Методы автоматического извлечения оценочной лексики для заданной предметной области

Семинар SIGMOD 2013

Четвёркин Илья

Аспирант ВМК

МГУ им. М.В. Ломоносова

План презентации

- **Введение**
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Автоматическая обработка

- Огромное количество субъективных данных
  - Структурированные отзывы и неструктурированные тексты
- Поиск и извлечение полезной информации
  - Для бизнеса: лояльность к продуктам и бренду
  - Для людей: помощь в принятии решений
- Сложная задача!
  - Применяется весь спектр методов NLP
  - Большое количество подзадач

Sentiment analysis

- Основные направления исследований
  - Классификация текста по субъективности \ объективности
  - **Классификация текста по тональности**
  - Классификация по выражаемым эмоциям в тексте
  - Выделение саркастических предложений
  - Аннотирование отзывов
- Разные уровни детализации (фрагмент, предложение, документ)

План презентации

- Введение
- **Классификация текстов по тональности**
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Классификация отзывов

Это какой-то ужас⁻. В рецензии все описано так, что от фильма ждешь минимум продолжения фильма "Адреналина" с Стэтхэмом, но нет. 80 минут бесмысленных⁻ бегов, туда-сюда. Единтсвенный вопрос - Зачем?!?!

Хороший+, трешовый+ фильм, с отличным+ чувством юмора. Для любителей гая ритчи самое то, вот только картинка нищенская⁻, но ничего страшного+. Это даже колорит+ какой-то придает.

Я в диком восторге+!!!!

Эта парочка неподражаема+!

Так всё красиво+, аккуратно+, технично+, с долей юмора, и ТАК ЗАХВАТЫВАЮЩЩЩЩЩЩЩЕ динамично+!!!! В экстазе+)

Классификация отзывов

Это какой-то ужас⁻. В рецензии все описано так, что от фильма ждешь минимум продолжения фильма "Адреналина" с Стэтхэмом, но нет. 80 минут бесмысленных⁻ бегов, туда-сюда. Единтсвенный вопрос - Зачем?!?!

Хороший+, трешовый+ фильм, с отличным+ чувством юмора. Для любителей гая ритчи самое то, вот только картинка нищенская⁻, но ничего страшного+. Это даже колорит+ какой-то придает.

Я в диком восторге+!!!!

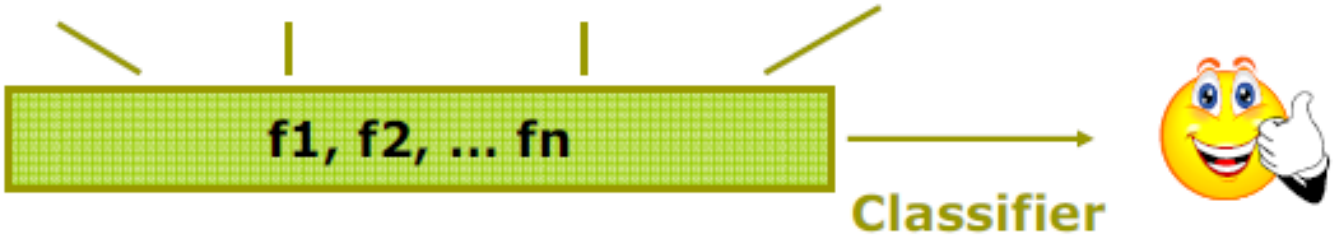
Эта парочка неподражаема+!

Так всё красиво+, аккуратно+, технично+, с долей юмора, и ТАК ЗАХВАТЫВАЮЩЩЩЩЩЩЩЕ динамично+!!!! В экстазе+)

Основные подходы к решению

- Агрегация оценочных выражений (+правила их комбинирования)
  - Хорошая история и отличное представление!

