

Автоматическое определение тональности объектов с использованием семантических шаблонов и словарей тональной лексики

Поляков П.Ю., Калинина М.В., Плешко В.В.
ООО «ЭР СИ О» (RCO)

Постановка задачи

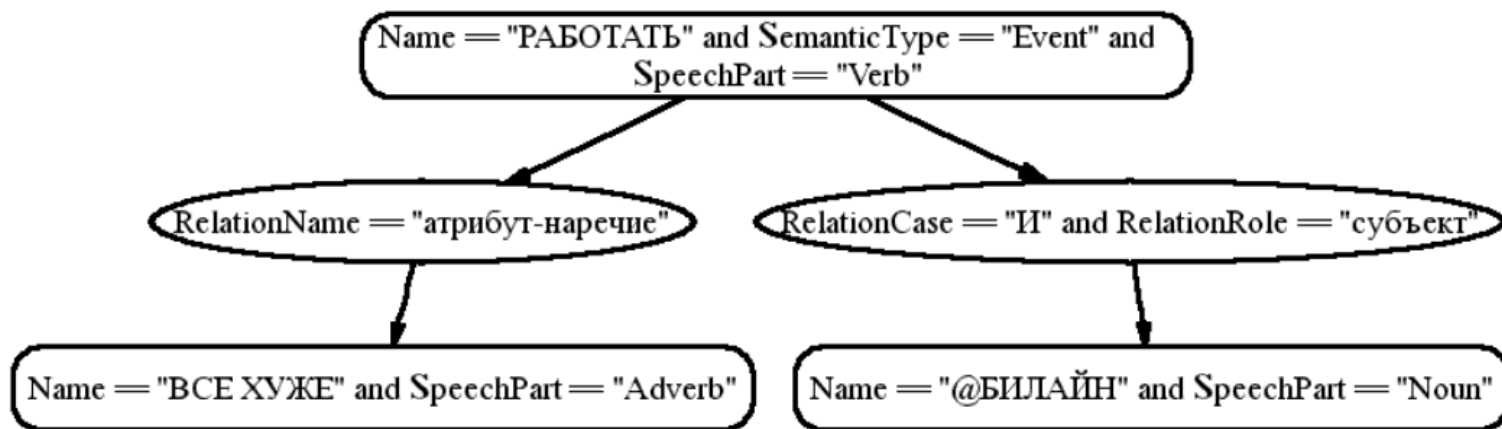
- Текстовая коллекция – сообщения Twitter
- Объекты оценки – операторы сотовой связи
- Задача: определить тональность упоминания операторов сотовой связи в сообщениях пользователей социальной сети.
- Тип тональности: позитивное, негативное, нейтральное упоминание.

Виды тональности

- Тональность мнения: оценка мнения пользователя об объекте
 - *Люблю Билайн - не теряйте марку!*
 - *RT @shahvv: Билайн работает всё хуже и хуже. Я буду скоро в Мегафоне.*
- Тональность информационного повода: новостные события с участием объекта
 - *«Билайн» первым запустил мобильную связь в 150 поселках Якутии #компьютер #новости*
 - *LTE Билайна заработал на iPhone 5s/5c*
- Доля новостных сообщений в тестовой коллекции: 17%.

