Анализ тональности сообщений

Лидия Михайловна Пивоварова

Системы понимания текста

Введение

- Opinion Mining – извлечение **мнений**, а не фактов:
  - Поиск отзывов о товарах и услугах (как потребителями, так и производителями)
  - Анализ мнений для политологических, социологических и др. исследований
- Другие приложения:
  - Рекомендательные системы
  - Извлечение информации
  - Вопросно-ответный поиск

Общая схема

- Объект O имеет (иерархический) набор свойств f_i
- Каждое свойство может выражаться набором слов/словосочетаний w_i - синонимов
- Субъект (opinion holder) высказывает свое мнение об O или о каких-то его свойствах

Основные задачи

На уровне документа:

- Классификация тональности
  - Классы: позитивный, негативный, нейтральный
  - Предполагается, что каждый документ содержит мнение только об одном объекте и только одного субъекта

На уровне предложения:

- Идентификация предложений, содержащих мнения
- Определение тональности предложения
  - Предполагается, что каждое предложение содержит только одно мнение

На уровне свойств:

- Определение свойств, которые оценивает субъект
- Сгруппировать синонимы (если они неизвестны)
- Идентифицировать тональность оценки

Классификация документов

- Классификация – классическая задача машинного обучения
- Различия с тематической классификацией только в используемых свойствах
  - Наличие терминов и их частота (часто взвешенная)
  - Части речи – для определения тональности принципиально важны прилагательные и наречия
  - Оценочные слова и словосочетания (словарь или более сложная структура типа WordNet)
  - Синтаксические зависимости – позволяют делать предположения о семантических отношениях между оценочными и тематическими словами
  - Отрицания – могут изменить мнение на противоположное

Категоризация документов

- Список оценочной лексики (прилагательные и наречия)
- Для всех упоминаний объекта и/или его свойств рядом с оценочной лексикой, подсчитывается коэффициент взаимной информации:

$$PMI(word_1, word_2) = \log_2\left(\frac{P(word_1 \wedge word_2)}{P(word_1)P(word_2)}\right)$$

- Итоговая оценка для данного упоминания:

$$SO(phrase) = PMI(phrase, \text{“excellent”}) - PMI(phrase, \text{“poor”})$$

- Оценка суммируется для документа в целом

Уровень документа и предложения

- Документ может быть очень противоречивым
- Требуется переход на уровень предложений
- Классификация предложений:
  - Объективные/субъективные
  - И затем негативные/позитивные
- Но: позитивная оценка объекта не означает позитивной оценки всех его свойств (и vice versa)
- Предложения могут быть очень сложными – нужно переходить на уровень отдельных свойств

Оценка свойств

- Идентификация свойств
- Группировка синонимов
- Определение оценок

- Подходы очень похожи на Information Extraction:
  - (Named) Entity Recognition + установление фактов (оценок)
  - Словари, образцы, машинное обучение

Сравнения

- Два вида оценок:
  - *X хороший (плохой, тяжелый, легкий, звонкий…)*
  - *X лучше (хуже, выше, ниже, толще, мощнее…) чем Y*
- Требуют более детальной обработки
- Типы сравнений:
  - Градации
    - *A лучше B*
    - *A такой же как B*
    - *A лучше всех*
  - Сравнения по свойствам
    - *У A есть характеристики, которых нет у B*
    - *У A одни свойства, у B другие*
    - *A похож на B не считая некоторых свойств*

Примеры сравнений

- Ex1: "*car X has better controls than car Y*"
  (relationWord = better, features = controls, entityS1 = car X, entityS2 = car Y, type = non-equal-gradable)
- Ex2: "*car X and car Y have equal mileage*"
  (relationWord = equal, features = mileage, entityS1 = car X, entityS2 = car Y, type = equative)
- Ex3: "*Car X is cheaper than both car Y and car Z*"
  (relationWord = cheaper, features = null, entityS1 = car X, entityS2 = {car Y, car Z}, type = non-equal-gradable)
- Ex4: "company X produces a variety of cars, but still best cars come from company Y"
  (relationWord = best, features = cars, **entityS1** = company Y, entityS2 = null, type = superlative)

Построение словарей

- Вручную
- На основе существующих словарей и тезаурусов (WordNet)
- Автоматически
  - Bootstrapping
  - *Она умная и красивая* vs. *Она умная, но вредная*
  - Возможно построение доменно-ориентированных словарей

Источники

- Liu B. Sentiment Analysis and Subjectivity // Handbook of natural language processing, Second Edition Editor(s): Nitin Indurkhya; Fred J. Damerau, Goshen, Connecticut, USA – 2010 – pp. 627-666
- Bing Liu Web Data Mining. Lecture Slides, Chapter 11 – http://www.cs.uic.edu/~liub/WebMiningBook.html
- Bing Liu Opinion Mining and Summarization, tutorial - http://www.cs.uic.edu/~liub/FBS/opinion-mining-sentiment-analysis.pdf
- Bo Pang and Lillian Lee Opinion mining and sentiment analysis // Foundations and Trends in Information Retrieval 2(1-2), pp. 1–135, 2008. – http://www.cs.cornell.edu/home/llee/opinion-mining-sentiment-analysis-survey.html