

Правительство Российской Федерации
Федеральное государственное бюджетное образовательное учреждение
высшего профессионального образования
«Санкт-Петербургский государственный университет»

Кафедра системного программирования

Михайлова Александра Алексеевна

Автоматическое извлечение ключевых фактов из
неструктурированных текстов для случая отзывов и
статей об отелях

Дипломная работа

Допущена к защите.

Зав. кафедрой:

д. ф.-м. н., проф. Терехов А.Н.

Научный руководитель:

ст. преп. Луцив Д.В.

Рецензент:

рук. отдела Горовой В.А., ООО «Яндекс»

Санкт-Петербург

2015

SAINT-PETERSBURG STATE UNIVERSITY

Software Engineering Chair

Aleksandra Mikhailova

Automatic key fact extraction from unstructured texts for the
case of hotel reviews and articles

Graduation Thesis

Admitted for defence.

Head of the chair:

Professor Andrey Terekhov

Scientific supervisor:

Sr. Lecturer Dmitry Luciv

Reviewer:

Head of Dept., Yandex LLC Vladimir Gorovoy

Saint-Petersburg

2015

Оглавление

Введение	4
Постановка задачи	6
1. Обзор существующих решений	8
1.1 Извлечение синтаксических фактов из текстов на естественном языке	8
1.2 Извлечение информации об отелях на основе онтологий	10
1.3 Коммерческие сервисы с данными про отели	10
2. Данные для экспериментов	14
3. Алгоритмы извлечения ключевых фактов	16
3.1 Алгоритм, основанный на частотности фактов	16
3.1.1 Особенности реализации	16
3.1.2 Практическое применение: построение рейтингов отелей	17
3.2 Алгоритм, основанный на статистической мере TF-IDF	19
3.2.1 Особенности реализации	19
3.2.2 Практическое применение: получение нестандартной важной информации	20
3.3 Проверка эффективности алгоритмов	21
3.3.1 Оценка алгоритма, основанного на частотности фактов	21
3.3.2 Оценка алгоритма, основанного на TF-IDF	22
4. Оценка качества базы данных по отелям	23
5. Эксперименты с машинным переводом	24
5.1 Расширение набора данных для экспериментов	24
5.2 Результаты	24
Заключение	26
Список литературы	28

Введение

По данным статистики, опубликованным Всемирным банком, за последние десять лет доля жителей Российской Федерации, имеющих доступ к сети Интернет, превысила 60% [1]. В то же самое время, согласно исследованиям, проведённым аналитическим центром Pew Research Center в 2014 году и посвящённым использованию информационных технологий и, в частности, Интернета в развивающихся странах, одним из основных сценариев использования Интернета является поиск информации о различных продуктах и совершение онлайн-покупок [2].

Для предоставления информации о товарах и услугах, которая соответствует поисковым запросам пользователей и обладает высокой степенью пертинентности¹, многие Интернет-порталы внедряют соответствующие рекомендательные сервисы. Простейшей и наиболее распространённой разновидностью таких рекомендательных сервисов являются системы фильтрации, основанные на фиксированном наборе параметров-фильтров, организованных в форму ввода на сайте. В этом случае процесс построения рекомендации состоит из обработки заданных пользователем параметров и последующего построения соответствующего рейтинга. Одним из самых популярных сервисов в русскоязычном Интернете, использующих описанную рекомендательную модель, является «Яндекс.Маркет»².