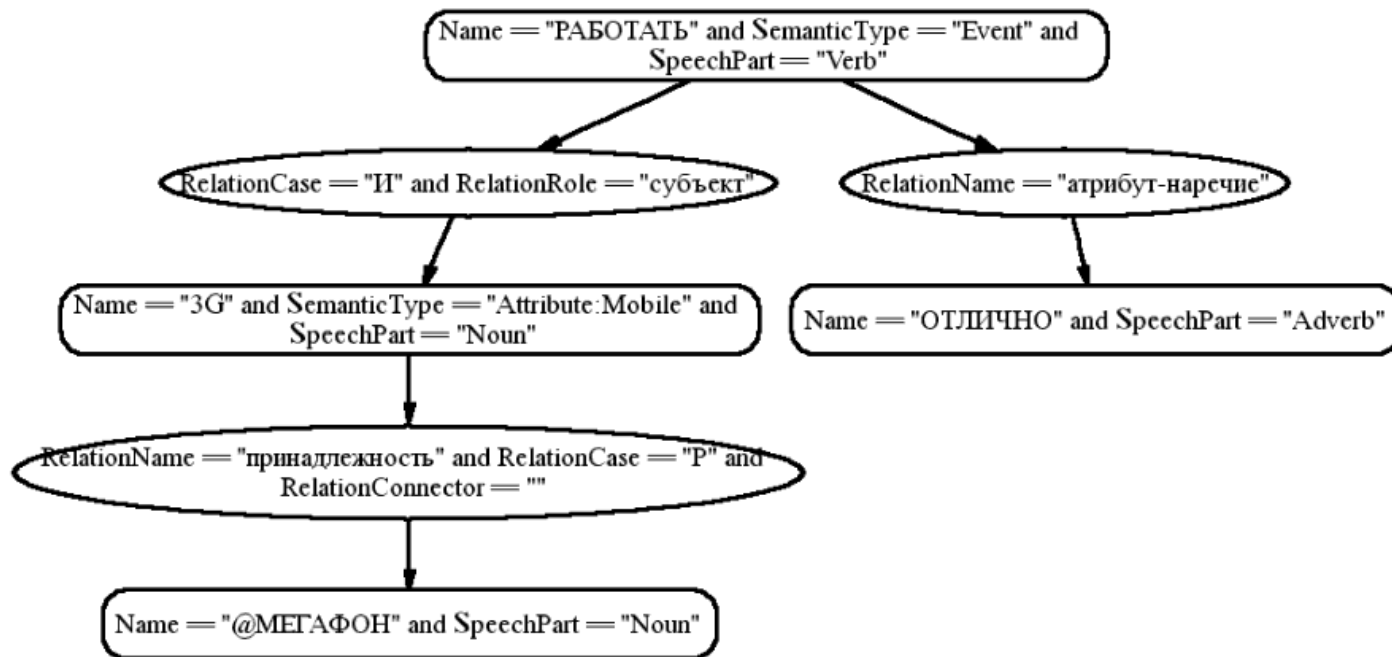
Способы выражения мнения пользователя об объекте

- Объект и его оценка синтаксически связаны:
 - Люблю **Билайн**;
 - **Билайн** работает всё хуже



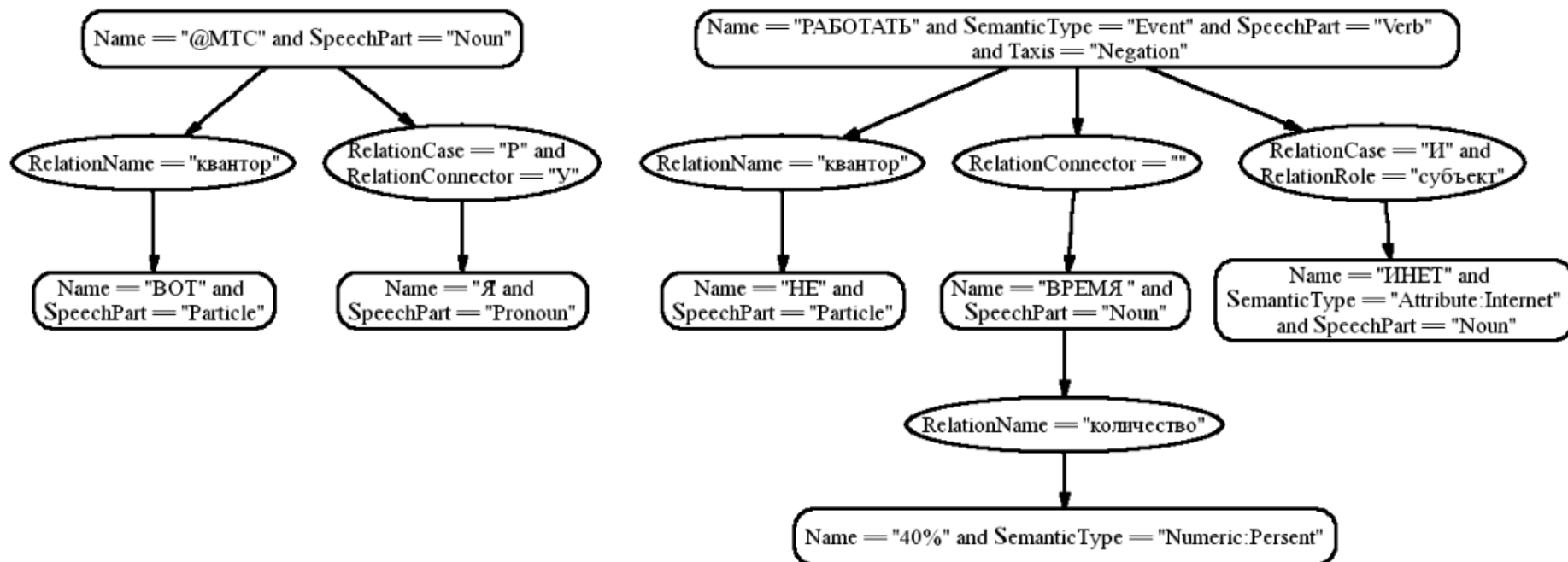
Способы выражения мнения пользователя об объекте

