ИСПОЛЬЗОВАНИЕ СЕМАНТИЧЕСКИХ КАТЕГОРИЙ В ЗАДАЧЕ КЛАССИФИКАЦИИ ОТЗЫВОВ О КНИГАХ

Отчет об участии в дорожке по анализу мнений на РОМИП’2012

Поляков П.Ю., Фролов А.В., Плешко В.В.

ООО «ЭР СИ О» (RCO)

План

- Постановка задачи – оценка тональности отзывов о книгах
- Описание подхода
  - Построение обучающей выборки
  - Извлечение классификационных признаков
  - Классификация на 2 и 3 класса
- Результаты
- Анализ ошибок и возможности улучшения
- Заключение

Оценка тональности отзывов о книгах

- Обучающая выборка
  - Imhonet.ru
  - 24 160 отзывов
  - Оценка от 1 до 10 баллов
- Задачи
  - разделить отзывы на положительные и отрицательные (2 класса)
  - разделить отзывы на 3 класса: "положительный", "средний" и "отрицательный"
- Тестовая выборка
  - ППБЯ
  - 16821 отзыв

Построение обучающей выборки

- Двое экспертов (expert-1, expert-2, expert-and)
  - негативный отзыв – только отрицательные характеристики
  - позитивный отзыв – только положительные характеристики
  - средний отзыв – как положительные, так и отрицательные, либо нейтральные характеристики
- Оценено порядка 4000 отзывов
  - Позитивные > Негативные > Средние
- Согласованность оценок экспертов – ок. 80%.

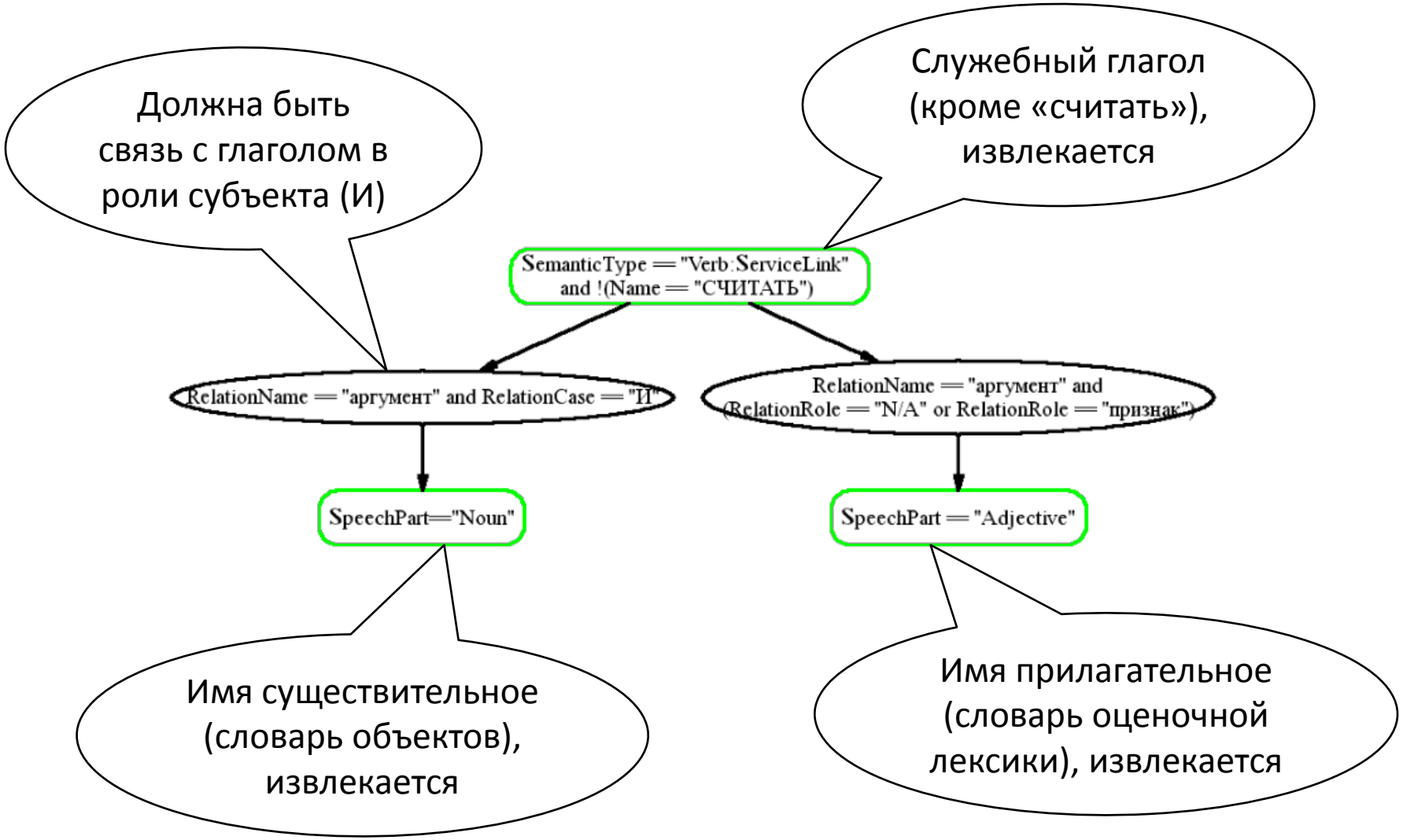
Извлечение классификационных признаков

- базовый метод
  - Леммы (*класс*, *отстой*, *удивлять*)
  - Простые именные группы (*глубокая мысль*, *классика жанра*)
  - Конструкций с предлогами в соответствии с моделями управления (*взгляд на мир*, *книга для детей*)
- термины, выделенные в рамках экспертно-лингвистического подхода с использованием словарей оценочной лексики