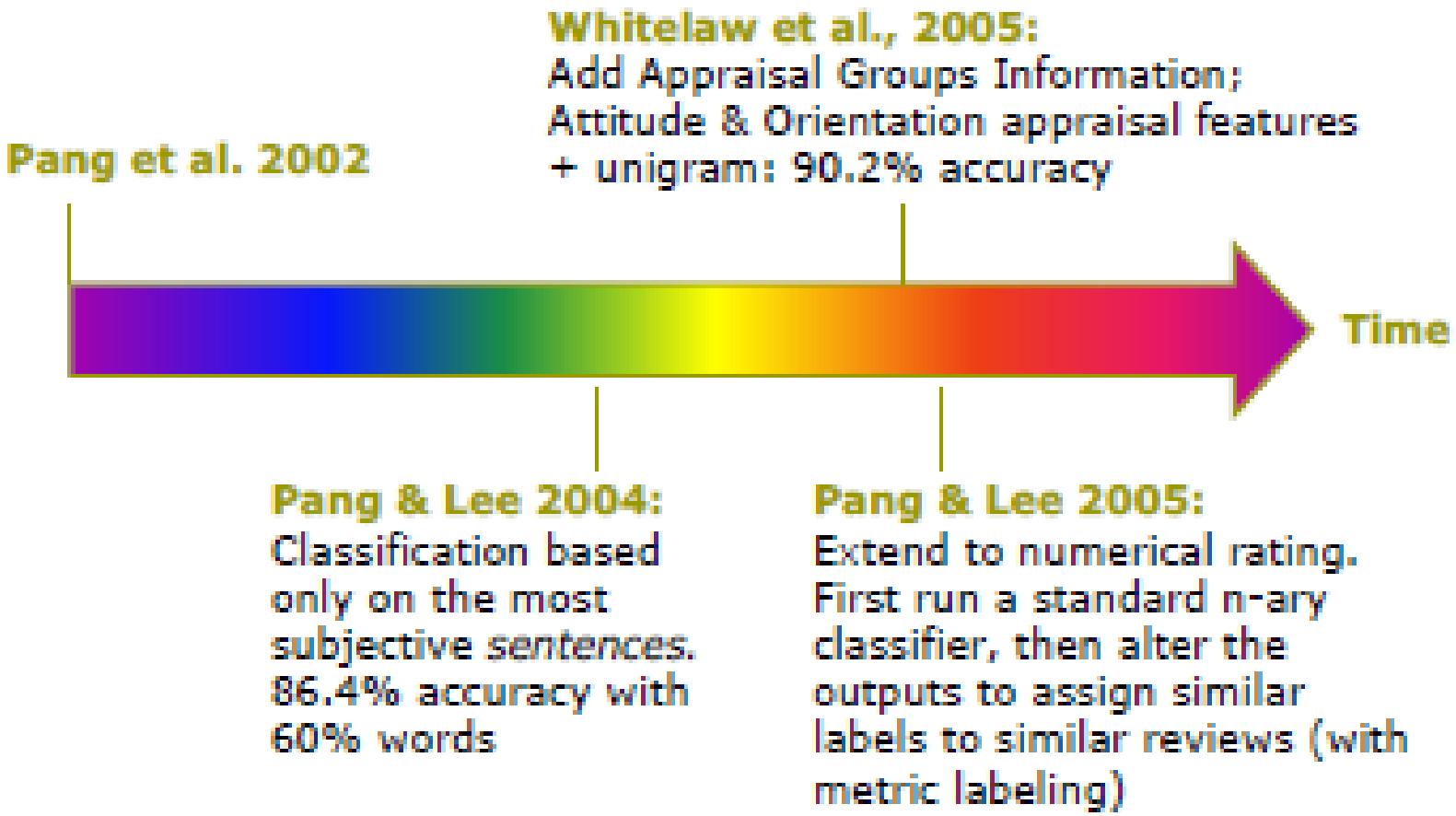
- Один глобальный классификатор
  - Хорошая история и отличное представление!



Использование глобального классификатора

Использование глобального классификатора

- Используется документ как целое
- Не требуется выделение оценочных выражений и их тональной ориентации
- Нужны признаки для классификации
- Нужна размеченная коллекция документов


Whitelaw et al., 2005:
Add Appraisal Groups Information;
Attitude & Orientation appraisal features
+ unigram: 90.2% accuracy
Pang et al. 2002
Time
Pang & Lee 2004:
Classification based only on the most subjective sentences.
86.4% accuracy with 60% words
Pang & Lee 2005:
Extend to numerical rating.
First run a standard n-ary classifier, then alter the outputs to assign similar labels to similar reviews (with metric labeling)