В связи с популярностью туризма как в России, так и во всём мире, а также проникновением телекоммуникационных технологий в туристическую отрасль отдельные услуги, туры и экскурсии дополнили

¹ Пертинентность (англ. pertinence) – соответствие полученной информации информационной потребности пользователя,

<https://ru.wiktionary.org/wiki/пертинентность>

² Яндекс.Маркет, <http://market.yandex.ru/>

перечень продуктов, информацию о которых люди предпочитают получать из сети Интернет. Этим обусловлено появление и стремительное развитие таких международных Интернет-сервисов, как Booking.com³ и TripAdvisor⁴. В российском сегменте можно выделить TopHotels⁵, а также запустившийся в марте 2015 года сервис «Яндекс.Путешествия»⁶. Все указанные сервисы используют фильтры при создании своих рекомендаций. Например, отели можно выбирать по таким характеристикам как близость к пляжу, наличие беспроводного или проводного Интернета, тип питания, месторасположение и др. В данной работе в качестве примера туристических продуктов будут рассматриваться такие виды гостиничных домов, как отели, хостелы, полупансионы и апартаменты. Для облегчения восприятия все перечисленные выше разновидности будут обозначаться словом «отель».

Для осуществления фильтрации туристических продуктов по фиксированному набору параметров, как правило, используется база данных, содержащая набор характеристик продукта. При этом внедрение такой базы знаний в сервис налагает обязательства по поддержанию её актуальности, а также регулярной проверке и корректировке содержащейся в ней информации. Как правило, осуществление данных задач требует привлечения контент-менеджеров для выполнения значительного количества «ручного» труда. В свете этого перспективным выглядит внесение изменений в существующую базу данных, основанное на автоматической обработке постоянно обновляющихся информационных ресурсов. В сфере туризма такими ресурсами являются

³ Booking.com, <http://www.booking.com/>

⁴ TripAdvisor, <http://www.tripadvisor.com/>

⁵ TopHotels, <http://www.tophotels.ru/>

⁶ Яндекс.Путешествия, <https://travel.yandex.ru/>

отзывы путешественников на специализированных сайтах, а также статьи в новостных Интернет-изданиях.

Как отзывы туристов, так и новостные заметки представляют собой неструктурированные тексты на естественном языке. Для эффективного анализа содержимого подобных текстов используются так называемые «факты» – биграмы вида «параметр» (главное, определяемое слово) + «характеристика» (зависимое слово, определение), – которые можно извлечь из текста с помощью различных инструментов синтаксического анализа. Одним из широко используемых инструментов такого рода является Томита-парсер⁷, разработанный компанией «Яндекс» на основе алгоритма GLR-парсинга, описанного японским учёным Масару Томитой [3]. Однако, число фактов, извлечённых из текста с использованием подобных «синтаксических» инструментов зачастую избыточно: к примеру, работающий на основе грамматик и словарей Томита-парсер извлекает из статьи длиной в 5000 символов порядка сотни словосочетаний, в то время как её содержание может быть основано не более чем на десяти основных фактах – так называемых «ключевых фактах». Автоматизация извлечения именно сокращённого множества фактов в рамках некоторой заданной тематики требует особого подхода при обработке текстов на естественном языке.

Постановка задачи

Целью данной дипломной работы является исследование возможности качественного извлечения ключевых русскоязычных фактов из отзывов и статей про отели. В данном исследовании выделяются два этапа: необходимо предложить новые алгоритмы для извлечения фактов и проанализировать области применимости данных алгоритмов в рамках

⁷ Томита-парсер, <https://tech.yandex.ru/tomita/>

сервиса «Яндекс.Путешествия». Для достижения этой цели были поставлены следующие задачи.

1. Исследовать существующие инструменты для работы с текстами на русском и английском языках.
2. Собрать данные для работы: отзывы об отелях со специализированных сайтов и статьи об отелях из результатов поиска.
3. Создать, реализовать и протестировать несколько различных алгоритмов извлечения ключевых фактов из неструктурированных текстов на русском языке.
4. Оценить качество базы данных фактов об отелях, полученной у партнёров компании «Яндекс», с помощью разработанных алгоритмов.
5. Проверить следующую гипотезу: можно ли улучшить качество работы созданных алгоритмов, расширив обучающее множество текстами, которые были автоматически переведены на русский язык с английского.