Экспертно-лингвистический подход

- Объекты
  - Книга (синонимы, гипонимы)
  - Составляющие книги: автор, концовка, сюжет, язык, герои, впечатления от прочтения (синонимы)
- Оценочная лексика
  - глагол, прилагательное
  - положительная, отрицательная, «средняя» (категории)
- Шаблоны
  - описывает реализацию связи между объектом и оценкой
  - поиск изоморфного фрагмента в дереве синтаксического разбора
- Результат = категория объекта + категория оценки
  - *Книга оказалась достаточно интересной.*
    → *книга, положительная характеристика*
  - *Эти писатели стали культовыми еще в 60-е годы.*
    → *автор, положительная характеристика*

Пример шаблона



Категории оцениваемых объектов и оценочная лексика

ЧИТАТЬ, ЧТЕНИЕ , НАСТРОЕНИЕ, ПРОЧТЕНИЕ, ВРЕМЯ, ЯЗЫК КНИГИ, СЮЖЕТ, ГЕРОЙ, АВТОР, КОНЦОВКА, ПРОИЗВЕДЕНИЕ

Subject	**Quality Verb**	**Quality Emotion**	**Quality Adjective**
КОНЕЦ КНИГИ *КОНЦОВКА* *ФИНАЛ* *РАЗВЯЗКА* *ХЭППИЕНД* *ХЭППИ ЭНД* *ХЭППИ* *ХЭППИ-ЭНД*	*УБИТЬ* *ИДТИ* *РАСТЯГИВАТЬСЯ* *СДЕЛАТЬ* *НЕ ПОНРАВИТЬСЯ*	*ЖДАТЬ* *УБИТЬ* *НЕ ЖДАТЬ* *ИДТИ* *РАСТЯГИВАТЬСЯ* *НЕ ПОНРАВИТЬСЯ* *НАЗВАТЬ*	*УМОПОМРАЧИТЕЛЬНЫЙ* *ОПТИМИСТИЧНЫЙ* *ДУРАЦКИЙ* *ЗАКРЫТЫЙ* *УТОМИТЕЛЬНЕЙШИЙ* *СКУЧНЫЙ* *ПЕЧАЛЬНЫЙ*

Классификация на 2 и 3 класса

- Линейная регрессия (regression)
  - Аппроксимация численных оценок (svm-regression)
  - Поиск пороговых значений для максимизации F-меры на обучающей выборке
- Классификация (one-per-class)
  - Редукция one-vs-rest (svm-linear) по числу классов
  - Класс по умолчанию – положительный

Результаты – 2 класса

2-class	**P-macro**	**R-macro**	**F-macro**	**Accuracy**
SVM	0.68	0.62	0.65	0.86
Regression	0.63	0.63	0.63	0.83

2-class	**P-pos**	**R-pos**	**F-pos**	**P-neg**	**R-neg**	**F-neg**
SVM	0.90	0.95	0.92	0.45	0.29	0.36
Regression	0.90	0.90	0.90	0.35	0.35	0.35

Результаты – 3 класса

3-class	**P-macro**	**R-macro**	**F-macro**	**Accuracy**
SVM	0.54	0.55	0.55	0.70
Regression	0.35	0.33	0.34	0.54

3-class	**P-pos**	**R-pos**	**F-pos**	**P-neg**	**R-neg**	**F-neg**	**P-neu**	**R-neu**	**F-neu**
SVM	0.34	0.70	0.46	0.40	0.22	0.29	0.89	0.74	0.81
SVM	0.28	0.55	0.37	0.20	0.11	0.14	0.87	0.74	0.80
Regression	0.15	0.25	0.19	0.09	0.11	0.10	0.86	0.72	0.78
Regression	0.12	0.25	0.16	0.08	0.11	0.10	0.86	0.64	0.74

Источники ошибок

- Автор большую часть рецензии пересказывает содержание книги. В этом случае, в тексте может содержаться достаточное количество шумовой лексики, чтобы классификатор выбрал неверное решение.
- Автор перечисляет положительные стороны произведения, но итоговую оценку даёт негативную. Пример: “Сюжет есть. И интрига присутствует. А вот то, как разворачиваются действия - не вдохновляет ни коим образом.” В итоге, положительная лексика перевешивает негативную за счет количества.
- Автор ссылается на положительные отзывы других людей, но собственную оценку даёт негативную. Пример: “С сожалением сообщаю: не для моих мозгов. Говорят, книга очень хорошая. Промолчу.”
- Обучающая выборка: lmhonet.ru, тестовая выборка: персональные блоги.

Сравнение результатов на прошлогодней дорожке

	Expert 1 F-macro	Expert 2 F-macro
New hybrid SVM	0.50	0.50
Old hybrid SVM	0.47	0.48

Возможности улучшения метода

- Для правильной оценки больших рецензий необходимо уметь корректно выделять резюмирующую оценку. Большую часть занимает пересказ сюжета или же отвлеченные рассуждения. В то время как реальная оценка делается либо в первых нескольких предложениях, либо в последних.
- Необходимо выделять объект рецензии. В одном тексте может “рецензироваться” и сравниваться между собой множество книг. Пример: “Сегодня я читал X и мне не понравилось. Гораздо хуже замечательной книги Y, которую я читал вчера”.
- При разделении отзывов на два класса необходимо отделять мнение автора рецензии от характеристик из внешних источников (“говорят книга хорошая, но мне не очень понравилась”). В этом случае мнение автора должно иметь больший вес.

Заключение

- Усовершенствован метод обогащения классификационных признаков в рамках лингвистического подхода с применением словарей оценочной лексики и машинного пополнения фильтров
- Проанализированы и классифицированы основные ошибки допускаемые классификатором

Вопросы…

info@rco.ru

www.rco.ru