Классификация мнений на основе оценочных выражений

Оценочные выражения

- Отдельные слова
  - Прилагательные – много оценочных слов (*пристойный,умилительный, мерзопакостный*)
  - Существительные (*нудятина, галиматья, удачность*)
  - Глаголы (*взбесить, восторгать*)
  - Наречия (*выигрышно, метко*)
- Фразы
  - *Обязателен к просмотру*
  - *Лишен достоинств*
  - *Убивать всю интригу*

Классификация на основе оценочных выражений

- Необходим словарь оценочных слов с тональностями
  - Общий
  - Для конкретной предметной области
- Правила комбинирования оценочны слов
  - Учет отрицаний
  - Учет слов-операторов (усиливающих или меняющих оценку)
  - Учет синтаксической структуры

План презентации

- Введение
- **Классификация текстов по тональности**
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Дорожки РОМИП 2011-2012

- Классификация документа целиком (например отзыва) по общей оценке выраженной автором
  - Классы: Положительный и Отрицательный (возможно Нейтральный)
  - 5 классов, от наиболее положительного до отрицательного
  - Вероятно наиболее широко изучаемая проблема
- Поиск мнений по запросу в коллекции блогов
- Классификация прямой и косвенной речи из новостных цитат

Некоторые выводы РОМИП

- Наиболее распространенным и эффективным алгоритмом классификации отзывов является метод опорных векторов
- Комбинирование SVM с различными лексическими и статистическими характеристиками улучшает результат
- При переносе классификатора на другую предметную область качество существенно падает
  - Инженерный подход работает более устойчиво

План презентации

- Введение
- **Классификация текстов по тональности**
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Проблемы и преимущества методов

- Методы машинного обучения
  + Высокое качество классификации
  – Необходима размеченная коллекция
  – Плохая переносимость модели
  – Сложно исправлять ошибки (дообучать)
- Методы на основе словарей и правил
  + Устойчивое качество работы
  + Легко исправлять ошибки
  – Необходим словарь оценочных слов
  – Низкое качество классификации

Переносимость классификаторов

- Оценочные слова в области товаров и услуг отличаются от оценочных слов в новостях
  - Слова *зло, предательство* не являются оценочными в области товаров и услуг
- Классификатор тональности обученный в одной области работает намного хуже в другой
- ***Актуальной*** является задача автоматического извлечения оценочных слов для заданной предметной области

План презентации

- Введение
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- **Постановка задачи извлечения оценочных слов**
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Постановка задачи

- Автоматизированное построение словаря оценочных слов для заданной предметной области
  - Выявить специфические черты оценочной лексики
  - Построить модель оценочных слов
- Проверить переносимость модели на различные предметные области
- Извлечь реальные словари оценочных слов
- Показать полезность извлеченных знаний на реальных задачах анализа мнений

План презентации

- Введение
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- **Постановка задачи извлечения оценочных слов**
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Обзор методов построения словаря

- Построение словаря экспертами вручную
- Использование различных словарей и онтологических ресурсов
- Использование различных правил и закономерностей в поведении оценочных слов для их извлечения из коллекции текстов
- Комбинация нескольких подходов

Словарь, созданный вручную

- Эксперты
  - Выбирают оценочные слова и выражения
  - Проставляют их тональность
- Плюсы
  - Перенос человеческого знания
  - Относительная надежность
- Минусы
  - Зависимость от предметной области
  - Эксперты не могут вспомнить все оценочные многословные выражения
  - Трудоёмкость

Подход на основе словарей

- Использование словарей содержащих отношения между словами (синонимы, антонимы, гипонимы) (Hu & Liu, 2004)
  - Вручную составляется базовое множество слов, затем расширяется с помощью словаря
  - Основной принцип: если слово оценочное, то его синонимы и антонимы будут оценочными
- Использование толковых словарей с описанием сущностей (Esuli & Sebastiani, 2005)
  - Слова имеющие одинаковую ориентацию имеют «похожее» толкование

Корпусный подход

- Поиск правил и закономерностей употребления оценочных слов в текстах
  - Поиск слов часто встречающихся со словами «хорошо» и «плохо» (Turney, 2002)
  - Использование союзов И, ИЛИ, НО (Hatzivasiloglou, 1997)
  - Поиск близких по контекстам слов (Velikovich et al, 2010)
- Методы машинного обучения для классификации слов (Chetviorkin & Loukachevitch, 2012)
  - Набор признаков качественно описывающий поведение оценочных слов