- Оценка синтаксически связана с атрибутом объекта, атрибут объекта синтаксически связан с объектом:
 - *Интернет билайна* слишком часто пропадает;
 - Отлично работает *3G Мегафона*



Способы выражения мнения пользователя об объекте

- Оценка синтаксически связана с атрибутом объекта, объект и атрибут синтаксически не связаны:
 - А пока хочу передать пламенный привет **билайну**, **связи** у меня нет второй день!
 - У меня вот **МТС**, **инет** не работает 40% времени



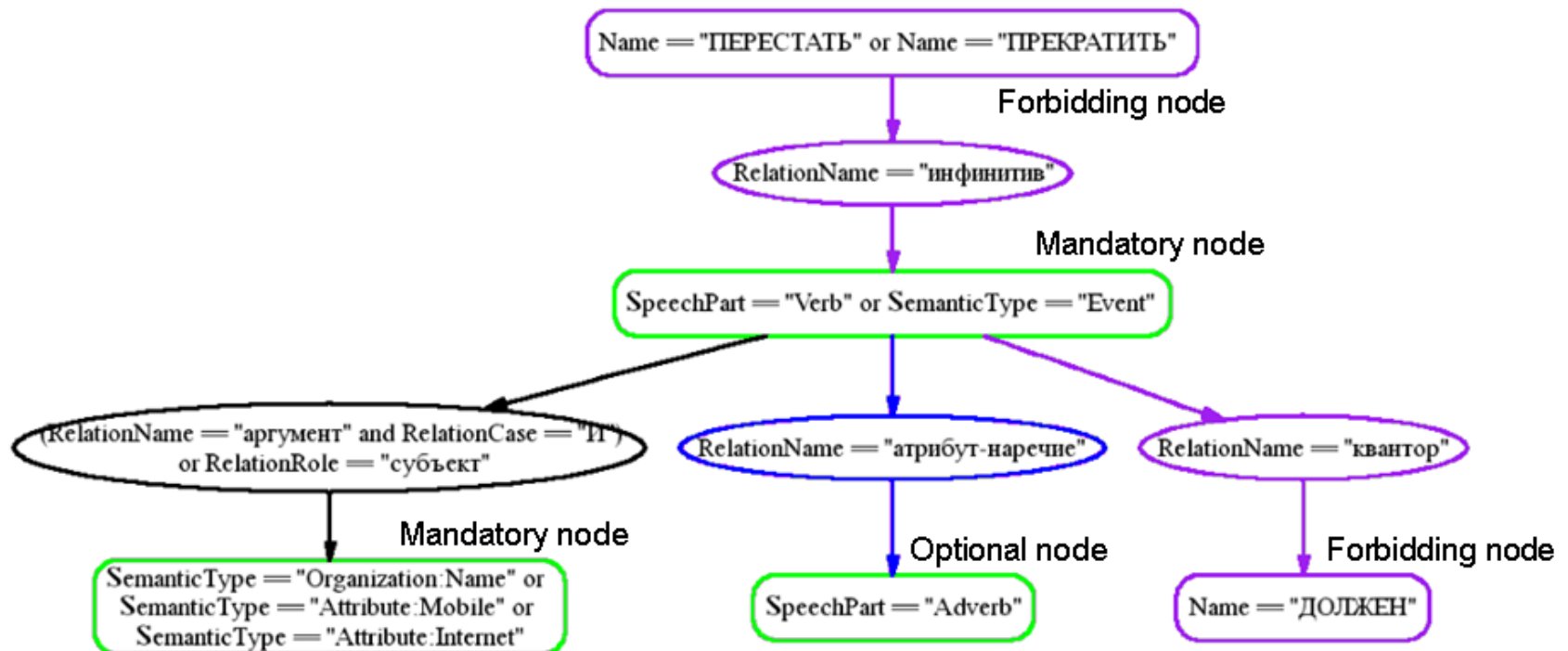
Способы выражения мнения пользователя об объекте