1. Обзор существующих решений

1.1 Извлечение синтаксических фактов из текстов на естественном языке

На сегодняшний день существуют инструменты для выделения из текстов на естественном языке различных синтаксически корректных словосочетаний, соответствующих грамматике данного языка. Ниже описаны конкретные инструменты для русского и английского языков, используемые в промышленных сервисах, а также в исследованиях наиболее часто.

Томи́та-парсер. В качестве примера такого инструмента для русского языка можно привести Томи́та-парсер, который был разработан компанией «Яндекс» и в настоящее время используется в актуальных пользовательских сервисах этой компании [4]. Данный инструмент был создан на основе алгоритма GLR-парсинга⁸ и позволяет выделять из текста цепочки слов (т.н. факты). Извлечение происходит согласно указанным пользователем правилам на языке контекстно-свободных грамматик и с использованием словарей ключевых слов (т.н. газеттиров⁹).

Грамматика для Томи́та-парсера представляет собой множество правил на языке контекстно-свободных грамматик, которые описывают структуру выделяемых цепочек. На листинге 1 приведён пример простейшей грамматики, с помощью которой можно извлечь все пары прилагательных и существительных.

⁸ GLR Parser, http://en.wikipedia.org/wiki/GLR_parser

⁹ Газеттир, грамматика, факты: основные понятия, <https://tech.yandex.ru/tomita/doc/dg/concept/about-docpage/>

```
#encoding "utf-8"  
#GRAMMAR_ROOT S  
S -> Adj Noun;
```

Листинг 1. Простейшая грамматика для Томита-парсера

Газеттиры представляют собой словари ключевых слов, разделённые на статьи, задающие множества слов и словосочетаний, объединённых общим свойством.

Результатом работы Томита-парсера является список словосочетаний из текста, соответствующих указанной грамматике и газеттирам. При этом слова в словосочетании остаются в согласованном виде и по возможности приводятся к нормальной форме.

Stanford CoreNLP. Stanford CoreNLP – набор библиотек для разработки средств анализа естественного языка, разрабатываемая студентами и сотрудниками Стэнфордского университета [5]. Данный инструмент был создан в первую очередь для анализа текстов на английском языке, но сейчас поддерживает также китайский, испанский, немецкий и арабский языки. Stanford CoreNLP – инструмент с открытым исходным кодом¹⁰.

Синтаксический анализ текстов позволяет извлечь все формальные словосочетания, соответствующие определённым правилам, однако количество извлечённых словосочетаний из одного текста может быть слишком велико, что будет затруднять анализ. Синтаксические парсеры никак не учитывают тематику текста и не способны отфильтровывать факты, относящиеся к заданной теме.

¹⁰ Stanford CoreNLP: A Java suite of Core NLP tools,

<https://github.com/stanfordnlp/CoreNLP>

1.2 Извлечение информации об отелях на основе онтологий

В работе [6] описан метод извлечения информации из текстовых источников данных, основанный на использовании семантической паутины¹¹ с опорой на онтологии. Авторы данной работы предлагают алгоритм построения онтологической базы знаний, проводя частотный анализ результатов работы текстового парсера синтаксических троек StanfordCoreNLP [5].

Аналогичным путём идут авторы исследования [7], расширяя область применимости описанных алгоритмов на статьи произвольной тематики и иллюстрируя работу алгоритмов на примере туристического сектора.

1.3 Коммерческие сервисы с данными про отели

Коммерческие сервисы в данной сфере, как правило, используют закрытые технологии, подробности которых не описаны в открытом доступе. Однако, в контексте данной работы представляется интересным для некоторых из таких сервисов дать краткое описание функциональности, связанной с построением рейтингов по набору фактов.

TrustYou. Сервис TrustYou (<http://www.trusty.com/>) предоставляет пользователям краткую информацию об отелях в виде коротких фактов, а также некоторый рейтинг, который строится на их основе. Это проиллюстрировано на рис.1.