План презентации

- Введение
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- **Постановка задачи извлечения оценочных слов**
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Модель оценочных слов

- На основе четырех текстовых коллекции для каждого слова извлекается набор признаков
- Основные характеристики оценочных слов, учтенные в предлагаемой модели:
  - Оценочные слова чаще встречаются в отзывах, чем в описаниях объектов, либо в новостях
  - Оценочные слова чаще встречаются в отзывах с некоторой тональностью, чем равномерно во всех типах отзывов
  - Оценочные слова чаще являются наречиями или прилагательными
- Применяется комбинация методов машинного обучения для классификации слов

Текстовые коллекции

- Коллекция отзывов о фильмах
  - 28773 отзыва о фильмах различных жанров, собранные с рекомендательного портала *www.imhonet.ru*
  - Пользовательская оценка для каждого текста (десятибалльная шкала)
- Коллекция описаний фильмов
  - 17680 описаний фильмов
- Новостная коллекция
  - Около двух миллионов новостных документов
- Малый корпус
  - Составлен из частей коллекции отзывов

Opinion Word Concentration

Коллекция мнений & Малая коллекция

глава государства в новогоднем поздравлении к украинскому народу, заявляет его пресс-служба в пятницу. По словам Януковича, украинцы, как честные и работящие люди, заслужили достойную жизнь, а у детей должно быть счастливое детство и радостное будущее. Президент также отметил, что старшее поколение, которое создавало независимость страны и создавало ее богатство, также имеет право на то, "чтобы осень и зиму своей жизни провести в атмосфере тепла, уважения, достатка". Вместе с тем, глава государства [illegible] в стране. [illegible] экономические преобразования " не будут проходить быстро и просто". "Даже в новогоднюю ночь не имею права рассказывать сказку. Скажу, как есть: будет у нас много работы. Но у нас есть и много энергии. У нас много веры в наше государство, у нас много надежд в свои силы и на божью помощь", - подчеркнул В.Янукович. "И с этим мы идем в Новый год. И с этим мы вновь победим и будем счастливы", - заявил Президент. В своем поздравлении он назвал уходящий 2010 год непростым, но счастливым. "Нам

Высокая концентрация оценочных слов!

Коллекция новостей & Коллекция описаний

Андрей Коновалов. Он сопровождает министра энергетики Сергея Шматко в поездке по Подмосковным населенным пунктам, отключенным от электроснабжения. По словам Коновалова, последними были запитаны подстанции 35 кВт Андреево (восток Подмосковья) и [illegible] (юг Подмосковья). [illegible] Владимира Путина Шматко уточнил, что неподключенными оставались именно эти две подстанции. "Теперь основная проблема лежит в зоне распределительных сетей 6 -10 кВт. Они наиболее пострадали от наледи", - сказал Коновалов. В целом, по его словам, восстановление электроснабжения Подмосковья выходит на финишную прямую.

Низкая концентрация оценочных слов!

Характеристики оценочных слов для классификации

Частотные характеристики

- Частотность леммы в коллекции
- Документная частотность
- Частотность слов с большой буквы
- Странность = $\frac{P_s(w)}{P_g(w)}$
- TFIDF

Всего15 характеристик на основе 4х коллекций

Примеры отзывов с оценками

Женя Филимонов 7 августа 2012 #

!!!!!!

★ Оценка: 9

Sanechka608 6 августа 2012 #

Много слышала положительных отзывов о фильме, но руки все не доходили, но все же я его посмотрела и не пожалела, гениальный фильм, прекрасная игра актеров, фильм до последневого держит в напряжении, ни секунды не пожплела что его посмотрела. Гениальное кино! Всем советую посмотреть!

★ Оценка: 10

british-idler 6 августа 2012 #

Один из лучших!

★ Оценка: 10

Obsuzhdatel 4 августа 2012 #

Фильм не может не понравится, любой зритель оценит его по достоинству. Книга тоже супер.
Читайте больше книг, там все интересней!

