения. На рисунке 1 можно увидеть схему получившейся нейронной сети.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
conv1d_102 (Conv1D)          (None, 46, 100)           150100
_________________________________________________________________
average_pooling1d_32 (Averag (None, 23, 100)           0
_________________________________________________________________
conv1d_103 (Conv1D)          (None, 19, 100)           50100
_________________________________________________________________
max_pooling1d_9 (MaxPooling1 (None, 9, 100)            0
_________________________________________________________________
conv1d_104 (Conv1D)          (None, 5, 10)             5010
_________________________________________________________________
flatten_23 (Flatten)         (None, 50)                0
_________________________________________________________________
dropout_36 (Dropout)         (None, 50)                0
_________________________________________________________________
dense_49 (Dense)             (None, 100)               5100
_________________________________________________________________
dense_50 (Dense)             (None, 500)               50500
_________________________________________________________________
reshape_23 (Reshape)         (None, 50, 10)            0
_________________________________________________________________
conv1d_105 (Conv1D)          (None, 50, 300)           3300
=================================================================
Total params: 264,110
Trainable params: 264,110
Non-trainable params: 0
_________________________________________________________________
```

Рисунок 1. Схема CNN encoder-decoder для получения векторов реплик.

Данный подход также дал некоторое улучшение. В приложении можно найти кластер, в который собраны фразы, где один из собеседников обращается к другому.

Основным его преимуществом является то, что на прошлых этапах для получения вектора предложения из набора вектора слов приходилось брать некоторую статистику, например, среднее. Из-за этого теряется много информации, которая содержится в самом предложении. Полученный же таким образом вектор содержит больше информации о предложении, за счет которой нейронная сеть пытается восстановить его.

Философия следующего метода векторного представления предложений во многом схожа с философией предыдущего. **Кросс-языковой Transformer/CNN-encoder для предложений** представляет из себя модель, обученную более чем на 8 миллионах предложений на 16 языках. В ее рамках сначала получают две векторные репрезентации предложения: одну с помощью трансформера [13], в другую с помощью сверточной нейронной сети [14]. Далее эти два вектора подаются на вход другой нейронной сети, которая, во-первых, окончательно кодирует исходное предложение, а, во-вторых, решает ту или иную задачу (например, стандартные задачи для SNLI[15]). [12]

По результатам ручной разметки данный подход дал наиболее хорошие результаты - для каждого кластера можно дать корректное описание. В приложении 2 приведены примеры для каждого из 50 получившихся кластеров.

Другим преимуществом именно подобных векторов для предложений является возможность сделать единую кластеризацию на все 16 языков, которые поддерживает данная модель. Например в таблице 1 можно увидеть примеры одного из кластеров на 4 европейских языках. В этот кластер объединились фразы о разговорах: о том, кто или что услышал, или просит, или хочет сказать.

Русский	Испанский	Немецкий	Английский
Покойной, вам легко **сказать!**	¿Y cómo te lo **dijo**?	Na muaß i d’ wahrheit drauß’n **hör’n.**	I **heard** the owl scream and the crickets cry.

Но **скажите** мне, **скажите**, что вы не в состоянии оскорбить женщину.	¡Lo que **digo**!		Did not you **speak**?
И ни слова мне не **скажете**?	¿Hay que **decir** las cosas dos veces?		Well, let's away, and **say** how much is done.
Вы меня не увидите более… Я вам **говорю** — я не переживу этой ночи.	yo no quiero **hablar** porque temo tus intenciones.		**Say**, if th' hadst rather hear it from our mouths,
Скажи, правду, умоляю тебя, **скажи** мне правду…	Hay quien cree que **habló** muchas noches con .		**Hear** his speech, but **say** thou nought.

Таблица 1. Пример кластера "молчание / разговор"в разных языках

В результате данной работы получилось создать модель, которая без помощи обучения с учителем разбивает все фразы корпуса на кластера объединенные семантикой или другими критериями. **Итоговым решением**, данного задачи, которую можно обозначить как кластеризацию реплик или как маркировку вершин диалогового графа, было принято решать с помощью **кросс-языкового Transformer/CNN-encoder для предложений** и метода кластеризации **KMeans**. Данный подход, во-первых, выдает по данным ручной разметки лучшие кластеры, а, во-вторых, позволяет расширить поле исследуемых языков без дополнительных затрат.

Метод кластеризации коротких диалогов

Основным этапом работы являлась кластеризации коротких диалогов. Задача заключается в объединении в отдельные кластеры диалогов из нескольких фраз. Так, получая на вход последовательность из коротких вопросов и ответов, стоит сделать вывод, что она скорее похожа на другую последовательность из коротких вопросов и ответов, чем на диалог, где персонажи просто обмениваются своими рассуждениями на ту или иную тему.

Дальнейший анализ был проведен для диалогов по три реплики. В то же время представленные в работе методы поддерживают возможность для работы и с более длинными диалогами. Всего в исследуемых пьесах содержится чуть более 100 тысяч диалогов из трех реплик.

Для того, чтобы решить данную задачу было решено превратить все диалоги в последовательности, где каждой фразе соответствует кластер полученный на предыдущем этапе работы. Далее было реализовано три подхода к анализу диалогов. Первый подразумевает создание отдельного класса для каждого варианта последовательности кластеров (далее ngram или триграмм для частного случая диалога из трех реплик). Второй подход основывался на графовых методах и векторном представлении вершин графа Node2Vec [7]. Третий подход предполагает создание векторной репрезентации короткого диалога и последующую их кластеризацию.

Для анализа подхода, который **объединяет в классы диалоги по признаку принадлежности их к одному триграмму** рассмотрим 10 наиболее частых триграммов в корпусе русской драмы (таблица 2).

Кластер первой реплики	Кластер второй реплики	Кластер третьей реплики	Количество вхождений
Чувство/ религия/ мифология	Чувство/ религия/ мифология	Чувство/ религия/ мифология	136
Князь/ царь/ король/ власть	Князь/ царь/ король/ власть	Князь/ царь/ король/ власть	113
Короткий ответ	Вопрос-уточнение	Короткий ответ	80

О деньгах	О деньгах	О деньгах	77
Эмоциональное высказывание	Эмоциональное высказывание	Эмоциональное высказывание	69
Мнение о человеке/ людях/ группе лиц	Мнение о человеке/ людях/ группе лиц	Мнение о человеке/ людях/ группе лиц	68
Вопрос-уточнение	Короткий ответ	Вопрос-уточнение	67
Длинное рассуждение (о высоких темах)	Чувство/ религия/ мифология	Чувство/ религия/ мифология	67
Чувство/ религия/ мифология	Чувство/ религия/ мифология	Длинное рассуждение (о высоких темах)	66
Чувство/ религия/ мифология	Длинное рассуждение (о высоких темах)	Чувство/ религия/ мифология	54

Таблица 2. Триграммы реплик с самым большим количеством вхождений в корпус.

Стоит заметить, что некоторым кластерам реплик свойственно делать переход в самих себя. Например, когда персонажи начинают рассуждать о какой-то теме, то далее весь диалог проходит в рамках одного кластера. В приведенных кластеров в 5 из 10 все три реплики соответствуют одному и тому же кластеру. Это как диалоги на резигиозно-мифологическую тему или обсуждение власти, так и более бытовой диалог о деньгах.

Также можно увидеть триграммы, в котором более одного кластера, состоящие из короткого ответа, вопроса и второго короткого ответа или из вопроса-уточнения, короткого ответа и второго вопроса-уточнения.

Приведем пример того, как выглядят последовательности фраз из одного кластера. Для краткости возьмем два диалога о деньгах:

«– Ан, нет, судыр... я вам лучше скажу... весь свет вместе и кажда человек поодиначке больше всего любыт деньга...

– Ха-ха-ха! Так и ты, значит, больше всего любишь деньги?

– Вестымо, судыр. Кто деньга любыт, тот, значит, всё хорошее любыт, потому что на деньга можно достать что есть наилучшего в свете, судыр.»

Диалог Кочергина и Татарина
из пьесы ”Актер” Некрасова Н.А.

Второй пример диалога из этого же триграмма:

«– Такой суммы, ей-Богу, нет. А нет ли у вас, Петр Иванович?

– При мне-с не имеется, потому что деньги мои, если изволите знать, положены в приказ общественного призрения.

– Да, ну если тысячи нет, так рублей сто.»

Диалог Бобчинского, Добчинского и Хлестанова
из пьесы ”Ревизор” Гоголя Н.В.

Также существуют последовательности вопросов и ответов. Например кластер ”короткий ответ, вопрос-уточнение, короткий ответ”, который можно проиллюстрировать следующим диалогом:

«–Играет-с.

–В преферанс?

–В свои козыри-с.»

Диалог Маши, Фонка и Пряжкиной
из пьесы ”Холостяк” Тургенева И.С.

Данный путь является с одной стороны самым простым и в некотором смысле довольно точным, так как триграммы состоят из очень похожих типов диалогов. Однако он производит слишком большое количество ”кластеров” диалогов. Так как у нас нет ограничения на то, чтобы после одного кластера обязательно следовал другой (а кластеры часто повторяются из реплики в реплику, как это видно по нашим примерам), всего может существовать 125 000 разных триграммов. В данном корпусе всего можно найти 53877 различных последовательностей из трех различных кластеров реплик. Из них более половины встречаются только один раз, а триграммов встречающихся более 5 раз более 2000. На рисунке 2 можно увидеть гистограмму триграммов по количеству вхождений.

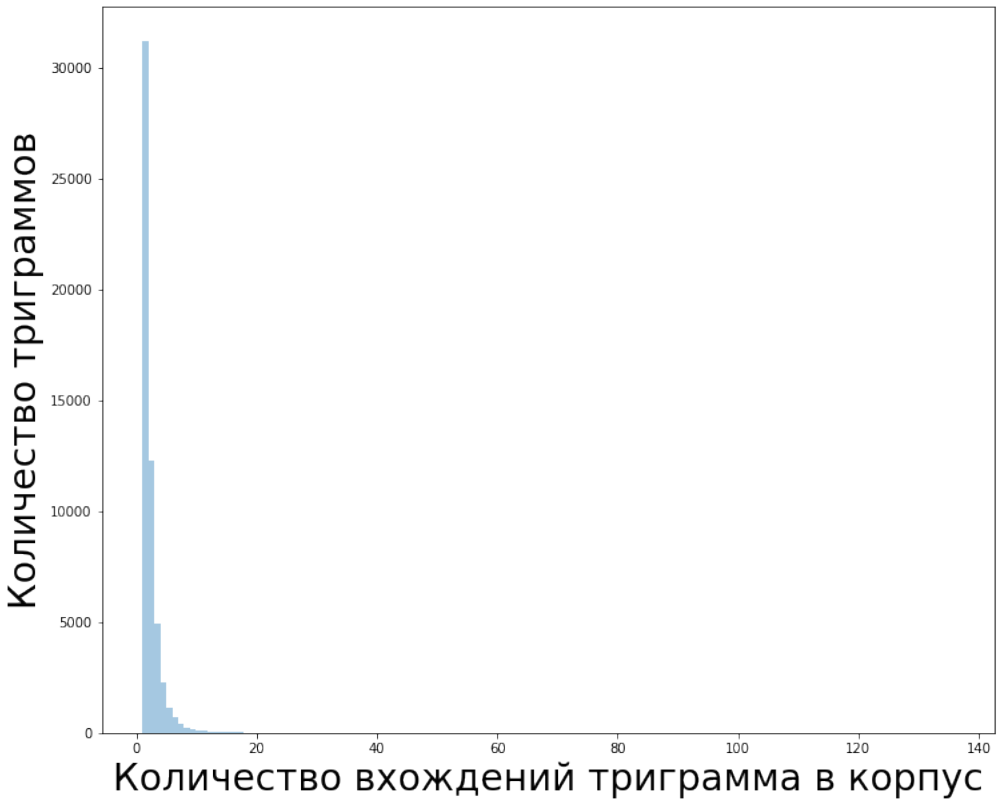

Рисунок 2. Распределение триграммов по количеству вхождений.