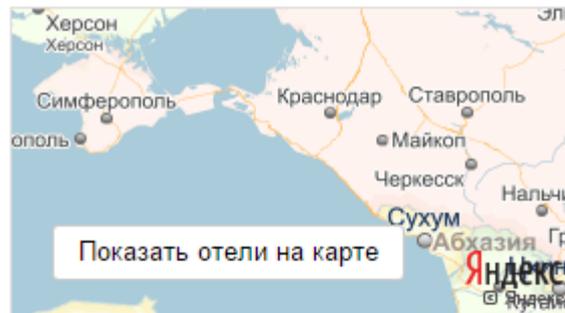
¹¹ Семантическая паутина (англ. Semantic Web),
https://en.wikipedia.org/wiki/Semantic_Web

Review highlights



Рис. 1. Построение рейтингов отелей по параметрам на основе фактов, TrustYou **Яндекс.Путешествия**. Запустившийся в конце марта 2015 года¹² сервис Яндекс.Путешествия (<https://travel.yandex.ru/>) предоставляет пользователям поисковые фильтры и рекомендации отелей, основываясь на существующей базе данных фактов. В данный момент актуальность этой базы поддерживается посредством ручного труда контент-менеджеров. Частичная автоматизация этого процесса может привести к существенной экономии ресурсов.

¹² Блог компании «Яндекс», 31 марта 2015 года, <https://blog.yandex.ru/post/93323/>



от 27 354 | до 158 146 ₽

Количество звёзд

5* 4* 3* 2* 1*

Расположение отеля

1-я линия 2-я или 3-я линии

Пляж

песчаный галечный
 собственный

Тип питания

любой ▼

Интернет

бесплатный Wi-Fi

Рис. 2. Фильтры поиска, Яндекс.Путешествия

Booking.com. Международный сервис **Booking.com** (<http://www.booking.com/>) также осуществляет фильтрацию отельных заведений по ключевым аспектам размещения (см. рис. 3). Помимо этого, для каждого конкретного отеля сервис предоставляет краткую информацию об особенностях объекта, что проиллюстрировано на рис. 4.

Выбрать по критериям:

- ▶ **Количество звезд**
- ▶ **Тип размещения**
- ▶ **Оценка по отзывам**
- ▼ **Удобства**
 - Wi-Fi (390)
 - Парковка (430)
 - Трансфер от/до аэропорта (20)
 - Фитнес-центр (102)
 - Номера для некурящих (402)
 - Крытый плавательный бассейн (21)
 - Спа и оздоровительный центр (23)
 - Семейные номера (182)
 - Открытый плавательный бассейн (11)
 - Допускается размещение домашних животных (202)
 - Номера/Удобства для гостей с ограниченными физическими возможностями (180)
 - Ресторан (187)

Особенности объекта размещения

Самое недавнее бронирование: 15 минут назад

 Бесплатный Wi-Fi

Ориентиры:
Автобусный вокзал Сититерминален (100 м)
Т-Централен (100 м)
Центральный вокзал Стокгольма (150 м)

Забронировать

Рис 3, 4. Сервис Booking.com

2. Данные для экспериментов

Описание. Данные для экспериментов были собраны на основе реального опыта пользователей поисковой системы «Яндекс» следующим образом. Был выбран топ наиболее популярных отелей, т.е. отелей, названия которых чаще всего фигурируют в запросах пользователей к поисковой системе. Для данных отелей были выгружены два типа документов: т.н. статьи как выдача поисковой системы на запрос с названием отеля и т.н. отзывы как результат аналогичного поиска с маркером «отзывы» в теле запроса. С помощью Томита-парсера из выгруженных текстов были извлечены все синтаксические биграмы. На данном шаге использовалась соответствующая грамматика для парсера и обширные словари, составленные вручную контент-менеджерами компании «Яндекс».

Структура данных. Полученные данные организованы в SQL-базу данных, схема которой приведена на рис. 5.