★ Оценка: 10

Характеристике на основе оценок

- Отклонение от средней оценки

$$Dev(w) = \left| E(c \mid w) - E(c) \right|$$

- Дисперсия оценки слова

$$Var(w) = E(c^2 \mid w) - E(c \mid w)^2$$

- Вероятность отнесения слова к классу

$$Lhc(w) = \log \frac{P(w \mid c)}{P(w)}$$

Морфологические характеристики

- Часть речи
  - Существительное, Прилагательное, Глагол, Наречие
- Есть ли у данной леммы морфологическая неоднозначность
  - Встречается ли лемма в разных частях речи в зависимости от контекста
- Найдено ли слово в морфологическом словаре
- Начинается ли слово на приставки: БЕЗ, БЕС, МАЛО, МНОГО, НЕ, НИЗКО, ОДНО, ПЕРЕ

Алгоритмы классификации и оценка качества

Алгоритмы

- Для обучения алгоритмов классификации 18,362 слов *в предметной области о фильмах* были вручную размечены двумя асессорами
  - 4079 слов были отобраны как оценочные
- Решалась задача классификации слов на два класса: оценочные слова и неоценончые
- Мы использовали следующие алгоритмы: *Logistic Regression*, *LogitBoost, Random Forest*

Оценка качества

- Для оценка качества извлеченных списков слов мы использовали метрику *Precision@n*
  - Удобна для оценки качества комбинации списков
  - Может быть использована с различными порогами

Logistic Regression	LogitBoost	Random Forest	Avg
75.7%	75.3%	72.4%	81.5%

Превосходит все характеристики (>= 20%)

План презентации

- Введение
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- **Перенос на другие предметные области**
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Перенос модели

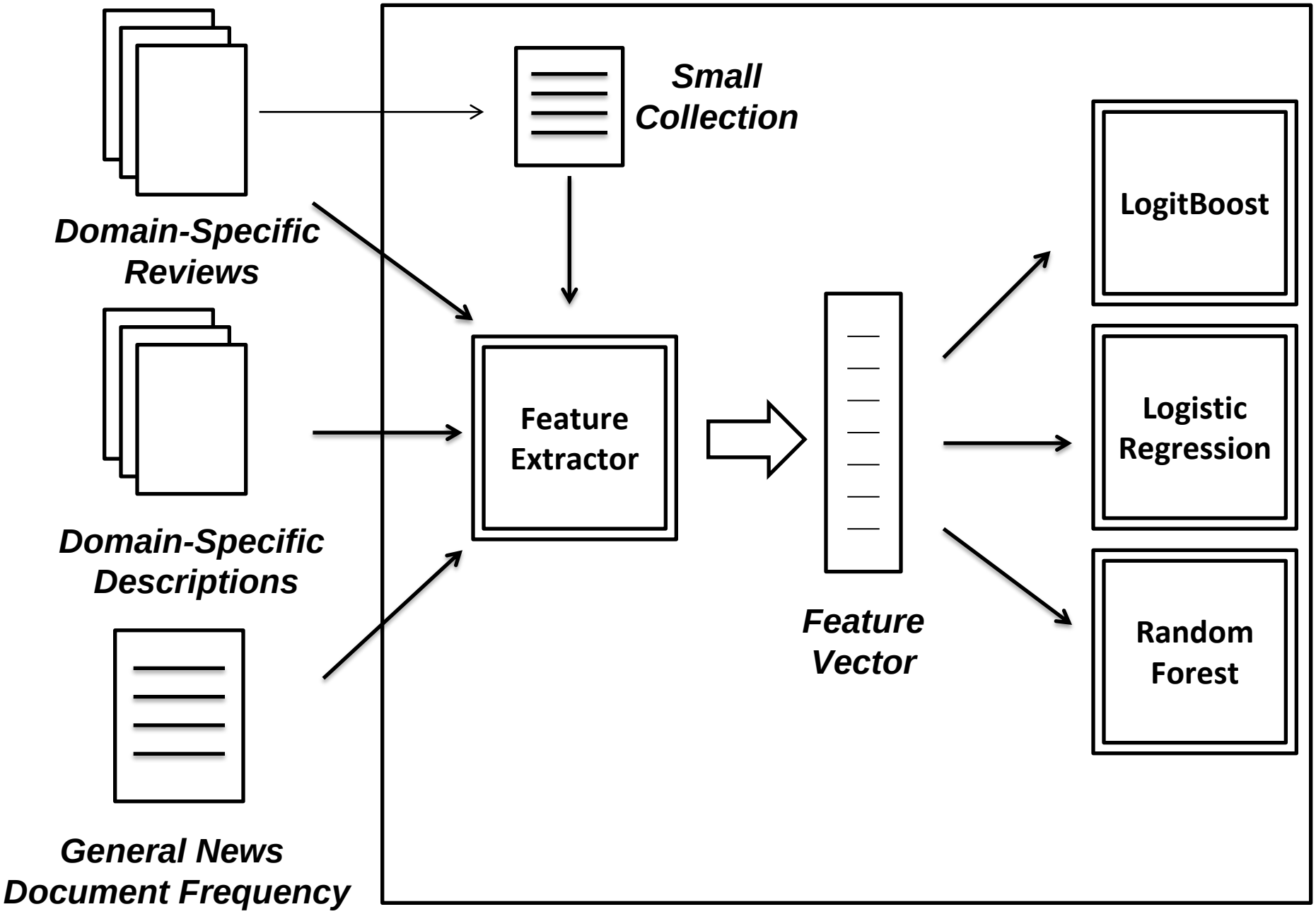
- Четыре новые предметные области:
  - Книги, компьютерные игры, мобильные телефоны и цифровые фотокамеры
  - Структура коллекций в этих областях такая же как и в предметной области о фильмах
- Каждое слово из этих предметных областей преобразуется в вектор характеристик
- Модель, обученная в предметной области о фильмах, используется для классификации слов в других областях

