Автоматическое извлечение ключевых фактов из неструктурированных текстов

(для случая отзывов и статей об отелях)

Михайлова Александра Алексеевна,
544 группа

руководитель: ст. преп. Луцив Д.В.
рецензент: рук. отдела Горовой В.А., ООО «Яндекс»

Анализ текстов с помощью фактов

- Неструктурированные тексты на естественном языке (рус)
- Рейтинги объектов по параметрам
  - Booking.com, TripAdvisor, TrustYOU
  - Яндекс.Путешествия: март 2015
- Примеры: «интернет: беспроводной true», «персонал: воровать false»
- Факты: синтаксические и ключевые
- Текущее решение: базы данных фактов, ручной труд

Постановка задачи

- Проанализировать особенности извлечения ключевых фактов
- Предложить и реализовать алгоритмы для извлечения ключевых фактов из текстов про отели (отзывы, статьи в Интернете), оценить качество
- Сравнить извлекаемые факты с партнёрской базой
- Проверить гипотезу: можно ли улучшить качество, расширив данные машинно переведёнными текстами

Ключевые факты

- Передают особенности предметной области
- Условно делятся на 2 типа
  - Общие для объектов предметной области: «пляж: собственный», «интернет: платный»
  - Специфические для конкретного объекта

    «По пляжу ползают черепахи» -> черепаха: ползать

    «После этого бассейна у меня зелёные волосы!» -> волос: зелёный

Данные

- На основе опыта пользователей поиска «Яндекса»
- Топ наиболее популярных отелей
- «Статьи»: результат поиска по названию отеля
  - 3 тыс. отелей, 150 тыс. статей ➔ 12.4 млн синтаксических фактов
- «Отзывы»: результат поиска с маркером «отзывы» + данные Booking.com
  - 8.3 тыс. отелей, 100 тыс. отзывов ➔ 1 млн синтаксических фактов
- Синтаксические факты: Томита-парсер
- Обучение – статьи

Алгоритм: общие факты

- Идея: общие характеристики – самые частотные факты
- Обучение:
  - Для каждого факта посчитать статистику употребления по всему обучающему множеству
  - Выбрать N% самых частотных фактов, сохранить
  - Эксперименты с порогами, добавление вспомогательных словарей
  - Чем выше порог, тем выше точность, но меньше фактов
- Субъективные факты: «отель: хороший», «отдых: шикарный»
- Оценка: семплинг, $P(0.74 < \text{точность} < 0.89) = 0.95$

Алгоритм: специфические факты

- Идея: модификация метрики TF*IDF – характерность факта для конкретного отеля
- Обучение: предподсчёт
- Для отеля по всем синтаксическим фактам считаем метрику, выбираем топ
- Объективно точность не оценить

Результаты

- Проанализированы особенности извлечения ключевых фактов
- Реализованы два алгоритма для извлечения КФ разных типов, проведена ручная оценка результатов с помощью семплинга
- Проведена оценка качества базы данных фактов про отели от партнёров по параметрам «пляж», «интернет». Совпадения: 54%, 9% несовпадений – неверная информация в базе
- Внедрение машинного перевода Яндекс.Перевод: новые отели, $P(0.71 < \text{точность} < 0.83) = 0.95$
- Работа рекомендована к представлению на RuSSIR 2015