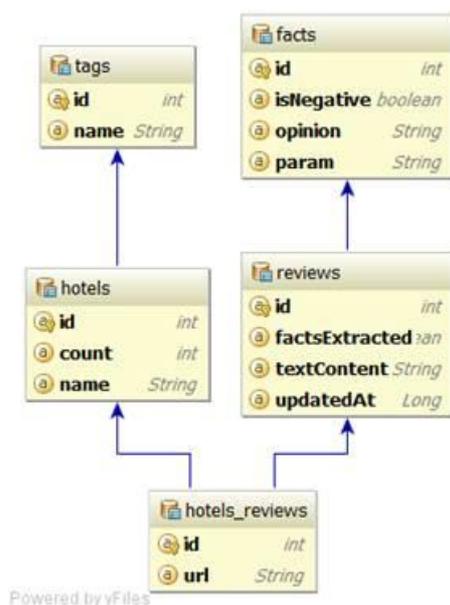


Рис. 5. Схема базы данных для экспериментов

Обучающее множество. Обучающее множество содержит 2 998 различных идентификаторов отелей. Для выбранных отелей было выгружено 149 507 статей, из которых было извлечено с помощью

Томи́та-парсера 12 434 262 синтаксических фактов. Построение моделей было выполнено только на основе статей, отзывы на этапе обучения не использовались ввиду их ограниченного размера и, как правило, малой информативности.

Тестовое множество. В тестовом множестве содержится 8 371 различных идентификатор отелей. Для выбранных идентификаторов отелей было выгружено 100 000 т.н. «отзывов». Отзывы – это результаты поисковых запросов пользователей с маркером «отзывы», а также тексты туристов с сайта Booking.com. Из данных текстов было извлечено порядка 1 000 000 синтаксических фактов.

3. Алгоритмы извлечения ключевых фактов

3.1 Алгоритм, основанный на частотности фактов

Данный подход направлен на сокращение множества фактов, извлечённых с помощью выбранного синтаксического парсера (в данном случае Томита-парсера). Ценность сокращённого подмножества фактов заключается в том, что оно помогает описать предметную область статей, не используя «лишних» словосочетаний, т.е. синтаксических словосочетаний, не относящихся к главной теме статьи. Алгоритм основан на подсчёте частотности словосочетаний и таким образом служит для выделения подмножества фактов, описывающие характеристики, общие для всех объектов предметной области. Например, для отелей в качестве примеров таких характеристик можно привести признаки «пляж», «интернет», «сервис».

3.1.1 Особенности реализации

Алгоритм состоит из следующих шагов.

1. Отфильтровать извлечённые синтаксические факты с использованием специально составленных словарей, содержащих только ключевые характеристики выбранной области (например, «пляж», «интернет» и др.). Данный шаг опционален: если не отфильтровывать характеристики по словарям, статистика будет считаться по всем извлечённым словосочетаниям. Несмотря на это, в топ наиболее употребляемых фактов выйдут словосочетания, чаще всего появляющиеся в статьях выбранной тематики (в данном случае это статьи и отзывы об отелях).
2. Подсчитать статистику появления оставшихся фактов по всей текстовой базе обучающего множества.
3. Отсортировать факты по убыванию частоты.

4. Взять первые N процентов фактов за искомое подмножество словосочетаний S .
5. На тестовом множестве оценить качество извлечения ключевых фактов предметной области с использованием множества S . Схема оценки качества алгоритма описана ниже в п. 3.3.1.
6. Изменяя параметр N и повторяя шаги 1-5 данного алгоритма, добиться желаемого качества извлечения фактов.

3.1.2 Практическое применение: построение рейтингов отелей

Поскольку цель данного алгоритма заключается в выделении фактов, которые относятся к характеристикам, общим для всех объектов выбранной предметной области, представляется возможным дальнейшее построение рейтингов объектов на основе извлечённых словосочетаний.

При извлечении факты f из текста об отеле h определим вес f для объекта h как отношение частоты появления f во всех текстах об объекте h к частоте появления f по всей текстовой базе (т.е. для всех объектов). В случае появления в результате противоречивых фактов предлагается сравнивать их веса для выбора более важного. Ниже представлен пример извлечённых фактов с весами для одного конкретного отеля.

