Автоматическое извлечение терминов из сообщений
Якупова М. М.

Якупова Марина Мансуровна / Yakupova Marina Mansurovna - магистрант,
направление: фундаментальная информатика и информационные технологии,
Институт информационных технологий,
Челябинский государственный университет, г. Челябинск

Аннотация: *в статье описывается эксперимент по автоматическому извлечению однословных и двухсловных терминов и фактов из входящих сообщений (неструктурированного текста), основанный на формулировании правил или шаблонов. Проведены эксперименты на 543 текстах сообщений, относящихся к различным областям (учеба, работа, личные сообщения).*
Ключевые слова: *извлечение терминов, обработка неструктурированных текстов, шаблонный метод извлечения фактов, синтаксические конструкции.*

Введение

Автоматическое извлечение терминов и фактов, то есть слов и словосочетаний, которые несут в себе значимую информацию, является приоритетной задачей, которая относится к области обработки текстов. В настоящее время представлено множество математических моделей и методов извлечения терминов, однако большинство существующих методов могут быть применимы только для конкретных задач и текстов на английском языке.

В данной работе термины и факты следует понимать, как основные характеристики назначения встреч для определения основных характеристик назначения встреч. Например: место встречи, название встречи, дата и время встречи.

Извлечение словосочетаний из текста представляет особую сложность.

Основные методы извлечения фактов

В настоящее время можно выделить три основных подхода, которые применяются в различных экспериментах для извлечения фактов из неструктурированных текстов:

1. Методы на основе онтологии,
2. Методы на основе лингвистических правил или шаблонов,
3. Метод на основе машинного обучения.

Применение методов на основе онтологий требует составления объемных словарей для определения терминов. Онтология описывает типы объектов (классы), взаимосвязи между ними (свойства), и способы совместного использования классов и свойств (аксиомы) [1]. Однако применение данного метода не может быть применимо в текущей задаче. Неструктурированные тексты на естественном языке (входящие сообщения) не поддаются структуризации или формализации, которые требует данный подход.

Методы на основе лингвистических правил или шаблонов является наиболее подходящими для извлечения фактов из неструктурированного текста на русском языке, т.к. если в полученных данных возникают ошибки, то очень просто по шаблонам найти причину и оперативно ее исправить или дополнить. Факты, которые планируется извлекать из текстов (дата и время, название встречи, место встречи), являются стандартизированными объектами, и шаблоны для таких типов информации составляются относительно просто в сравнении с остальными объектами. Шаблоны для стандартизированных объектов при правильной реализации легко дополнять и заменять части, которые не дают высокой точности.

Методы на основе машинного обучения для разметки неструктурированных текстов на русском языке также подходят для данной работы, однако требуют больших научных знаний в области нейронных сетей. Данный подход в настоящий момент только начинает развиваться. Инструменты для автоматической разметки русскоязычных текстов пока не очень развиты, а существующие не всегда легкодоступны [2, с. 7]. Потребуется много времени для создания размеченного корпуса, который бы при необходимости мог сам перенастраиваться и переобучаться. Метод на основе лингвистических шаблонов является наиболее подходящим благодаря быстроте реализации и возможности расширения алгоритмов. Вместо сложного расчета ядер выполняется более простая с точки зрения вычислительной сложности процедура сопоставления с лингвистическим шаблоном [3, с. 268].

Метод извлечения фактов при помощи лингвистических шаблонов

Главной задачей для извлечения фактов из входящих сообщений при помощи метода шаблонов является выделение конкретных сущностей, упомянутых в тексте [4, с. 2]. Под сущностью в тексте следует понимать объект, который обладает всеми или несколькими выявленными атрибутами (Табл. 1).

Таблица 1. Атрибуты для объекта встречи

Атрибут	**Способ извлечения**

Дата встречи	Словарь + Лингвистический шаблон
Время встречи	Лингвистический шаблон
Название встречи	Лингвистический шаблон
Место встречи	Словарь + Лингвистический шаблон
Отправитель	Лингвистический шаблон

Для извлечения даты и времени встречи из входящего сообщения используется малый специализированный словарь, который включает в себя названия месяцев, дней недели и времен года. Для извлечения факта формата «23.05.16» или «23 мая 2016 года» применяется шаблон, основанный на регулярных выражениях: '[0-9]{4}-(0[1-9]|1[012])-(0[1-9]|1[0-9]|2[0-9]|3[01])', '(0[1-9]|[12][0-9]|3[01])[- /.](0[1-9]|1[012])[- /.](19|20)\d\d'.

Для извлечения времени встреч было применено регулярное выражение: '/^([0-1][0-9]|[2][0-3]):([0-5][0-9])$/'.

Для определения сущности и извлечения типа «название встречи» был разработан простой алгоритм, работа которого основывалась на определении в тексте словосочетаний находящихся в кавычках («Конференция N») или начинающихся с заглавной буквы. Если кавычки отсутствуют, то алгоритм ищет слова, начинающиеся с заглавной буквы или аббревиатуры, и определяет прилагательные, которые к нему относятся. Если определенное словосочетание заканчивается на прилагательное, то последующее за ним существительное, также включается в сущность. Для определения атрибута названия встречи было определено 4 основных лингвистических шаблона (Табл. 2).