В сумме для того, чтобы разметить хотя бы 10% всех диалогов входящих в корпус, необходимо проанализировать чуть более 900 триграммов. Это делает данный метод малоприменимым. Количество данных для ручной разметки становится очень большим, а количество классов не позволяет использовать их в дальнейшем.

Следовательно, необходимо реализовать подход, который будет объединять похожие диалоги в меньшее количество отдельных кластеров. Для этого было решено создать векторные представления для коротких диалогов и кластеризовать уже их.

Для того, чтобы получить векторное представление кластеров реплик, было решено испробовать **графовые методы**. Корпус был представлен, как ориентированный взвешенный граф. В качестве вершин находятся кластера фраз, а вес ребра идущего от одного класса к другому соответствует тому, сколько раз после первого класса следовал второй в итоговой выборке.

Для векторного представления вершин графа существует множество методов. В рамках данной работы был использован подход Node2vec, который является реализацией схож с методов CBoW для получения эмбеддингов слов. Для векторного представления в нем берется выход внутреннего слоя нейронной се-

ти, которая пытается предсказать вершину графа, по ее соседям по случайному блужданию по графу. Для перехода от вектора реплики к вектору диалога из трех реплик брался вектор средних. Далее они кластеризовалис с помощью метода KMeans.

К сожалению, из экспериментов по данному подходу не получилось создать сколько-нибудь значимые кластера диалогов. По результатам разметки выяснилось, что в кластерах не наблюдается внутренней логики. Например, если взять случайный кластер, то в нем можно найти как последовательности из кластеров про свадьбу (триграмм ”Свадьба/ жених/ невеста” , ”Свадьба/жених/невеста” , ”О себе”), так и про власть (триграмм ”Обращение/ восклицание” , ”Князь/ царь/ король/ власть” , ”Князь/ царь/ король/ власть”).

При этом, если кластер пример из диалога про выглядит следующим образом:

«– А к тому-то оно, что я хочу жениться; а это счастие сделаю тебе и учиню тебя участницею моего имения и моего сердца.

– Едакой женитьбе и куры смеяться станут; мне семнадцать лет, а вам семьдесят.

– Да я так бодр, как лучше быть нельзя, и молодого детину заткну за пояс.»

Диалог Чужехвата и Нисы
из пьесы ”Опекун” Сумарокова А.П.

То пример из диалога про власть уже совершенно другой:

«– В оковы, воины!

– Все пра́ва разрушаешь И князя моего величье оскорбляешь.

– Коль хочешь, возвратись ко князю твоему, Как мало я его страшусь, сказать ему.»

Диалог Христиерна и Любомира
из пьесы ”Росслав” Княжнина Я.Б.

После ручной проверки как в этом, так и в оставшихся кластерах не получилось найти зависимостей. Данный подход не совсем подходит для решения данной задачи. По мнению автора работы это связано с тем, что, во-первых, у модели нет информации о семантике анализируемых кластеров, а, во-вторых, вектор средних от векторов реплик не является в полной мере вектором самого диалога.

В дальнейшем в работе по классификации коротких диалогов было решено отказаться от готовых графовых методов и перейти к ручному созданию нейронных сетей для решения задачи. В целом, языковая модель, которая будет представлена далее, может быть интерпретирована в терминах графовой модели (как векторное представление подграфа), однако при ее создании не было опоры на теорию графов и современные методы работы с ними.

При переходе к анализу **языковой модели на рекуррентной нейронной сети** стоит отметить, что те или иные кластера предложений могут быть более или менее похожи друг на друга. Например, на рисунке 3 можно увидеть, что кластера, в которых реплики представляют из себя длинные рассуждения, содержат мысли о власти или выражают высокие (в том числе, мифологические и религиозные) чувства, семантически ближе друг к другу, нежели кластеры, в которых спрашивается или рассказывается о передвижении в пространстве.

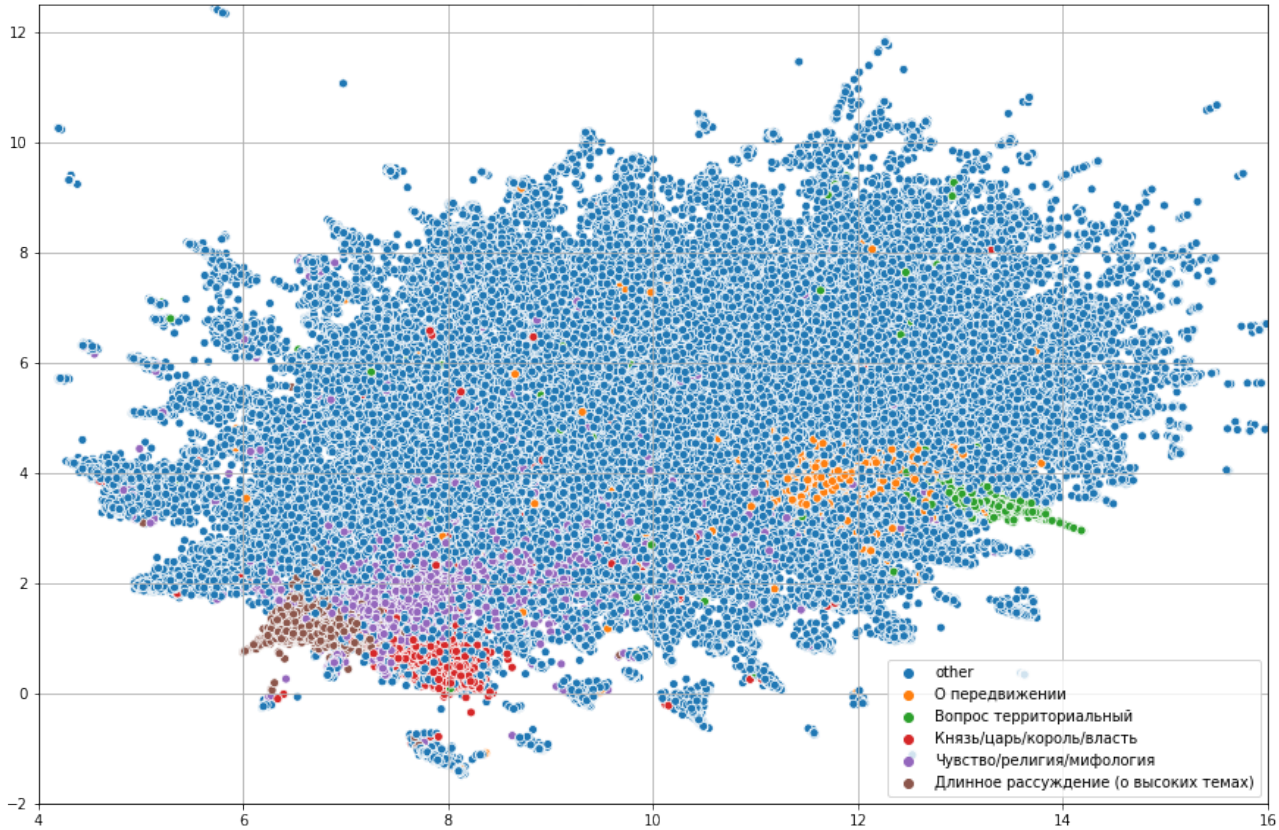

Рисунок 3. Распределение векторов предложений, размерность которых была снижена с помощью метода UMAP [16]

Разработанная для получения векторов диалогов языковая модель, получает на вход в том или ином виде последовательность кластеров и предсказывает, какой кластер будет в этой последовательности следующим. После обучения из

данной нейронной сети можно удалить последний слой и вместо кластера следующей реплики она будет выдавать вектор, который будет схожим для похожих последовательностей кластеров.

В рамках работы были обучены и проанализированы два варианта данной нейронной сети. Они различаются по тому, как были представлены кластеры на входе в модель. В первом варианте на вход ей подавались one-hot вектора, где все значения, кроме одного, были равны нулю. Оставшееся значение было равно единице и его индекс был равен номеру кластера реплики. Подобные вектора перед попаданием в LSTM модуль сети проходили через Embedding слой, который трансформировал их в вектор, который должен был быть похожим у похожих кластеров.

Второй вариант предполагает отсутствие предварительного Embedding слоя в начале нейронной сети. В нем на вход модели подается не соответствующий кластеру one-hot вектор, а его центроид. Благодаря этому на вход модели подаются данные о семантике входящего предложения и ей не надо решать дополнительную задачу по получению ”векторной репрезентации кластера”. На рисунке 4 можно увидеть схему этого второго варианта архитектуры нейронной сети.

```
_________________________________________________________________
Layer (type)                 Output Shape              Param #
=================================================================
lstm_2 (LSTM)                (None, None, 300)         975600
_________________________________________________________________
dropout_4 (Dropout)          (None, None, 300)         0
_________________________________________________________________
time_distributed_2 (TimeDist (None, None, 128)         38528
_________________________________________________________________
dense_9 (Dense)              (None, None, 64)          8256
_________________________________________________________________
dropout_5 (Dropout)          (None, None, 64)          0
_________________________________________________________________
dense_10 (Dense)             (None, None, 32)          2080
_________________________________________________________________
dense_11 (Dense)             (None, None, 50)          1650
=================================================================
Total params: 1,026,114
Trainable params: 1,026,114
Non-trainable params: 0
_________________________________________________________________
```

Рисунок 4. Схема финального варианта языковой модели на рекуррентной нейронной сети для кластеров реплик.

При этом, за счет архитектуры (в начале сети идет Long short-term memory (LSTM) модуль), языковая модель для решения задачи классификации использует не только информацию о кластере последней реплики, но и информацию о предыдущих кластерах, если это необходимо. Более того, благодаря рекуррентности, данная архитектура позволяет достаточно легко перейти от кластеризации триграммов реплик (которые были наиболее полно разобраны в данной работе) к кластеризации последовательностей из двух, четырех, пяти и более реплик.

После обучения языковой модели из нее был удален последний слой в котором делалось финальное предсказание. Далее для каждого короткого диалога был получен вектор, обучена методом KMeans модель кластеризации и получены кластера для коротких диалогов.

Ручная разметка показала, что если нейронная языковая модель, на вход которой подается one-hot вектора, не показывает себя значительно лучше, чем Node2Vec подход. Однако кластеризация векторов из модели, получающей центроиды кластеров фраз дает уже гораздо лучший результат. Далее приводятся примеры различных кластеров для русского языка.

Например, в кластер ”диалоги с территориальными вопросами”собрано множество диалогов, где тем или иным образом участвуют вопросы о местонахождении или передвижении людей. Это может быть как триграмм ”Вопрос” , ”Короткий ответ” , ”Вопрос территориальный”:

«– Ты, Лёв, в лавку пойдешь?

– Нет, уж я забрался.

– Дома, что ль, будешь?.»