Amata Resort #о. пхукет #Таиланд				
параметр	определение	истинно ли	вес для отеля	частота по базе
интернет	платный	false	0.02	0.25
завтрак	бесплатный	false	0.02	0.67
wi-fi	бесплатный	true	0.08	0.68

Табл. 1. Пример частотных фактов для конкретного отеля

На данном примере видно, что факт «wifi: бесплатный» имеет больший вес, нежели факт «интернет: платный», как для конкретного отеля, так и по всей базе извлечённых фактов в целом. Таким образом, при построении рейтингов, а также при оценке качества извлечения фактов мы

можем учитывать только факты с наибольшим весом для предотвращения противоречий.

3.2 Алгоритм, основанный на статистической мере TF-IDF

Данный подход направлен на извлечение фактов, описывающих особенности конкретного объекта в противовес общим характеристикам предметной области. В основе предложенного алгоритма лежит статистическая мера TF-IDF [8], позволяющая оценить важность слова в контексте документа, который является частью некоторой коллекции документов. Важность слова для документа пропорциональна частоте употребления этого слова в данном документе, и обратно пропорциональна частоте употребления слова во всех остальных документах из коллекции.

В контексте данной работы в роли слов выступают синтаксические факты, извлечённые с помощью Томита-парсера, а в качестве одного документа принимается объединение всех текстов, относящихся к одному отелю.

3.2.1 Особенности реализации

Обозначения. Примем следующие обозначения. $Articles(h)$ – все статьи, относящиеся к отелю h . $Articles(H)$ – все статьи, относящиеся к отелям из текущей базы данных.

Алгоритм. Данный алгоритм состоит из следующих шагов.

1. Для данного отеля h взять все тексты, относящиеся к данному объекту, извлечь все синтаксические факты с помощью парсера.
2. Для каждого синтаксического факта f вычислить величину
$$R(f) = tf(f, Articles(h)) * idf(f, Articles(H)).$$
3. Отфильтровать все факты с величиной $R > T$, где T – выбранный порог. Принять получившееся подмножество за искомое множество фактов.
4. Оценить качество извлечения характерных фактов. Стоит заметить, что в данном случае не представляется возможным

выделить отдельные этапы обучения и тестирования модели. Для интегрирования описанного подхода в рекомендательные системы, работающие в реальном времени, необходимо осуществить предварительный подсчёт фактов данного вида для выбранного множества объектов.

5. Изменяя параметр T и повторяя шаги 1-4 алгоритма, добиться желаемого качества извлечения фактов.

3.2.2 Практическое применение: получение нестандартной важной информации

Специфика метрики TF-IDF позволяет выделять особенности, характерные только для одного объекта коллекции и таким образом выделяющие данный объект среди всех остальных элементов. В контексте кратких характеристик отелей, в роли которых в данной работе рассматриваются факты, разработанная метрика предоставляет возможность извлекать отличительные черты отдельных отелей, которые были бы интересны пользователям, но не вышли бы в топ частотных словосочетаний по итогам первого алгоритма, описанного в п. 3.1. Ниже приведены примеры извлечения подобных характеристик, полученные в результате работы данного алгоритма.

текст	параметр	определение	истинно ли
«По пляжу ползают черепахи»	черепаха	ползать	true
«После этого бассейна у меня зелёные волосы!»	волос	зелёный	true

Табл. 2. Пример специфичных фактов для отелей.

3.3 Проверка эффективности алгоритмов

3.3.1 Оценка алгоритма, основанного на частотности фактов

Оценка алгоритма, основанного на частотности фактов, проводилась с использованием модификации стандартной метрики «точность»¹³. Также ввиду большого количества исходных данных и извлечённых фактов было использовано случайное семплирование: были построены 5 случайных выборок по 30 уникальных идентификаторов отелей, и оценивалась точность извлечения фактов из соответствующих этим идентификаторам статей.