Коллекции

	Коллекция отзывов	Коллекция описаний	Источник
Фильмы	28,773	22, 321	Imhonet
Книги	23, 883	22, 321	Imhonet
Игры	7, 928	1, 853	Imhonet
Цифровые Фотокамеры	10, 208	920	Yandex Market
Мобильные Телефоны	30, 620	890	Yandex Market

Результаты переноса

Предметная область	**P@1000**
Фильмы	81.5%
Книги	86.0%
Игры	72.2%
Цифровые Фотокамеры	62.0%
Мобильные Телефоны	73.2%


Small Collection
Domain-Specific Reviews
Domain-Specific Descriptions
General News Document Frequency
Feature Extractor
Feature Vector
LogitBoost
Logistic Regression
Random Forest

План презентации

- Введение
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- **Обобщенный список оценочных слов**
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

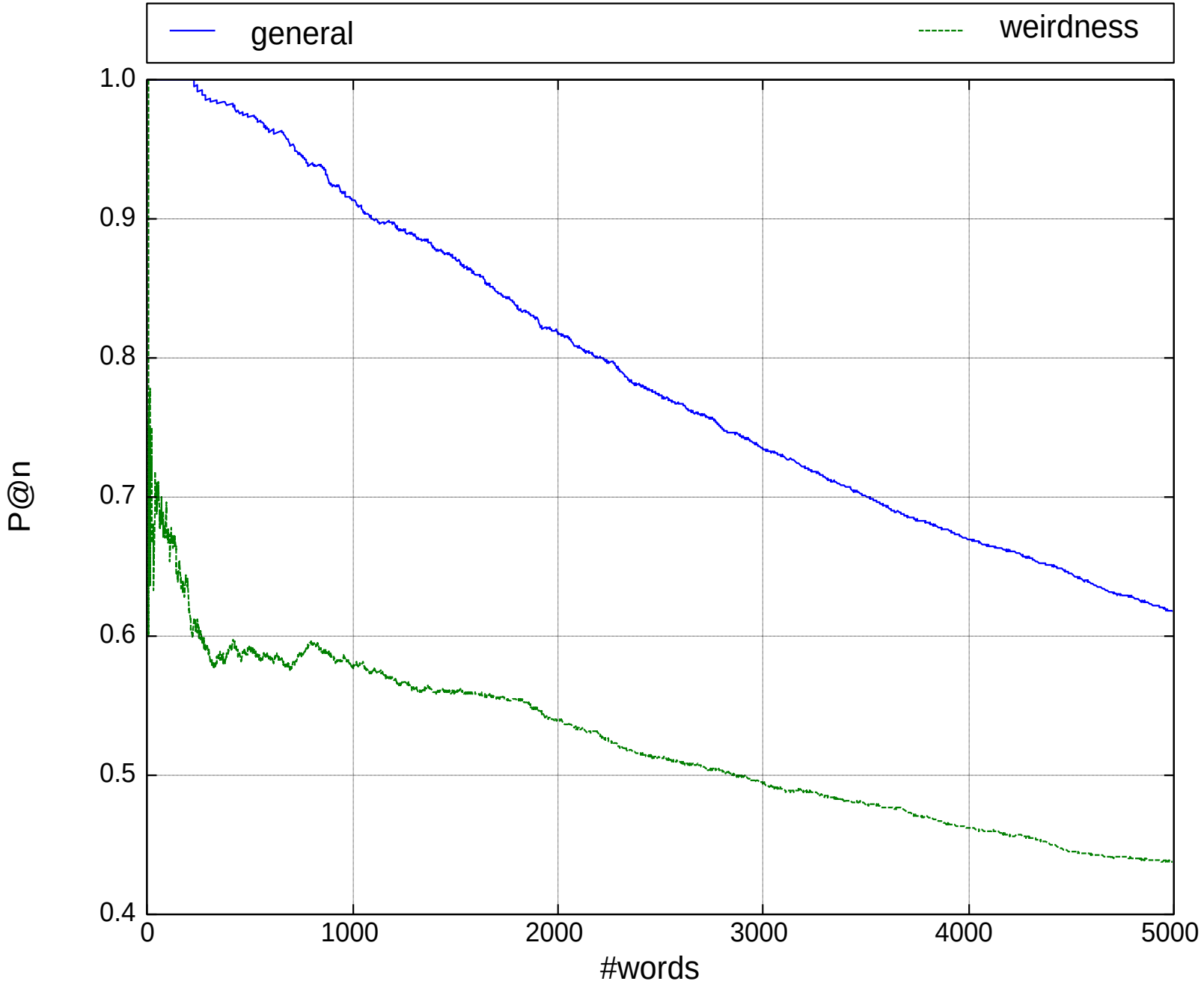
Обобщенные список

- Для создания обобщенного словаря оценочных слов для продуктов и услуг мы скомбинировали списки оценочных слов из *5 предметных областей*