- Свободная тональность: тональная лексика не связана синтаксическими связями ни с объектом, ни с его атрибутами, но задаёт контекст упоминания объекта и его атрибутов:
 - *@centr_mts Если ваша компания не может нормально работать, убирайтесь с рынка к чертовой матери со своей компанией!!!*
 - *Уррряяя, теперь в **Билайне** можно заказать гудок с @yakovlevsinger !!!*
 - *Вот блин... Пользуюсь интернетом от **МТС** уже два года, но такого еще не видел!*
 - *Браво, **Билайн**!*

Алгоритм анализа текста

- Токенизация;
- Морфологический анализ;
- Извлечение объектов интереса;
- Синтаксический анализ;
- Извлечение фактов (применение семантических шаблонов);
- Связывание объектов с их атрибутами
- Поиск свободной тональности, связывание с объектами и атрибутами.

Семантические шаблоны



Предложения, на которых шаблон срабатывает:

+ *Билайн хорошо ловит*

Предложения, на которых шаблон не срабатывает:

- *Билайн перестал хорошо ловить*
- *Билайн должен хорошо ловить*

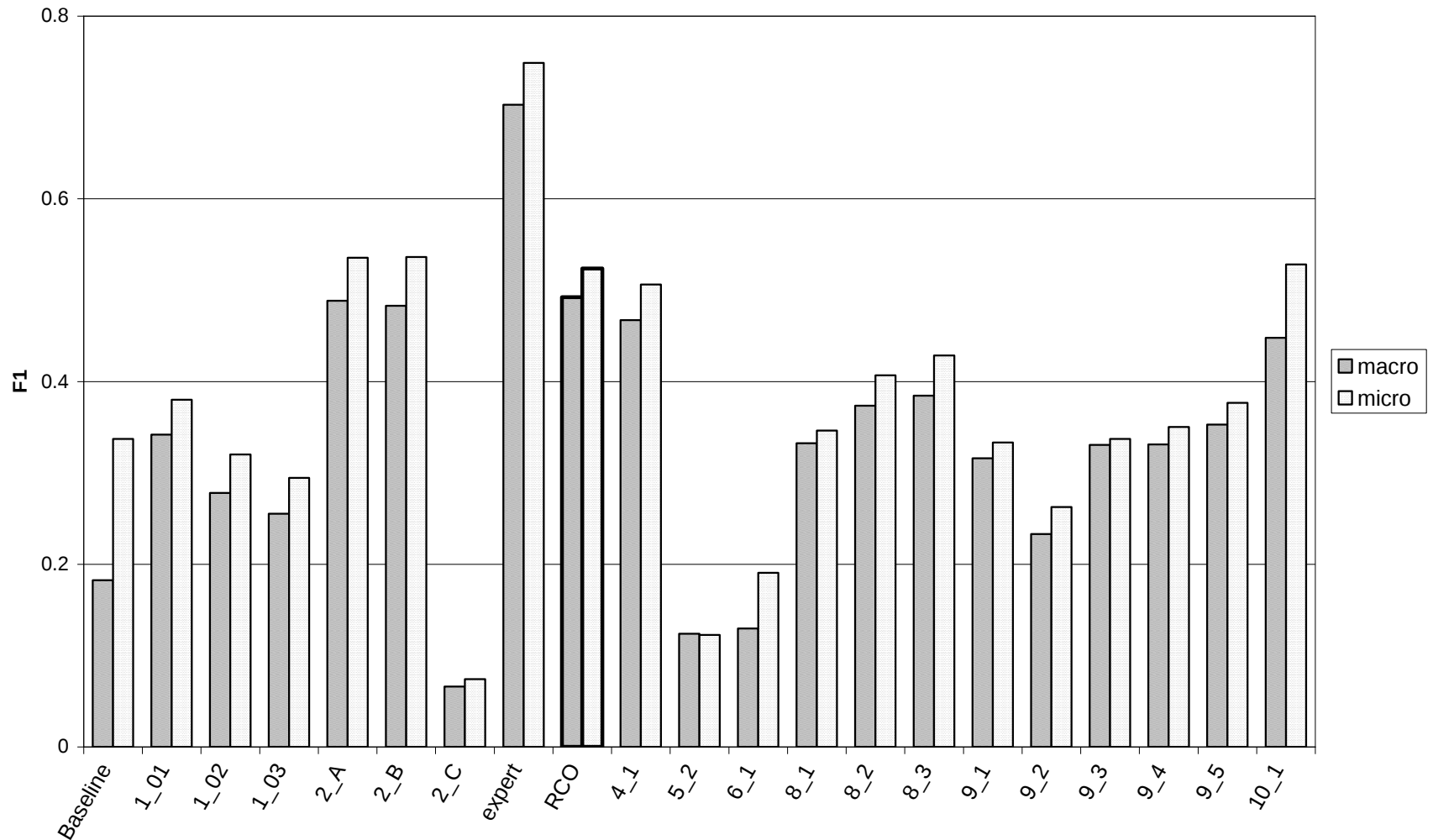
Связывание объектов с их атрибутами

- Правила связывания атрибутов с объектами учитывают:
 - расстояние в словах между объектом и атрибутом;
 - разбиение текста на предложения и клаузы;
 - порядок следования в тексте атрибутов и объектов;
 - тип атрибута (безусловный или может относиться к другому типу объектов)

Свободная тональность

- Используются словари позитивных и негативных терминов:
 - *не надо так!*
 - *что за нах;*
 - *блин;*
 - *ну как так можно*
 - *ура!*
 - *браво;*
 - *крутотень;*
 - *супер!*
- Определяется ближайший объект или атрибут, к которому может относиться данная тональная лексика, при этом учитывается:
 - расстояние в словах и клаузах;
 - порядок следования в тексте;
 - наличие поблизости конкурирующих объектов или атрибутов;
 - прочее.

Результаты



- Лучшие методы показали сравнимый результат
- Уровень согласованности экспертной разметки низкий, но он существенно выше показателей лучших методов

Направление дальнейшей работы

- Учёт веса достоверности определения тональности
- Оценка полноты и точности определения тональности по отдельности для каждого способа выражения мнения пользователя об объекте
- Дополнение лингвистического подхода методами машинного обучения:
 - в части генерации словарей тональной лексики;
 - в части генерации шаблонов;
 - в части генерации правил связывания объектов с атрибутами и отнесения свободной тональности.