Диалог Архипа и Краснова
из пьесы ”Грех да беда на кого не живет” Островского А.Н.

А может быть триграмм ”О передвижении” , ”Вопрос территориальный ” , ”Вопрос территориальный”:

«– А раз добрый, пойдем со мною, Яшенька.

– Куда?

– Куда-нибудь.»

Диалог Маргариты и Якова
из пьесы ”Океан” Андреева Л.Н.

Также благодаря тому, что векторная репрезентация фраз едина для 16 языков, есть возможность рассмотреть этот же кластер и в других языках. Например, диалог из этого кластера из драмы на испанском языке:

«– ¡Su nombre!

– ¡El!

– ¡El! ¿Ha estado aquí el?... ¿Pero dónde está, dónde?»

Диалог Томаса и Пласидо
из пьесы "A fuerza de arrastrarse" Хосе Эчегарай-и-Эйсагирре

Или другой пример того же кластера на английском языке:

«Here is a place reserved, sir.

Where?

Here, my good lord. What is't that moves your highness?»

Диалог Леннокса и Макбета
из пьесы "Макбет" Уильяма Шекспира

По итогам работы над кластеризацией коротких диалогов было получено два значимых результата.

Во-первых, были проанализированы **триграммы реплик**. Те диалоги, у которых совпадают кластеры всех реплик, очень похожи друг на друга по структуре. И хотя у данного способа есть определенные ограничения, связанные с тем, что количество триграммов очень велико, существуют варианты анализа наиболее частых последовательностей для первоначального знакомства с новым корпусом.

Во-вторых, с помощью **языковой модели на основе рекуррентной нейронной сети** был реализован подход к кластеризации различных триграммов. Данный подход позволяет выделить меньшее, ограниченное количество кластеров диалогов, что может быть полезно для будущей работы с корпусом.

Обсуждение результатов

Первоначальное накопление данных является проблемой для множества разработок в области искусственного интеллекта и, в частности, автоматической обработки текстов. Как в коммерческой, так и в академической деятельности размеченные данные представляют из себя огромную ценность. В том числе, ценностью являются размеченные диалоги, а так же явно или не явно выделенные структуры диалогов. [1]

В рамках данной работы были выработаны и оценены различные подходы к решению двух задач: кластеризации отдельных реплик и кластеризации коротких диалогов.

Для **решения задачи по кластеризации реплик** были испробованы различные методы векторного представления предложений: нейронные сети архитектуры BERT; эмбеддинги Word2Vec; сверточная нейронная сеть, кодирующая предложение в вектор; кроссязыковой Transformer/CNN-encoder для предложений. В результате:

- подход к решению данной задачи с помощью BERT оказался наименее успешным (получившиеся кластера не связаны синтаксически или семантически);
- лучше себя показали эмбеддинги Word2Vec (по результатам ручной разметки для большей части кластеров получилось подобрать описание, описывающее кластер);
- следующим в порядке улучшения качества оказался CNN-encoder, обученный на текстах самих диалогов;
- **лучший результат** в рамках данной работы был получен с помощью кроссязыкового Transformer/CNN-encoder для предложений (в рамках ручной разметки все кластеры оказались достаточно семантически связными и получили описание).

В качестве дополнительного преимущества важным является возможность для **работы** с помощью результирующей модели сразу **с множеством языков**. Кроме русских пьес были проанализированы и отрывки из пьес на английском,

немецком и испанском языках. В результате данного анализа можно сделать вывод о том, что полученные описания для кластеров валидны и для других языков. При этом важно отметить, что при работе с пьесами на иностранных языках, полученная модель при обучении ”не видела” ни этих текстов, ни данных языков как таковых.

На следующем этапе работы анализировались **короткие диалоги из трех реплик**. Для анализа использовалось три подхода: анализ триграммов; векторные представления графов; языковые модели, основанные на рекуррентных нейронных сетях. По итогам анализа:

- графовые методы (кластеризация векторов, полученных с помощью метода Node2Vec) не показали значимых результатов - кластера, полученные с их помощью, не обладают какой-либо семантической или синтаксической связью;

- анализ триграммов диалогов показал, что может быть полезен в определенных целях - анализ частотных триграммов помогает получить представление о датасете и выделить наиболее стандартные шаблоны диалогов в корпусе. Однако, количество триграммов очень велико и не решает проблему необходимости ручной разметки большого количества текстов;

- языковые модели на рекуррентных нейронных сетях позволяют генерировать ограниченное количество кластеров, которые, по данным ручной разметки, объединяют семантически и/или структурно схожие диалоги. При этом важным является то, что для лучшей работы на вход языковой модели необходимо подавать не просто one-hot вектора, а центроиды кластера, в которых находится информация о том, какие фразы в них содержатся.

Далее полученные кластеры могут использоваться в различных целях.

С одной стороны, это может быть **полуавтоматическая разметка произвольного корпуса** диалогов по намерениям (intent), как отдельных фраз, так и выделенных коротких диалогов. Для этого необходимо обучить модель на интересующем корпусе диалогов или на схожем корпусе, а после вручную разметить сами кластера - выделить то, по какому признаку полученные реплики или диалоги объединились.

При решении подобной задачи может быть полезным кроссязыковое свойство модели - если корпус, с которым проводится работа является недостаточно

большим по количеству реплик или диалогов, то можно найти похожий корпус на языке, отличном от оригинального. При этом, если найденый корпус на иностранном языке является размеченным и есть интерес в переносе подобных текстов на исходные диалоги, имеет смысл от методов обучения без учителя перейти к методам обучения с учителем при помощи все той же модели для получения векторной репрезентации реплик диалогов.

Также полученные результаты могут найти применение в **цифровых гуманитарных исследованиях (digital humanities)**. Реализуя с помощью полученных кластеров метод ”дальнего чтения”, можно проанализировать то, как меняется стиль и тематика пьес от года к году или от автора к автору.

Например, мы можем рассмотреть то, как менялись частоты кластеров от одной эпохи к другой. Для этого построим тепловые карты частотности кластеров, где на вертикальной шкале будут отложены кластеры, а по горизонтальной шкале будут идти части пьес (5 частей, где первая это начало пьес, а последняя конец пьес). Для удобства чтения карта была разделена на две части, с разным набором кластеров. На рисунке 5 можно увидеть такую карту для пьес 1725-1775 годов.

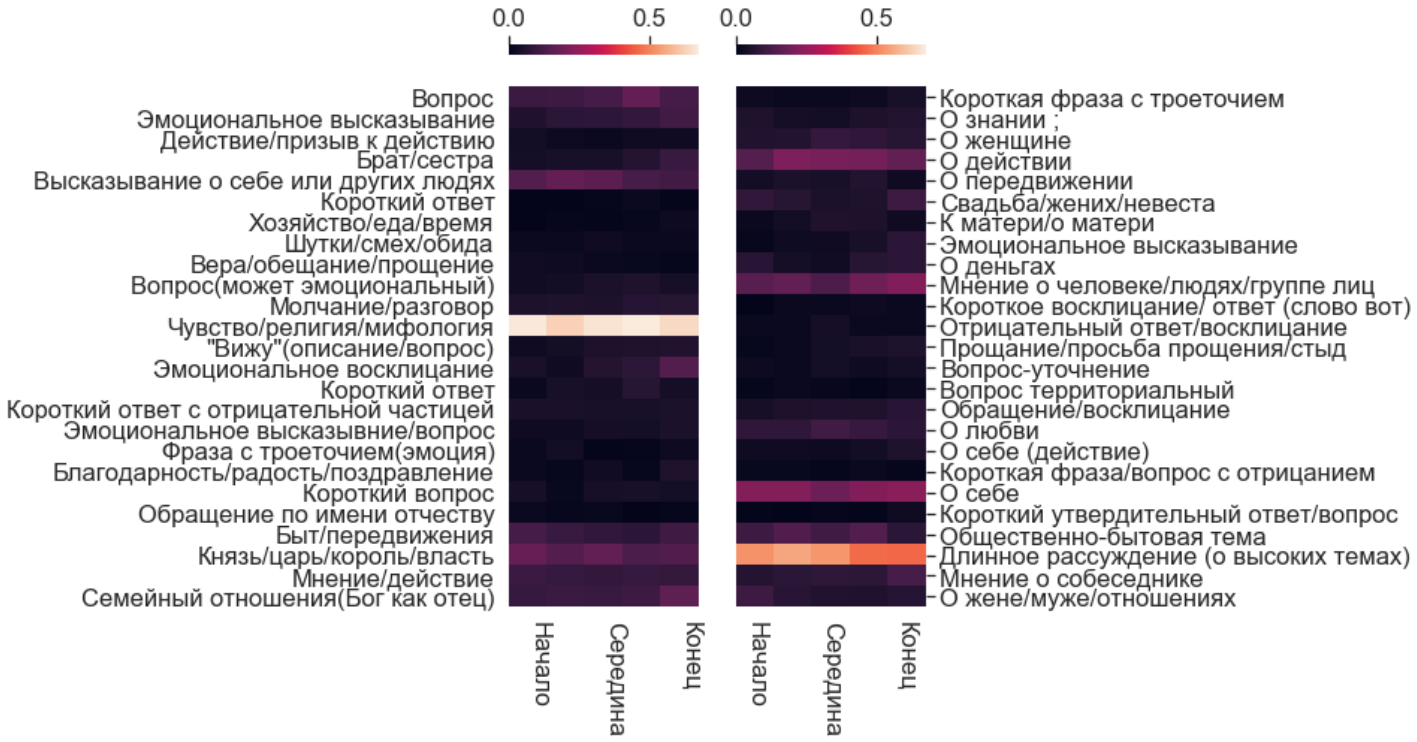

Рисунок 5. Тепловая карта кластеров фраз в пьесах 1725-1775 годов.

Здесь можно видеть, что эти пьесы (в основном, за авторством Сумарокова

А.П. и Княжнина Я.Б.) наполнены фразами с рассуждениями, в том числе написанных ”высоким стилем” (они объединились в кластер под названием ”чувство/религия/мифология”). В то же время бытовая тема и короткие вопросы или ответы представлены довольно мало. Эта ситуация меняется с течением времени. Если рассматривать тепловую карту для следующего периода в 50 лет (рисунок 6), то можно увидеть, то как появляются общественно-бытовые темы, а частота фраз написанных ”высоким стилем” снижается.