Все извлечённые ключевые факты были разбиты на 4 типа, соответствующие понятиям True Positive, True Negative, False Positive и False Negative, обычно используемым в оценки качества бинарной классификации. Оценка истинности факта проводилась вручную с использованием подтверждённых источников: официальный сайт отеля и описание отеля на сайтах Booking.com и TopHotels. Далее по приведённой ниже формуле была вычислена метрика ACC (от англ. accuracy – «точность») отдельно для каждой случайной выборки.

$$ACC = (TP + TN) / (TP + TN + FP + FN)$$

В результате было получено, что с вероятностью 95% доверительный интервал (0.74, 0.89) содержит ACC алгоритма на случайной выборке. ACC пяти построенных выборок попали в данный доверительный интервал.

Одной из трудностей, проявившейся в процессе оценки, оказалось большое количество субъективных характеристик отеля, корректность которых не представляется возможным оценить независимо. В качестве примера фактов с подобными характеристиками можно привести такие

¹³ Англ. Accuracy (не путать с Precision),

http://en.wikipedia.org/wiki/Accuracy_and_precision

словосочетания, как «отдых: лучший», «пляж: шикарный» и т.п. Данная проблема была решена фильтрацией словосочетаний, содержащих определения, выраженные качественными прилагательными, обозначающими отношение к объекту (хороший, роскошный, шикарный и т.п.).

3.3.2 Оценка алгоритма, основанного на TF-IDF

В случае алгоритма, в основе которого лежит модификация метрики TF-IDF, не представляется возможность оценить объективные величины. Зачастую факты, извлечённые по данной схеме, не подлежат объективной «удалённой» оценке. Однако, была выявлена следующая зависимость: в случае, если в построенном рейтинге фактов первые 2-3 места занимали факты, вес которых более чем в 4 раза превышает вес следующих в топе фактов, то с вероятностью 89% данные факты передают особенности конкретного отеля.

4. Оценка качества базы данных по отелям

С использованием ключевых фактов, полученных в результате работы алгоритма, основанного на частотном подходе и описанного в п. 3.1., была проведена предварительная оценка качества базы данных фактов об отелях, полученной компанией «Яндекс» от партнёрских организаций. Под оценкой качества в данном контексте понимается корректность и актуальность информации, содержащейся там в виде фактов.

Сравнение двух множеств фактов (результаты работы предложенного алгоритма и информация из базы данных), относящихся к параметрам «пляж», «интернет» (также «wi-fi») по построенным для оценки случайным выборкам, описанным выше, показало:

- 54% полных совпадений по идентификаторам отелей: все извлечённые факты про отели из этой группы совпали с информацией из партнёрской базы данных.
- 9% фактов, не совпавших с информацией из базы, оказались верны. Таким образом, было выявлено наличие неверной информации в партнёрской базе данных.

5. Эксперименты с машинным переводом

5.1 Расширение набора данных для экспериментов

В настоящее время одним из наиболее востребованных туристических направлений являются европейские страны. Особенно популярными среди туристов из Москвы и Санкт-Петербурга являются короткие поездки в столицы стран Европы. В то же самое время отличительной особенностью туризма по данному направлению является большое количество мелких отелей, рассчитанных на малое число постояльцев, а также редко освещаемых в средствах массовой информации. В результате для данных отелей существует достаточно малое число отзывов в Интернете в целом, при этом отзывов на русском языке может не быть вообще.

В связи с данной проблемой была выдвинута следующая гипотеза: можно использовать существующие инструменты машинного перевода текстов на естественном языке, чтобы собрать достаточное количество отзывов и статей на русском для применения Томита-парсера и алгоритмов извлечения фактов, описанных выше. Появление данной гипотезы обусловлено тем, что структура фактов в русском и английском языках одна:

параметр: характеристика <модификатор истинности утверждения>.

В данной работе для описанных выше целей было выбрано множество из 100 тысяч отзывов на английском языке. Отзывы были собраны с сайта Booking.com. Для автоматического перевода текстов был написан модуль, использующий API сервиса Яндекс.Перевод [9].