- Мы хотели увеличить вес слов, которые встречаются в большом количестве областей и имеют высокий вес в каждой области

$$R(w) = \max_{d \in D}(prob_d(w)) \cdot \sum_{d \in D} \frac{1}{|D|} \cdot \left(1 - \frac{pos_d(w)}{|d|}\right)$$

Обобщенные список

- Оценка качества списка согласно метрике *Precision@1000* равна **91.4%.**
- Этот обобщенный список оценочных слов состоит из слов реально используемых в отзывах пользователей
  - Для создания такого рода списков не требуется никаких словарных ресурсов
  - http://www.cir.ru/SentiLexicon/ProductSentiRus.txt
- В качестве baseline был взят комбинированный список *Weirdness*


general
weirdness
1.0
0.9
0.8
0.7
0.6
0.5
0.4
P@n
0
1000
2000
3000
4000
5000
#words

Примеры слов из обобщенного списка

- *БЕСПОДОБНЫЙ*
- *НЕВНЯТНЫЙ*
- *ОТЛИЧНЕЙШИЙ*
- *ОБАЛДЕННЫЙ*
- *БЕЗУМНО*
- *НЕПОНЯТНО*
- *НЕПРИЯТНО*
- *ОТВРАТНЫЙ*
- *НЕЖНЫЙ*
- *ПОСРЕДСТВЕННЫЙ*
- *ШИКАРНЫЙ*
- *НЕЛОГИЧНЫЙ*
- *КЛЕВЫЙ*
- *СРЕДНЕНЬКИЙ*
- *НЕПОНЯТНЫЙ*
- *СЛАБЕНЬКИЙ*
- *НЕЕСТЕСТВЕННЫЙ*
- *НЕПЛОХО*
- *СУПЕРСКИЙ*
- *НЕПЛОХОЙ*