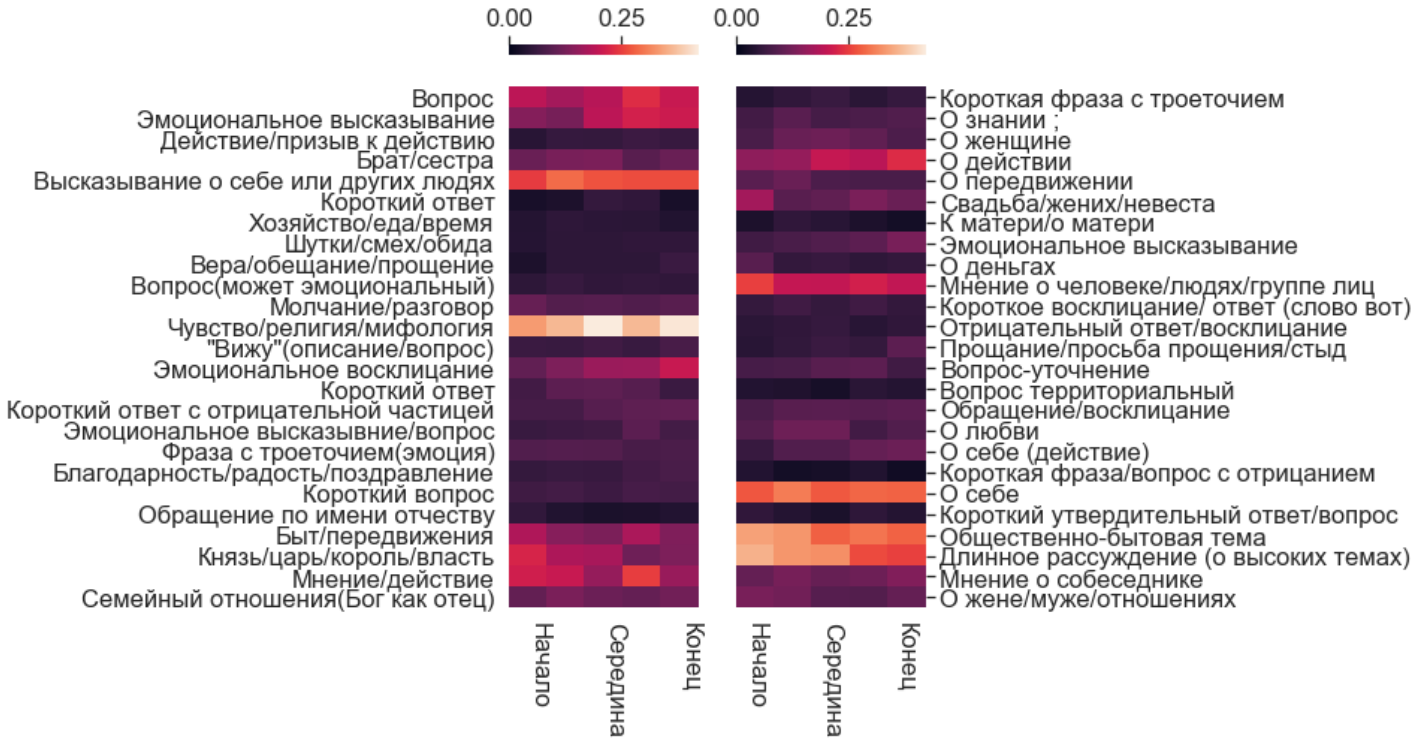

Рисунок 6. Тепловая карта кластеров фраз в пьесах 1775-1825 годов.

При этом тенденция с течением времени идет все дальше в этом направлении. После 1825-ого года (рисунок 7) написанных ”высоким стилем” фраз становится настолько мало, что этот кластер перестает выделяться на фоне остальных, как делал на протяжении 100 лет. В то же время количество реплики на общественно-бытовые темы и высказывания о себе или других людях становятся гораздо более частыми. В целом можно увидеть, что тепловая карта становится более однородной, что связано как с изменениями в обществе, языке и стиле, так и с увеличением количества авторов, участвующих в творческом процессе (если в первом рассматриваемом периоде их 7, то во втором и третьем уже 20 и 18 соответственно).

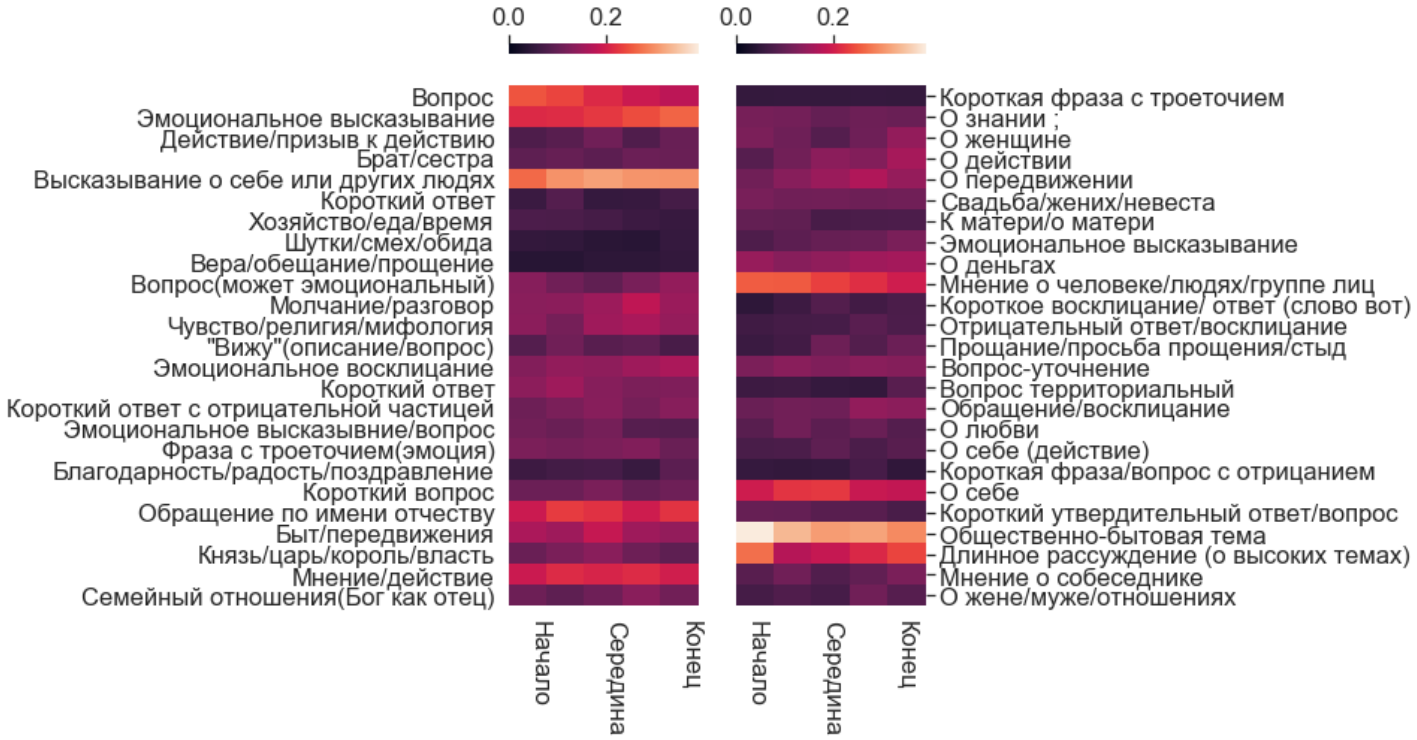

Рисунок 7. Тепловая карта кластеров фраз в пьесах 1825-1875 годов.

Эта же тенденция наблюдается и в конце XIX - начале XX веков (рисунок 8).

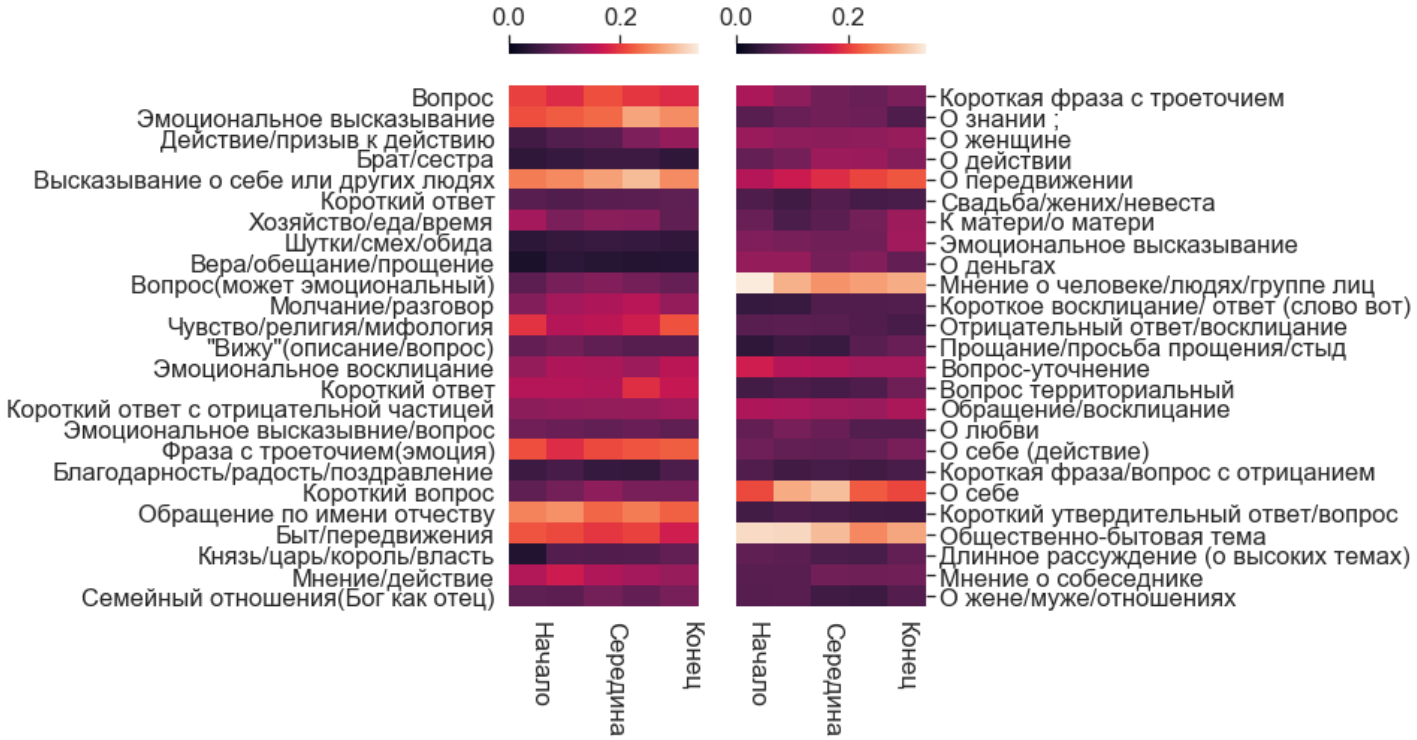

Рисунок 8. Тепловая карта кластеров фраз в пьесах 1875-1925 годов.

С каждым описываемым периодом можно увидеть, как увеличивается в корпусе частота кластеров, в которых фразы посвящены личным отношениям и

общественно-бытовым темам, увеличивается число эмоциональных восклицаний и рассказов о себе. Практически отсутствующий до 1825 года кластер с обращениями по имени-отчеству становится все более ярким (иногда даже самым частым в частях пьес за авторством, таких драматургов как Тургенев, Мамин-Сибиряк, Салтыков-Щедрин, Андреев). Можно увидеть каким разнообразным становится язык и стиль текстов, с течением времени, увидеть смещение фокуса к более активным и коротким фразам от длинных рассуждений.

Также данный инструмент можно использовать для сравнения между собой авторов. У некоторых, таких как, например, Горький, Екатерина II, Глинка, Толстой А.К., в текстах превалирует один из кластеров. Например, по тепловой карте для творчества Алексея Константиновича Толстого (рисунок 9) можно увидеть, что он писал в первую очередь о людях, наделенных некоторым титулом. И это легко доказать тем, что в корпусе он представлен тремя произведениями: «Смерть Иоанна Грозного», «Царь Фёдор Иоаннович», «Царь Борис».