Таблица 2. Лингвистические шаблоны для атрибута «Название встречи»

Шаблон	**Пример**
[«Сущ. + Прил.»] или [«Прил. + Сущ.»]	«Конференция N»
[Сущ. с заглавной буквы + Прил.] или [«Прил. с заглавной буквы + Сущ.»]	Всероссийская конференция
[Сущ. с заглавной буквы + Прил.] или [«Прил. с заглавной буквы + Прил. + Сущ.»]	Всероссийская конференция или Всероссийская научная конференция
[Аббревиатура + Сущ.] или [Сущ. + Аббревиатура]	IT конференция

Для определения атрибута «Место встречи» все слова из текста сообщения приводятся к нормальной форме функцией normal_form библиотеки pymorphy2 для языка Python. Затем по преобразованному тексту осуществляется поиск на соответствие слов из словаря («аудитория», «университет» и т.д.). Если соответствия найдены, то в атрибут включается найденное слово, последующее слово или если последующее слово прилагательное, то последующие два слова, включая существительное. Например: «аудитория №132», «университет ЧелГУ».

Результаты эксперимента

Для проверки работоспособности разработанной модели извлечения фактов из входящих сообщений (неструктурированных текстов) была составлена тестовая выборка из входящих сообщений различной тематики 10 человек за полгода. Общее количество текстов составило 543 штуки, длиной от 10 до 200 слов.

Из представленных текстов извлечению подвергались факты:

1. Дата и время встречи (встречающихся в тексте сообщения),
2. Название встречи (Например: «пара английского языка», «студенческая конференция»),
3. Место встречи (Например: «кафе Апельсин», «аудитория № 132»).

Успех проведения эксперимента основывался на подсчете численной оценки качества алгоритма для полученных данных (фактов).

Эффективность алгоритма оценивалась с помощью классических метрик: точности и полноты.

$P = (correct + 0{,}5*partial)/\ actual$

$R = (correct + 0{,}5*partial)/\ possible$

Где P – точность; R – полнота;

- correct – количество корректно извлеченных строк базы данных;
- partial – количество частично корректных;

- actual – количество заполненных строк, имеющих пропуски только тех значений атрибутов, которые отсутствуют в тексте;
- possible – количество строк, которые можно извлечь из текстов.

Числовые характеристики эффективности извлечения каждого из атрибутов по отдельности, а также всего события целиком, приведены в таблице (Табл. 3).

Таблица 3. Числовые характеристики эффективности извлечения атрибутов

Атрибут события	Дата	Время	Название	Место
Точность	86	81	63	71
Полнота	70	76	69	64

Оценке качества работы метода проводилась на 3 кластерах: учеба, работа, личные сообщения. Для каждого кластера были выделены факты, которые могут в нем содержаться. Для кластеризации данных был применен метод kNN.

Численная оценка качества обученной модели высчитывалась по формуле: Accuracy=P/N, где P — количество верно принятых решений, N — размер обучающей выборки [5].

При обучающей выборке в 50 текстов параметр accuracy принимал значение 0,65.

Однако увеличив обучающую выборку до 150 текстов, параметр accuracy принял значение 0,87.

При обучающей выборке в 200 слов параметр accuracy принимал значение 0,71. Поэтому для эффективного решения экспериментально определен размер обучающей выборки на 543 текста в 150 единиц.

Заключение

В данной работе предложена модель для извлечения фактов, атрибутов из сообщений на естественном языке и результат алгоритма их кластеризации. Модель извлечения основывается на составлении и реализации лингвистических шаблонов для выделения необходимых атрибутов и на составлении и применении тематических словарей.

Представленный метод может применяться для обработки входящих сообщений из различных сервисов. Например: Яндекс Почта, Google Mail, сообщений из сервиса ВКонтакте или Facebook.

В настоящее время ведутся работы по разработке системы для извлечения сообщений за 6 месяцев из нескольких сервисов, для определения в текстах основных фактов и для визуализации полученных данных в календарь событий.

Литература

1. Linked Data Glossary. W3C Working Group Note 27 June 2013. [Электронный ресурс]. - Режим доступа: https://www.w3.org/TR/ld-glossary/#ontology (дата обращения 28.05.2016)
2. *Пантелеев Ф. М.* Построение автоматизированной системы поиска топонимов в тексте на русском языке. М.: Молодежный Научно-технический Вестник, 2015.
3. Dan Moldovan Domain-Specific Knowledge Acquisition from Text // ACM, 2000.
4. *Котельников Д. С.* Итерационное извлечение шаблонов описания событий по новостным кластерам // Труды конференции RCDL-2012, 2012.
5. Блог Дениса Баженова. Оценка классификатора (точность, полнота, F-мера). [Электронный ресурс] Режим доступа: http://bazhenov.me/blog/2012/07/21/classification-performance-evaluation.html (дата обращения: 16.05.2016).