План презентации

- Введение
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- **Обобщенный список оценочных слов**
  - Использование словаря в разных задачах
- Вычисление оценок для оценочных слов

Задача переноса классификатора

- Для оценки полезности обобщенного списка оценочных слов мы использовали его в задаче переноса на разные области классификатора тональности
  - Области: *Фильмы, Книги, Цифровые камеры, Мобильные телефоны*
- Во всех парах предметных областей извлеченный словарь показывал лучшее качество
- Наилучший результат был получен при переносе классификатора с *Мобильных телефонов* на *Цифровые камеры* ***+3.25%***

Задача извлечения мнений

- Наилучшие результаты в задаче извлечения мнений были получены с использованием обобщенного списка оценочных слов
- Каждый пост получал вес в соответствии:

$$Weight = \alpha \cdot (\sum_{w \in q} tfidf_w + \sum_{w \in q} tfidf_w^{header}) + (1-\alpha) \cdot SentiWeight$$

  - Оптимальный α = 0.6
  - *SentiWeight* вычислялся как доля слов из списка в тексте

	Фильмы	**Книги**	**Камеры**
p@1	30.0%	44.0%	49.4%

План презентации

- Введение
- Классификация текстов по тональности
  - Подходы к решению
  - Семинар РОМИП
  - Проблемы и преимущества методов. Переносимость
- Постановка задачи извлечения оценочных слов
  - Обзор методов построения словаря
  - Признаки и модель оценочных слов
- Перенос на другие предметные области
  - Система по извлечению оценочных слов *DomEx*
- Обобщенный список оценочных слов
  - Использование словаря в разных задачах
- **Вычисление оценок для оценочных слов**

Вычисление оценок

- Используем модель Марковских случайных полей для моделирование связей между словами
- Построение графа связей между оценочными словами
  - Если два слова часто встречаются близко → схожая тональность
  - Чем дальше слова друг от друга, тем больше распределение приближается к равномерному
- Начальная тональность каждого слова задается его средней оценкой по коллекции текстов

Вычисление оценок

- Общая энергия системы:

$$E(x,W) = -\sum_{ij} w_{ij} x_i x_j - \sum_i h_i x_i$$

  - w_{ij} вес связи между словами i и j, обратно пропорциональный среднему расстоянию между ними
  - h_i отклонение от средней оценки слова в коллекции
- В построении не участвуют какие-либо словарные ресурсы или ручная разметка
  - Алгоритм может быть использован в любой предметной области!

Вычисление оценок

- Для поиска MAP оценки распределения вероятностей использовался алгоритм распространения доверия
- Правильность кластеризации составила *82.7%*, прирост относительно baseline ***5.5%***
- Всего в оценке принимало участие 700 слов
  - Слабая связанность графа
  - Нейтральные и неоднозначные слова

Примеры оценочных слов с тональностью

- ОТЛИЧНЕЙШИЙ +
- НЕСПЕШНЫЙ +
- БЕЗДАРНЫЙ -
- НЕПРИЯТНО -
- НЕПОВТОРИМЫЙ +
- ОБАЛДЕННЫЙ +
- НЕПОНЯТНО -
- ОДНООБРАЗНЫЙ -
- НЕЛЕПО -
- ЗАТЯНУТЫЙ -
- НЕЛОГИЧНЫЙ -
- БЕСПОДОБНЫЙ +
- ПОСРЕДСТВЕННЫЙ -
- ВОСХИТИТЕЛЬНЫЙ +
- НЕЕСТЕСТВЕННЫЙ -
- ШИКАРНЫЙ +
- НЕОБЫЧНО +
- НЕОДНОЗНАЧНЫЙ +
- СКУЧНЫЙ -
- НЕПРЕДСКАЗУЕМЫЙ +

Заключение

- Был предложен новый метод автоматического извлечения оценочных слов на базе нескольких текстовых коллекций
- Были построены словари оценочных слов для нескольких предметных областей и создан обобщенный список оценочных слов для продуктов и услуг
- Данный словарь был оценен асессорами, с качеством P@1000 = **91.4**% и использован в разных задачах обработки мнений: задаче переноса классификатора, извлечение мнений из блогов
- Для заданного списка оценочных слов был предложен и оценен алгоритм определения тональности слова

Спасибо за внимание!

Вопросы?