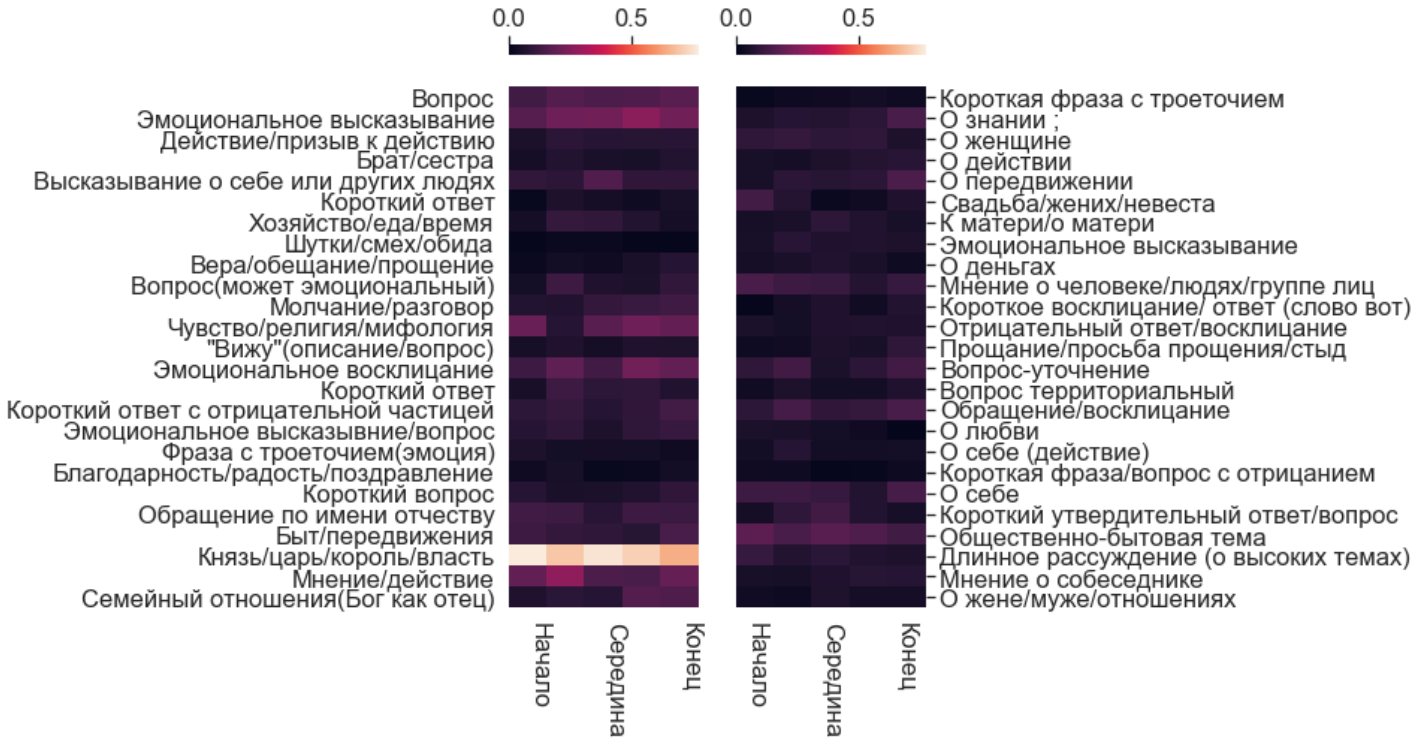

Рисунок 9. Тепловая карта кластеров фраз в пьесах Алексея Константиновича Толстого.

У других авторов наблюдается большее разнообразие. Например, к таким драматургам относится Антон Павлович Чехов. Как можно увидеть на тепловой карте (рисунок 10) в его творчестве нет какого-то кластер, который выглядит

сильно ярче других. Да, общественно бытовая тема является наиболее частотным, однако параллельно выделяются и более личные кластеры с высказываниями о себе и других людях. К тому же можно заметить, что в начале пьес общество-бытовая тема выражена гораздо больше, а в конце учащаются различные кластеры с эмоциональными фразами, которые связаны с активным действием персонажей.

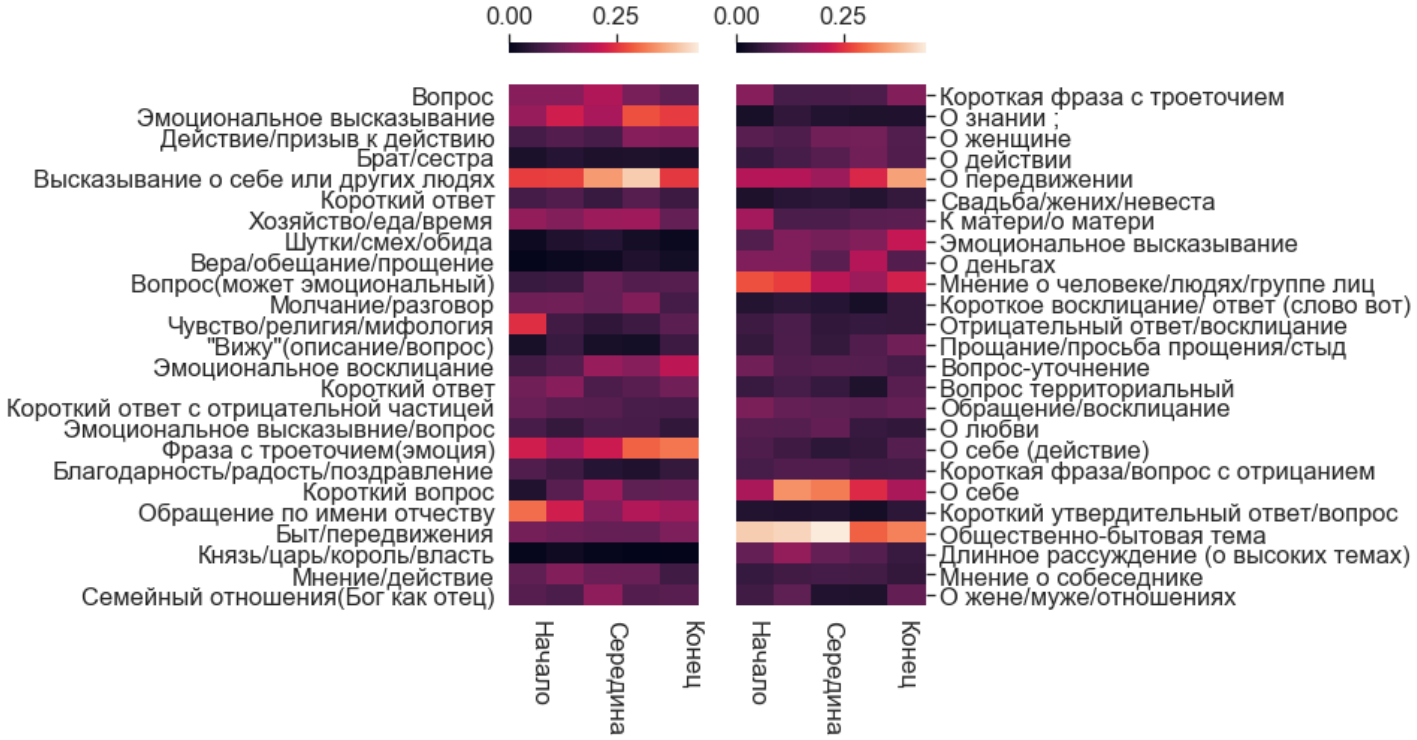

Рисунок 10. Тепловая карта кластеров фраз в пьесах Антона Павловича Чехова.

Помимо российских драматургов интересно рассмотреть и примеры из иностранной драматургии. При этом было решено одновременно посмотреть на то, как могут выглядеть карты для разных пьес одного автора. В качестве примера был выбран Уильям Шекспир и его пьесы «Ромео и Джульетта»(рисунок 11) и «Макбет» (рисунок 12).

Можно заметить несколько ключевых моментов, общих для обеих пьес. Во-первых, как и в русских пьесах XVIII века и начала XIX века большая часть фраз были определены в кластер ”чувство/религия/мифология”, связанный с поэтичным и возвышенным стилем драматургии того времени. Во-вторых, в отличии от русских пьес середины XVIII века у Шекспира уже заметно присутствие общественно-бытовой темы, видны кластеры эмоциональных фраз.

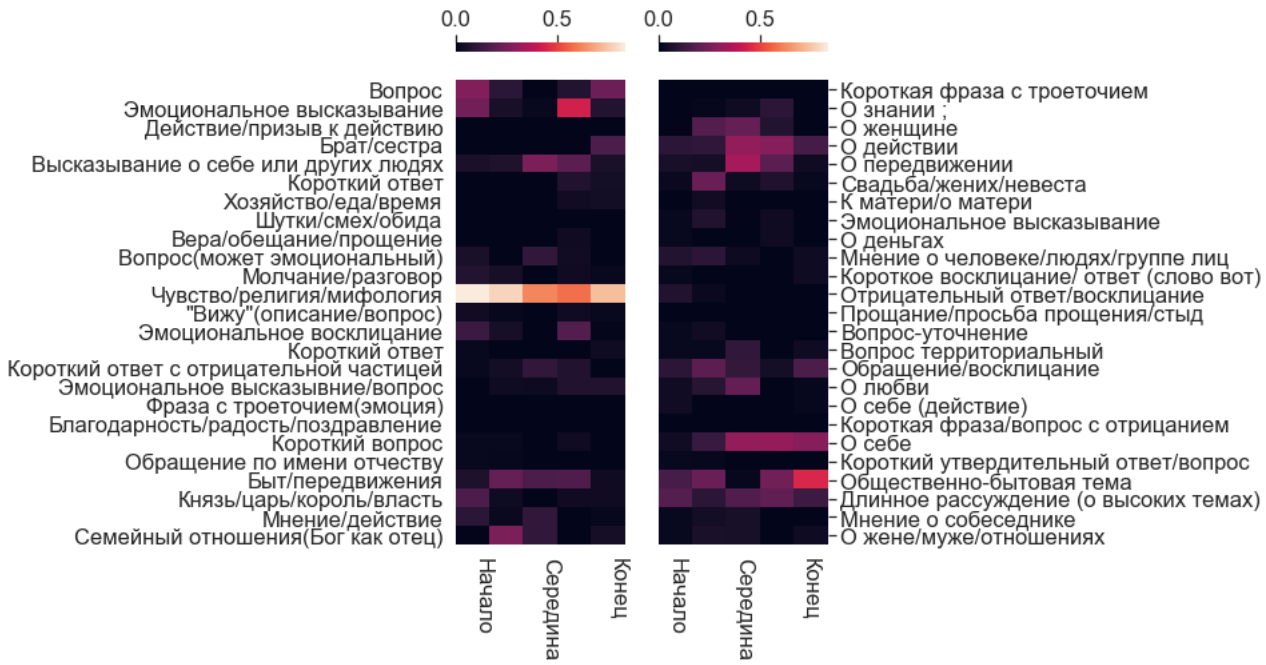

Рисунок 11. Тепловые карты кластеров фраз в пьесе «Ромео и Джульетта».

При этом заметны и различия между картами. Это могут быть как достаточно очевидные вещи, что в «Ромео и Джульетте» говорят о свадьбе и любви, а в «Макбете» нет, так и то, что персонажи первой пьесы более направлены на себя и чаще говорят о своих чувствах и о своем положении, нежели о других людях. Также стоит заметить, что ”Ромео и Джульетта”более активная пьеса - в ней больше эмоциональных фраз, чаще говорят о действиях и передвижениях, при этом практически не поднимают тему власти.