5.2 Результаты

В ходе экспериментов по внедрению машинного перевода были получены следующие результаты.

Было расширено множество уникальных идентификаторов отелей, о которых возможно извлекать факты. Из 6.6 тысяч уникальных идентификаторов отелей, англоязычные тексты про которые были собраны с Booking.com, только для 1 тысячи в имеющемся множестве текстов были найдены отзывы на русском языке.

Доверительный интервал для точности извлечения частотных русскоязычных фактов из машинно переведённых текстов – (0.71, 0.83), что показывает перспективность дальнейшего развития данной методики и вероятного её использования в коммерческих сервисах в будущем.

Заключение

В ходе данной работы были получены следующие результаты.

1. Сформулированы требования к данным, необходимым для проведения экспериментов. Собраны соответствующие наборы данных.
2. Разработаны и реализованы два алгоритма для извлечения ключевых фактов различных типов из русскоязычных отзывов и статей об отелях (Java, Python, MySQL).
3. Измерена эффективность данных алгоритмов и описаны области их возможного применения.
4. На основе извлечённых фактов проведена оценка качества базы данных фактов об отелях, полученной от партнёров компании "Яндекс".
5. Проведены эксперименты по извлечению ключевых фактов из набора русскоязычных текстов, дополненного текстами, машинно переведёнными с английского языка на русский. Выполнено сравнение работы первого из описанных алгоритмов (извлечение частотных для выбранной тематики фактов) с использованием и без использования машинного перевода. Показана перспективность дальнейшего развития описанной методики.
6. Результаты данной работы были рекомендованы к представлению в формате постера на конференции RuSSIR 2015.

В качестве дальнейших направлений работы можно выделить:

- проведение аналогичных экспериментов над данными о курортных объектах (страны, города и т.п.);
- использование машинного перевода в более широких масштабах;
- перенос исследований на тексты на других языках;

- оптимизация алгоритмов для работы с более объёмными наборами данных;
- улучшение точности извлечения фактов.

Список литературы

- [1] The World Bank. Internet users (per 100 people). – 2015. – URL: <http://data.worldbank.org/indicator/IT.NET.USER.P2> (дата обращения: 27.05.2015).
- [2] Internet Seen as Positive Influence on Education but Negative on Morality in Emerging and Developing Nations. – 2015. – URL: <http://www.pewglobal.org/files/2015/03/Pew-Research-Center-Technology-Report-FINAL-March-19-20151.pdf> (дата обращения: 23.05.2015).
- [3] Masaru Tomita. Efficient parsing for natural language // Kluwer Academic Publishers, Boston, 1986.
- [4] Что такое Томита-парсер, как Яндекс с его помощью понимает естественный язык, и как вы с его помощью сможете извлекать факты из текстов. – 2014. – URL: <http://habrahabr.ru/company/yandex/blog/219311/> (дата обращения: 01.06.2015).
- [5] Stanford CoreNLP: A Suite of Core NLP Tools. – 2015. – URL: <http://nlp.stanford.edu/software/corenlp.shtml> (дата обращения: 01.06.2015).
- [6] Amy Aung, May Phyo Thwal. Onthology Based Hotel Information Extraction From Unstructured Text // International Conference on Advances in Engineering and Technology (ICAET'2014) March 29-30, 2014 Singapore.
- [7] Raghu Anantharangachar, Srinivasan Ramani, S Rajagopalan. Ontology Guided Information Extraction from Unstructured Text // International Journal of Web & Semantic Technology (IJWesT) Vol.4, No.1, January 2013.

- [8] Christopher D. Manning, Prabhakar Raghavan, Hinrich Schütze. Introduction to Information Retrieval. – Cambridge University Press, 2008.
- [9] Яндекс. Как работает машинный перевод. – 2015. – URL: <https://tech.yandex.ru/translate/doc/intro/concepts/how-works-machine-translation-docpage/> (дата обращения: 02.06.2015).