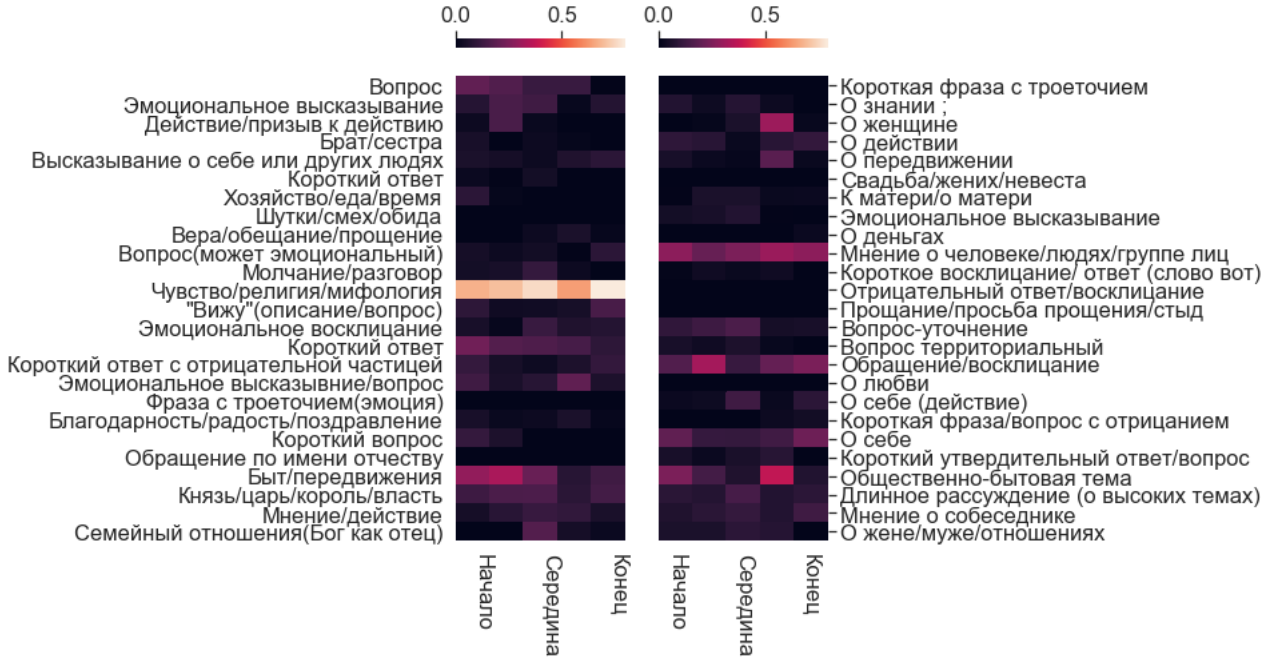

Рисунок 12. Тепловые карты кластеров фраз в пьесе «Макбет».

Беглый анализ того как меняются распределения кластеров между историческими периодами, авторами и пьесами одного автора позволяют сформировать тематический и стилистический образ для них. При этом данный метод может позволить проанализировать большое количество различных пьес, больше, чем содержится в корпусе, использованном для этой работы, выделить тренды тех изменений, что происходят в драматургии и отличия между авторами.

Также данный инструмент может быть использован для сравнения пьес из разных языков без поправки на влияния переводчика. Общие кластера позволяют анализировать пьесы на 16 поддерживаемых языках ”в общей шкале” и сравнивать то как изменялись тематики в творчестве драматургов разных стран.

В результате работы была получена система, которая при помощи обучения без учителя автоматически размечает по тематикам и типам, как отдельные фразы, так и короткие диалоги. Последняя версия системы состоит из 4 частей:

- кроссязыковой Transformer/CNN-encoder для получения векторов фраз;
- кластерная модель для фраз, построенная методом KMeans;
- языковая модель для кластеров, построенная на рекуррентных нейронных сетях;
- кластерная модель для коротких диалогов, построенная методом KMeans.

Данная система может быть использована, с одной стороны, в проектах по построению чат-ботов и решению различных задач связанных с диалогами в качестве метода для полуавтоматической разметки диалогов или источника дополнительных характеристик (**features**). С другой стороны, свое применение полученная модель может найти и в анализе пьес при проведении цифровых гуманитарных исследований, для анализа тематического состава пьес в различных срезах (по авторам, годам и странам). Важной особенностью модели является ее кроссязыковой функционал, позволяющий одновременно проводить анализ диалогов на разных языках и создавать для них единые кластеры.

При этом текущие результаты являются шагом на пути к будущим исследованиям. Для развития модели можно двигаться в различных направлениях, и фактически отдельно развивать каждую из 4 частей представленной модели. Это могут быть как новые подходы к векторизации фраз и диалогов, так и различные методы кластеризации полученных векторов. При этом наиболее

перспективным местом для развития является векторное представление коротких диалогов, которое нуждается в большем количестве как экспериментов так и данных для обучения. Также желательно в будущем собрать больше данных (пьес) и расширить область языков доступных модели в наборе данных для обучения.

Заключение

Список литературы

[1] https://arxiv.org/pdf/2005.02233.pdf

[2] https://www.aclweb.org/anthology/D18-1072.pdf

[3] https://arxiv.org/pdf/2005.02233.pdf

[4] https://arxiv.org/pdf/1810.04805.pdf

[5] https://arxiv.org/pdf/1808.03314.pdf

[6] https://tfhub.dev/google/universal-sentence-encoder-multilingual-large

[7] https://cs.stanford.edu/ jure/pubs/node2vec-kdd16.pdf

[8] https://dracor.org/rus

[9] http://docs.deeppavlov.ai/en/master/features/models/bert.html

[10] https://rusvectores.org/ru/models/

[11] https://tfhub.dev/google/universal-sentence-encoder-multilingual-larg

[12] https://arxiv.org/abs/1907.04307

[13] http://papers.nips.cc/paper/7181-attention-is-all-you-need.pdf

[14] https://www.aclweb.org/anthology/D14-1181.pdf

[15] https://nlp.stanford.edu/projects/snli/

[16] https://arxiv.org/pdf/1802.03426.pdf

Приложения

Приложение 1. Таблица с примерами кластеров для различных методов векторного представления предложений.

Метод	Пример кластера
BERT для отдельных слов, без дообучения с использованием текущего датасета	Что ж за барин, коли уж пенсиона слуге не выдаст за службу? Что ж мне лгать? Досадно. Неизвестно. Хи, хи, хи! Барин ушел, чего бы, кажется, лучше, — нет, сейчас привалит этот черт, брюхач-дворецкий. Ведь вы совсем подлец после этого, Григорий Павлович. Почему ж не заснуть? всего два, три каких-нибудь подсвечника вычистить. Оно примерно, вот изволите видеть, складчина.
BERT для отдельных слов, с дообучением с использованием текущего датасета	Вставай, Лизанька; да ну же, вставай!.. не надо!.. Таким же образом, как мы поручены ему Валерий и брат его. А презренны из них только те, которые этого недостойны имени. Даром-то, что ты ее теперь видела, однако она в жестокой горячке и бредит, и в уме совсем повредилася. А я, за келью между нами молвить, к Богу-то никакого усердия не имею и в этом вам, как добрый человек и православный христианин, чистосердечно признаваюсь. Что с невежей и говорить.

	Обращения ты никакого не знаешь, как есть дура, невоспитанная! Зачем ты ушел? Миша!
Word2vec модель, построенная алгоритмом fastText CBOW на корпусе Araneum	Так вот она что значит, смерть-то! Вот это значит: прямо Писание исполнить! Кашлять перестала, значит. Что это значит всё? Что значит видеть свет! Вот и, значит, грех. Развращаете, значит, понемножку. Раз поет, значит, все хорошо. Это что же значит? Что значит это «уже»?
Convolutional Neural Network encoder-decoder модель, обученная на корпусе диалогов из русских пьес	Будет вам проказничать-то, уморили со смеху! Поцелуемтесь. Что за грабеж, а ступайте с богом, вот и все тут. Государь, мы тебя не узнали, Не суди же покорных рабов, Но скажи, чтобы мы разметали Этих низких и злобных волков. Приданое? — Какая ужасная весть! Верно, выдают Лизу! Надобно все узнать и во что бы то ни стало разбить эту свадьбу. Позвольте сударыня, изъявить вам мою радость... Мерси. Не откажусь. Жарко, знаете, а тут все на ногах да на ногах. Ну, что, брат, нравится ль? Что вы говорите? Как вы, сударь, можете?.. Помню, помню, ваше сиятельство: ты княжна Тройкина и, проживши девичьи лета, нейдешь замуж для того, что надеешься быть графинею.

Кросс-языковой CNN-encoder для предложений	Вы уж лучше простите меня; помиримтесь как следует, тогда я сам останусь. Тысячу раз прошу прощения, если я обеспокоил вас собою Особливо перед вами, сударыня, я так виноват! Прости, колечко золотое! Ошибаюсь? Вы говорите, что я ошибаюсь? Максим! Максим! Пусти, я буду каяться! Прощай, денежки! Ох, эти трагики! Благородства пропасть, а смысла никакого. Ну постойте, я вас помирю. Дерби кто взял? Это, конечно, очень неприятно. Я вполне вам сочувствую. Но что делать? Я никак не мог уехать без того, чтобы не засвидетельствовать вам лично моей благодарности... и не извиниться перед вами. Он нынче ж побывает. Прощайте.

Приложение 2. Таблица с примерами каждого кластера получившегося в результате применения кросс-языкового encoder’а.

Название кластера	Случайный пример
Вопрос	Что бы вы сделали, прекрасная сеньйора?
Эмоциональное высказывание	Как?.. Стой… держи… трррррррр… молодой человек… Врешь, старуха!..
Действие/ призыв к действию	Да помилуйте…
Брат/ сестра	Что тебе, братец, до моего чина? Какого ни есть.
Высказывание о себе или других людях	Вообразите себе, что вы одним вашим присутствием даруете другому человеку, то есть мне, — такое блаженство, какое… словом — высочайшее блаженство… Не будьте же жестоки, останьтесь, умоляю вас.
Короткий ответ	Да.
Хозяйство/ еда/ время	К крайнему моему сожалению, я не в праве отложить наш разговор до завтра… Не угодно ли вам выпить стакан воды?
Шутки/ смех/ обида	Стыдитесь, сударь, стыдитесь! Если б я была мужчиной, вы бы не дерзнули смеяться надо мной!
Вера/ обещание/ прощение	О, я не сомневаюсь!
Вопрос (может эмоциональный)	Помилуйте, что вы делаете?!
Молчание/ разговор	Покойной, вам легко сказать!
Чувство/ религия/ мифология	Послушайте. Вы меня не знаете. Вы не знаете, какими опасностями я пренебрегал, как часто я жертвовал честью, жизнию, — и все для того, чтоб хоть изредка, хоть издали увидеть вас, услышать голос ваш… или… любоваться, мучительно любоваться вашим безмятежным сном.
”Вижу”(описание/ вопрос)	Как… обожатель… Я вас вижу в первый раз.

Эмоциональное восклицание	Сеньйор дон Бальтазар д’Эстуриз!
Короткий ответ	Тотчас.
Короткий ответ с отрицательной частицей	Я ее не боюсь…
Эмоциональное высказывние/ вопрос	Это что значит? Бальтазар…
Фраза с троеточием	Ах, останьтесь, останьтесь… Если б вы знали…
Благодарность/ радость/ поздравление	Ни полслова… даже не поблагодарю вас.
Короткий вопрос	Это что?
Обращение по имени отчеству	Дон Бальтазар д’Эстуриз, друг мой, извольте предложить ваше мнение. Мы вас слушаем.
Быт/ передвижения	В сад… да до них высоко.
Князь/ царь/ король/ власть	Я прислан от графа Касандра к вашему благоро... Да вить вы, сударь, дворянин?
Мнение/ действие	Да он сумасшедший!.. Он на дворе, бежит в сад… стучится в дверь. Ах, я пропала, пропала! Пойду, запрусь в своей комнате… авось, его не увидят… Нет, решительно отказываюсь от всяких необыкновенных приключений.
Семейный отношения (Бог как отец)	. Это ты, ты, мой спаситель, отец… Сангре, спаси, заступись… скорей… Поймай его, поймай… Вообрази себе… Да как он забрался, а? Отчего ты не кричала, а? Ты сама с ними в заговоре, старая ведьма…
Короткая фраза с троеточием	Но…
О знании	Я, право, не знаю…

О женщине	О, невинная голубка! Где? Вы спрашиваете, где? Здесь… и не только здесь, но даже… , там… Надобно ее удивить…
О действии	Да я не могу вас выдать за вора.
О передвижении	Да как уйти? Я не птица, не могу перелететь через трехаршинный частокол… Ваш муж вернулся?
Свадьба/ жених/ невеста	Достойного жениха скудной или, лучше сказать, неимущей девке трудно иметь: скудный и достойный меня не возьмет, а за недостойного богача я не пойду.
К матери/ о матери	А ты, матушка, что так печальна?
Эмоциональное высказывание	. А-га!
О деньгах	Понимаю, но мне деньги не нужны.
Мнение о человеке/ людях/ группе лиц	Мы его пропустим… Притом, не забудьте, в случае опасности, вы в одно мгновенье можете скрыться.
Короткое восклицание/ ответ	. Да я тут шею себе сломлю.
Отрицательный ответ/ восклицание (нет/ нельзя/ невозможно)	Не может быть…
Прощание/ просьба прощения/стыд	. Как она мила! Не гневайтесь на меня… Смотрите, я стал на колени, я прошу вашего прощения…
Вопрос-уточнение	Сеньйора?
Вопрос территориальный	Но куда вы меня поведете?
Обращение/ восклицание	Господин мой, сеньйор!
О любви	И я тоже. — Сеньйора… я давно вас люблю… что я говорю: люблю! я страстно, я отчаянно в вас влюблен… Вы меня не замечали; но я сам всячески старался не быть замеченным вами… Я боялся навлечь на себя и на вас подозрение вашего супруга.

О себе (действие)	. Это я.
Короткая фраза/ вопрос с отрицанием	Ничего, ничего; я сама скоро лягу спать. Ступай, ступай, бедняжка; мне, право, жаль тебя…
О себе	Пресильный? И его я не боюсь.
Короткий утвердительный ответ/ короткий вопрос	Ну?
Длинная фраза про общество/ страну/ общественное положение/ работу/ быт	Что же могут подумать? Разве это не улица? Разве не всем позволено ходить по этой улице? Я прохожу мимо… и вздумал вернуться. Что же тут предосудительного… или подозрительного? Мне это место понравилось… А вы… вы сидите на балконе…
Длинное рассуждение (о высоких темах)	А! бывали! Не правда ли, как приятно притаиться и ждать, долго ждать? Вот, птички, красивые, веселые птички, начинают понемногу слетаться; сперва дичатся, робеют; потом начинают поклевывать корм ваш, ваш собственный корм; наконец, совершенно успокоятся и уж посвистывают, да так мило, так беззаботно!.. Вы протягиваете руку, дергаете веревочку: хлоп! сеть упала — все птицы ваши; вам только остается придавить им головки — приятное удовольствие! Пойдем, Бальтазар! Сети расставлены, птицы слетелись; пойдем, пойдем!
Мнение о собеседнике	. Сеньйора, ваш смиренный и почтительный обожатель ждет вашего ответа.

О жене/ муже/ отношениях	. Вы хотите уйти?.. А сами сейчас жаловались на одиночество, на скуку… Да помилуйте, если вы станете избегать всякого знакомства, как же вы хотите избавиться от скуки? Правда, наше знакомство началось довольно странным образом… что за беда! Вот, я уверен, с вашим супругом вы познакомились самым обыкновенным образом

Приложение 3. Названия и примеры итоговых кластеров коротких диалогов.

Название кластера: Про передвижение в пространстве (в основном короткие фразы)
Пример:
– На Старой Басманной.
– И мы там тоже...
– Одно время я жил на Немецкой улице. С Немецкой улицы я хаживал в Красные казармы. Там по пути угрюмый мост, под мостом вода шумит. Одинокому становится грустно на душе. А Здесь какая широкая, какая богатая река! Чудесная река!

Название кластера: Комбинация коротких вопросов и ответов
Пример:
– Так не соизволите ли, ваше высокорейсграфское превосходительство, хотя рюмочку рейнского или церковного?
– Нет, сударыня, благодарствую.
– Ин медку или бражки?

Название кластера: Рассуждения на общественные темы
Пример:
– А сверх того сам прикажи, что варить, жарить, печь, только бы всего было довольно. Салат подай не с конопляным, да с ореховым маслом.
– Знатные господа больше к салату деревянное масло употребляют: так не прикажите ли лучше к салату лампатнова положить масла?
– Фу, батька! вить я не басурманка! А после кушанья поставьте стручков, бобов, моркови, репы да огурцов и свежих, и свежепросольных, а кофе подавайте с сахаром, а не с патокою. Исправь же все как надобно, да пошли на базар купить золоченых пряников, да паутины вели обместь, а двери-то вели подмазать, чтобы не скрипели, да людей вели накормить.

Название кластера: Длинные рассуждения на высокие темы
Пример:
– Да я его высокой милости, покуль душа в теле, не позабуду. И коли бы он такую мне многогрешной показал отеческую милость и велел бы маляру красками

написать персону свою, я бы ее у себя поставила пред кроватью и не спустила бы с нее глаз.
– Как будто слышало это мое сердце! Да почему ты знаешь его и какую сделал он тебе милость?
– А вот, сердечушко, я тебе донесу. Как я нынешнею зимою была без тебя в Москве, так расхвалили мне какую-то интермецию и уговорили меня туда съездить. Бывает и на старуху проруха. Поехала, вошла я в залу, заиграли и на скрипицах, и на гобоях, и на клевикортах; вышли какие-то и почали всякую всячину говорить, и уж махали, махали руками, как самые куклы; потом вышел какой-то, а к нему какую-то на цепи привели женщину, у которой он просил не знаю какого письма, а она отвечала, что она его изодрала; вышла, ему подали золоченый кубок, а с каким напитком, этого я не знаю; этот кубок отослал он к ней, и все было хорошо; потом какой-то еще пришел, поговорили немного, и что-то на него нашло; как он, батька, закричит, шапка с него полетела, а он и почал метаться, как угорелая кошка, да выняв нож, как прыснул себя, так я и обмерла. А граф этот, сидя тогда со мною в одном чулане и разговорився прежде еще интермедии, что я его соседка, меня тогда мунгальской водкой, как я от страха обмерла, от смерти избавил.

Название кластера: Комбинация из коротких вопросов и ответов
Пример:
– Какого это енарала адъютант у нас был?
– Не адъютант, егерь был. По-нашему, слуга, который стреляет ходя птиц.
– Какой слуга; весь в прозументах.

Название кластера: Диалоги из коротких вопросов и ответов
Пример:
– Нет, сударыня, благодарствую.
– Ин медку или бражки?
– Нет, сударыня.

Название кластера: Длинные рассуждения о чувствах/религии/мифологии
Пример:
– Странный он человек. Мне и жаль его, и досадно, но больше жаль. Мне ка-

жется, он застенчив... Когда мы вдвоем с ним, то он бывает очень умен и ласков, а в обществе он грубый человек, бретер. Не ходите, пусть пока сядут за стол. Дайте мне побыть около вас. О чем вы думаете? Вам двадцать лет, мне еще нет тридцати. Сколько лет нам осталось впереди, длинный, длинный ряд дней, полных моей любви к вам...
– Николай Львович, не говорите мне о любви.
– У меня страстная жажда жизни, борьбы, труда, и эта жажда в душе слилась с любовью к вам, Ирина, и, как нарочно, вы прекрасны, и жизнь мне кажется такой прекрасной! О чем вы думаете?

Название кластера: Один из персонажей говорит, а второй немного вмешивается в диалог
Пример:
– В горшечке, да в муравленом, и покройте его веницейскою тарелкой; с морковью пироги, пирожки с солеными груздями, левашники с сушеною малиной, фрукасе из свинины с черносливом, французский пирог из подрукавной муки, а начинка из брусничной пастилы. Да есть ли у нас калужское тесто?
– Имеется.
– А сверх того сам прикажи, что варить, жарить, печь, только бы всего было довольно. Салат подай не с конопляным, да с ореховым маслом.

Название кластера: Комбинации из коротких вопросов и ответов (не обязательно коротких
Пример:
– Как будто слышало это мое сердце! Да почему ты знаешь его и какую сделал он тебе милость?
– А вот, сердечушко, я тебе донесу. Как я нынешнею зимою была без тебя в Москве, так расхвалили мне какую-то интермецию и уговорили меня туда съездить. Бывает и на старуху проруха. Поехала, вошла я в залу, заиграли и на скрипицах, и на гобоях, и на клевикортах; вышли какие-то и почали всякую всячину говорить, и уж махали, махали руками, как самые куклы; потом вышел какой-то, а к нему какую-то на цепи привели женщину, у которой он просил не знаю какого письма, а она отвечала, что она его изодрала; вышла, ему подали золоченый кубок, а с каким напитком, этого я не знаю; этот кубок отослал он к ней,

и все было хорошо; потом какой-то еще пришел, поговорили немного, и что-то на него нашло; как он, батька, закричит, шапка с него полетела, а он и почал метаться, как угорелая кошка, да выняв нож, как прыснул себя, так я и обмерла. А граф этот, сидя тогда со мною в одном чулане и разговорився прежде еще интермедии, что я его соседка, меня тогда мунгальской водкой, как я от страха обмерла, от смерти избавил.
– А хорош граф-ат?

Название кластера: Диалог, с короткими фразами. Зачастую бытовой
Пример:
— И это не худо. Что ж ты, князь, так задумался?
– Думаю о том, что слышу, да ничего сам придумать не могу; а признаюсь, что обедать пора, и потому прошу вас, господин предводитель, и вас, государь мой, у меня откушать.
– Как вам угодно.

Название кластера: Разговор с обращениями друг к другу по имени/отчеству
Пример:
– Одной рукой я поднимаю только полтора пуда, а двумя пять, даже шесть пудов. Из этого я заключаю, что два человека сильнее одного не вдвое, а втрое, даже больше...
– При выпадении волос... два золотника нафталина на полбутылки спирта... растворить и употреблять ежедневно... Запишем-с! Так вот, я говорю вам, пробочка втыкается в бутылочку, и сквозь нее проходит стеклянная трубочка... Потом вы берете щепоточку самых простых, обыкновеннейших квасцов...
– Иван Романыч, милый Иван Романыч!

Название кластера: Обмен информации, запрос на информацию (о знании)
Пример:
– Да он ее и не знает.
– Он сосед ваш, так ему известно имя вашей сожите... супру... ну как ни есть.
– Почему известно?

Название кластера: О движении/ местонахождении

Пример:

– Да, это ужасно. Он всегда делает глупости.

– У лукоморья дуб зеленый, златая цепь на дубе том... Златая цепь на дубе том...

– Ты сегодня невеселая, Маша. Куда ты?

Название кластера: Один/ оба говорят о себе, своем положении/действиях

Пример:

– Это уж очень низко.

– Нижайшего поклона ничего нет ниже. А что всенижайший поклон, этого я уже и не понимаю.

– Что тебе еще приказано?

Название кластера: Комбинации из коротких вопросов и ответов

Пример:

– Уже студеное на стол, сударь, поставили.

– Я тебе этого не приказывала.

– Как же без этого?

Название кластера: Активное действие

Пример:

– Он еще в постеле. Да от кого ты прислан и зачем?

– К кому я прислан, тому я и скажу, от кого я прислан и зачем.

– Фу, батька, какой спесивый!

Название кластера: Активное действие

Пример:

– А ты, душенька, так хороша, что я едаких хорошеньких мало видал. Знаешь ли ты, девушка, что я в тебя смертно влюбился.

– Перестань же балагурить-то.

– Какое балагурство! Ежели это ложь, так ты повесь меня.

Название кластера: Рассуждения на разные темы

Пример:

– У покровителей зевать на потолок,
Явиться помолчать, пошаркать, пообедать,
Подставить стул, поднять платок.
У покровителей зевать на потолок,
Явиться помолчать, пошаркать, пообедать,
Подставить стул, поднять платок.
– Он вольность хочет проповедать!Он вольность хочет проповедать!
– Кто путешествует, в деревне кто живет...Кто путешествует, в деревне кто живет...

Название кластера: Разговор о том что можно/ нельзя или возможно/ невозможно
Пример:
– Положите, сударыня на меня, так я о вашем счастии, сколько можно, буду иметь попечение.
– Я очень благодарна и принимаю ваше доброе и великодушное намерение за исполнение, хотя бы я от вас и никакого никогда в перемене моей жизни успеха не получила. Да только не станет сил ваших ко вспоможению бедных, когда вы, увидя кого в первый раз, толиким великодушием наполняетесь.
– Для всех многого сделать не можно, да вы не в том числе.

Название кластера: Диалог о семье/ любви/ свадьбе
Пример:
– Не к тому клонится.
– Еще ты молода; так может быть, выйдешь за такого мужа, который все твои нынешние грусти превратит в веселость.
– Достойного жениха скудной или, лучше сказать, неимущей девке трудно иметь: скудный и достойный меня не возьмет, а за недостойного богача я не пойду.

Название кластера: Женщины/ быт/ дом
Пример:
– Нет, братец, помещик я, а не она. А ей принадлежит только седьмая часть из недвижимого моего имения. И то еще тогда достанется ей, ежели она меня переживет.

– Мне приказано и ей отдать нижайший поклон.
– Хорошо, друг мой, я ей этот поклон отнесу.

Название кластера: Разговор о матери /с матью
Пример:
– Я вашу матушку знал.
– Хорошая была, царство ей небесное.
– Мама в Москве погребена.

Название кластера: Короткие фразы, высказывания о себе/ других людях
Пример:
– Хотя и не богатый... Да зачем и от кого ты прислан?
– Я прислан от графа Касандра к вашему здоровью. Граф приказал вам нижайший отдать поклон.
– Это уж очень низко.

Название кластера: Действие, короткие эмоциональные фразы
Пример:
– Какой слуга; весь в прозументах.
– Ныне у господ такой манер. Это был егерь от графа Касандра: его сиятельство к нам заехать изволит.
– Его сиятельство!

Название кластера: Обмен личными мнениями о ситуации
Пример:
– Жена, кто говорит о ревности.
– Что это меня прорвало! Да полно, конь о четырех ногах, да и тот спотыкается, а я баба безграмотная, так как не промолвиться.
– Да ты не в слове, да в деле промолвилася.

Название кластера: Рассказ о себе/ других людях
Пример:
– Титулуй как хочешь. Да что графу до моей жены?
– То дело, чтобы засвидетельствовать ей свое почтение.

– Да он ее и не знает.

Название кластера: Диалог, где один из персонажей говорит о себе
Пример:
– Это для вас теперь... Пожалуйте-с!..
– Вона, мать, гарнитуры-то какие: мы и не видывали здесь этаких. На-ка, и бархатцу-то на оторочку привез. Словно кукла нарядная, будешь ходить у нас в шелках да в бархате.
– А вас, извините на том, не чаял здесь захватить... Позвольте, по крайности, хоть полтинничком поклониться...

Название кластера: Длинные диалоги про общество/ страну/ общественное положение/ работу/ быт
Пример:
– Я прислан от графа Касандра к вашему сия... к вашему превосхо... к вашему высоко... Какого, сударь, вы чина?
– Что тебе, братец, до моего чина? Какого ни есть.
– Я прислан от графа Касандра к вашему высокоблаго... Вить вы, сударь, имеете майорский чин?

Название кластера: Фразы с троеточиями
Пример:
– Что, девочка моя, радость моя?
– Скажите мне, отчего я сегодня так счастлива? Точно я на парусах, надо мной широкое голубое небо и носятся большие белые птицы. Отчего это? Отчего?
– Птица моя белая...

Название кластера: Кто-то говорит о себе
Пример:
– Он, если вы позволите, также будет участником в нашем деле.
– Согласна, Николай Иваныч, я на все согласна. По мне, хоть весь уезд, всю губернию созовите: у меня совесть чиста, Николай Иваныч. Они, я знаю, за меня заступятся. Они не дадут меня в обиду… А вы как в своем здоровье, Евгений Тихоныч?

– Хорошо. Что мне деется! Покорно благодарю.

Название кластера: Шутки/смех/обида

Пример:

– Да опять медведю в лапы попади.

– Ты мне страшнее медведя. Я с ним оправиться умею, а твоего сердца ни дробью, ни пулей не добудешь.

– Лих ты шутить.

Название кластера: Вера/ обещание/ прощение

Пример:

– Да я ничем особливого вашего снисхождения не заслужила.

– Я бы желал того, чтобы и вы такое усердие ко мне получили, какое я к вам в это получил короткое время.

– Мы и сердца наши закрыты! Я вам верю, да поверьте и мне, что и я не меньше к вам усердия имею.

Название кластера: Фразы с троеточиями, которые обрамляют/прерывают более длинные рассуждения

Пример:

– Не забудь и нас, сиятельнейшая графиня!

– Я еще не графиня, а вашей дружбы никогда не забуду.

– Высокосиятельнейшая графиня! Не оставь нас, ежели какая нужда...

Название кластера: Диалог из вопросов и ответов (в основном вопрос-ответ-вопрос)

Пример:

– Да есть пословица, что гром-ат гремит не всегда из небесной тучи, да иногда и из навозной кучи.

– Типун бы табе на язык; какая навозная у табя я куча?

– Что это, сударыня, такое?

Название кластера: О власти/ служении

Пример:

– Я прислан от графа Касандра к вашему благоро... Да вить вы, сударь, дворянин?
– Хотя и не богатый... Да зачем и от кого ты прислан?
– Я прислан от графа Касандра к вашему здоровью. Граф приказал вам нижайший отдать поклон.

Название кластера: О любви
Пример:
– А что? Разве он тебе знаком?
– Да я его высокой милости, покуль душа в теле, не позабуду. И коли бы он такую мне многогрешной показал отеческую милость и велел бы маляру красками написать персону свою, я бы ее у себя поставила пред кроватью и не спустила бы с нее глаз.
– Как будто слышало это мое сердце! Да почему ты знаешь его и какую сделал он тебе милость?

Название кластера: Быстрое действие (короткие и эмоциональные фразы)
Пример:
– Полно, дурища.
– Полно табе, дурачища.
– Постыдитеся.

Название кластера: Короткие вопросы и ответы
Пример:
– Простите меня, милостивый государь.
– Да что этому причина, что ты не поехал?
– Любовь.

Название кластера: Разговор о прошлом/ будущем/ нынешнем разговоре
Пример:
– Перестань же балагурить-то.
– Какое балагурство! Ежели это ложь, так ты повесь меня.
– Пора мне идти к барам, скоро барыня встанет. Так что же мне о тебе сказать?

Название кластера: Высказывание различных мнений о людях

Пример:

– Конечно, ты балагур?

– А ты, душенька, так хороша, что я едаких хорошеньких мало видал. Знаешь ли ты, девушка, что я в тебя смертно влюбился.

– Перестань же балагурить-то.

Название кластера: Быстрый диалог с короткими ответами

Пример:

– В горшке прикажешь, барыня-государыня, или на блюде?

– В горшечке, да в муравленом, и покройте его веницейскою тарелкой; с морковью пироги, пирожки с солеными груздями, левашники с сушеною малиной, фрукасе из свинины с черносливом, французский пирог из подрукавной муки, а начинка из брусничной пастилы. Да есть ли у нас калужское тесто?

– Имеется.

Название кластера: Диалог с короткими эмоциональными фразами

Пример:

– Нисколько!

– А я хочу тебя назвать: влюбленный скрипач!

– Или влюбленный профессор!..

Название кластера: О свадьбе

Пример:

– Я готов раз пять обвенчаться с тобою.

– И пять раз изменить.

– Многие бы мужья и жены постоянными еще назывались, ежели бы только друг другу по пяти раз изменяли.

Название кластера: Быт и деньги

Пример:

– Почти все по миру ходят, не здесь-то и не вам-то сказано.

– Отчего это?

– Боярыня наша праздности не жалует и ежечасно крестьян ко труду понуждать

изволит. Щегольство и картежная игра умножилися, и ежели крестьяне меньше работать будут, так чем нашим помещикам и пробавляться. А мои господа хотя ни щегольства, ни картежной игры и не жалуют, однако, собирая деньги, белую денежку на черный день берегут.

Название кластера: Комбинации из коротких вопросов и ответов
Пример:
– Я не приглашала.
– И прекрасно.
– Самовар! Это ужасно!

Название кластера: Короткие вопросы/ответы на тему передвижений
Пример:
– У лукоморья дуб зеленый, златая цепь на дубе том... Златая цепь на дубе том...
– Ты сегодня невеселая, Маша. Куда ты?
– Домой.

Название кластера: Активное действие (призывы к действию)
Пример:
– Он у нас и ученый, и на скрипке играет, и выпиливает разные штучки, одним словом, мастер на все руки. Андрей, не уходи! У него манера – всегда уходить. Поди сюда!
– Иди, иди!
– Оставьте, пожалуйста.

Название кластера: Быстрый диалог с короткими эмоциональными фразами
Пример:
– Ну, что ж! Очень рада.
– Он старый?
– Нет, ничего. Самое большее, лет сорок, сорок пять. По-видимому, славный малый. Неглуп, это – несомненно. Только говорит много.

Название кластера: Описания

Пример:

– Что ты так весела, Ниса?

– А ты, матушка, что так печальна?

– Коли ты меня веселою видишь?

Название кластера: Диалоги с территориальными вопросами (о движении/местонахождении)

Пример:

– Ряженые!

– Скажи, нянечка, дома нет никого. Пусть извинят.

– Никого нет... А где же все?