Комбинирование фактов, семантических ролей и тональных слов в генеративной модели для поиска мнений

Фельдман Даниил

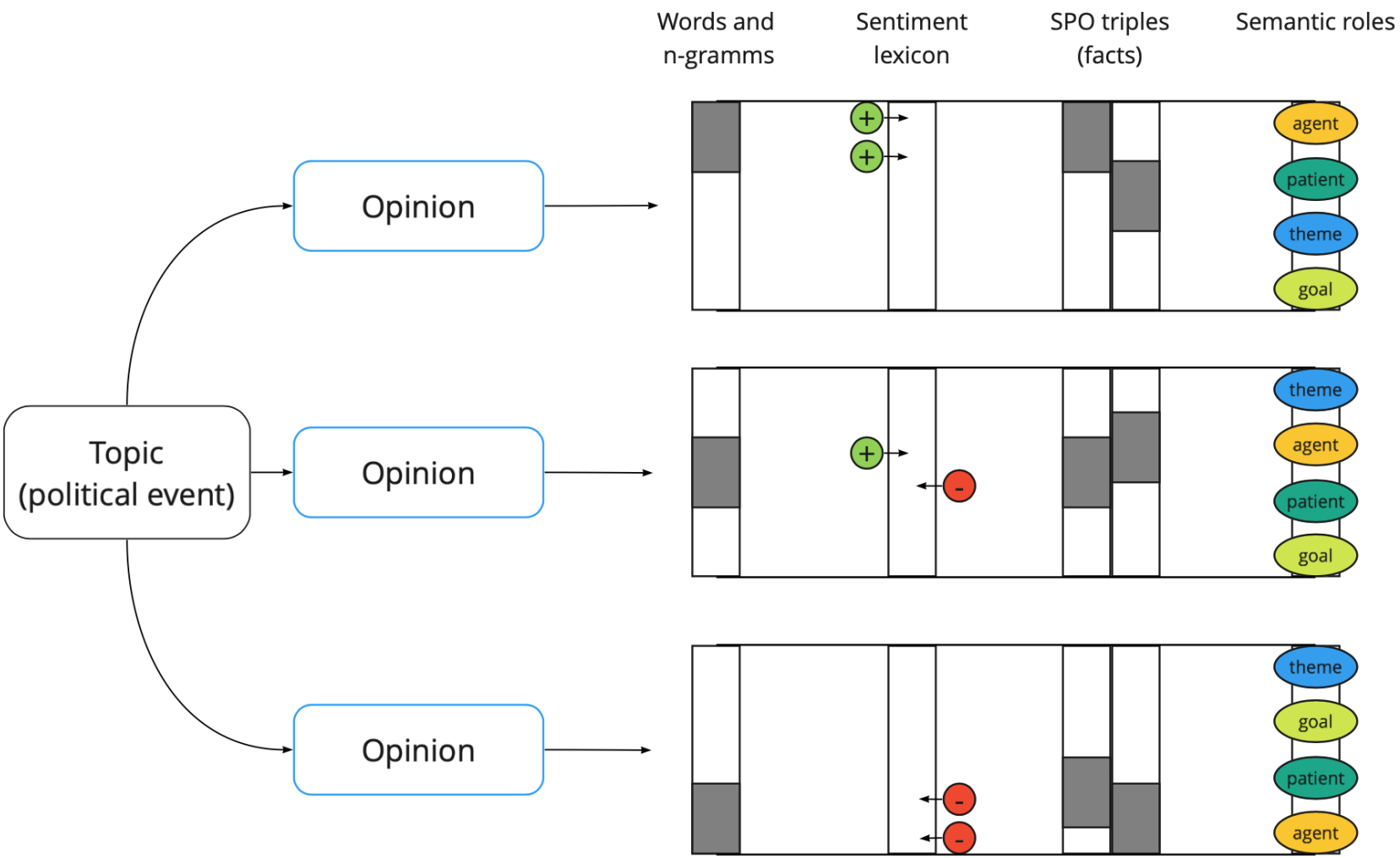
Научный руководитель: д.ф.-м.н. К. В. Воронцов

Московский физико-технический институт
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра интеллектуальных систем

Москва, 2020

Формализация понятия мнения

Мнение может быть определено через распределение слов, тональных слов, фактов или семантических ролей



Признаки для разделения мнений

... Президент Петр Порошенко заявил, что Россия де-факто конфисковала украинские предприятия, которые находятся на неподконтрольной Киеву территории. Сегодня ДНР и ЛНР "национализировали" украинские предприятия ... При этом Кремль защитил конфискацию предприятий в ЛДНР ... Украина потребует расширить санкции ... За все эти действия обязательно наступит наказание. Украина потребует расширения санкций на тех, кто украл украинские предприятия ... ***(Kiev opinion)***

... По словам Захарченко, Киев встретит свой "ужасный конец"... Киев возьмется за ум, и в целях спасения собственной промышленности снимет блокаду ... Обстановка, которую искусственно создала Украина с блокадой Донбасса, вынудила ... кошмарит свой народ ... если в Киеве были приняты какое-либо постановление ... положительные результаты, как в республиках, так и в России ... Если им удастся сместить Порошенко и при этом не развалить Украину, то все вернется на свои места ... ***(Moscow opinion)***

Subject | Object | Agent | Locative | Negative lexicon | Dependent word

Две новости с противоположными мнениями

- Слова «Порошенко», «Россия», «Украина» употреблены одинаково часто
- «Порошенко» - субъект в первом, объект во втором
- «Россия» - agent в первом, locative во втором
- Отрицательно окрашенные слова относятся к словам «Россия», «Кремль» в первом, к словам «Киев», «Украина» во втором

Постановка задачи

Предположения

- Корпус D состоит из новостей на заданную тему или политическое событие
- Общее количество мнений $|O|$ известно

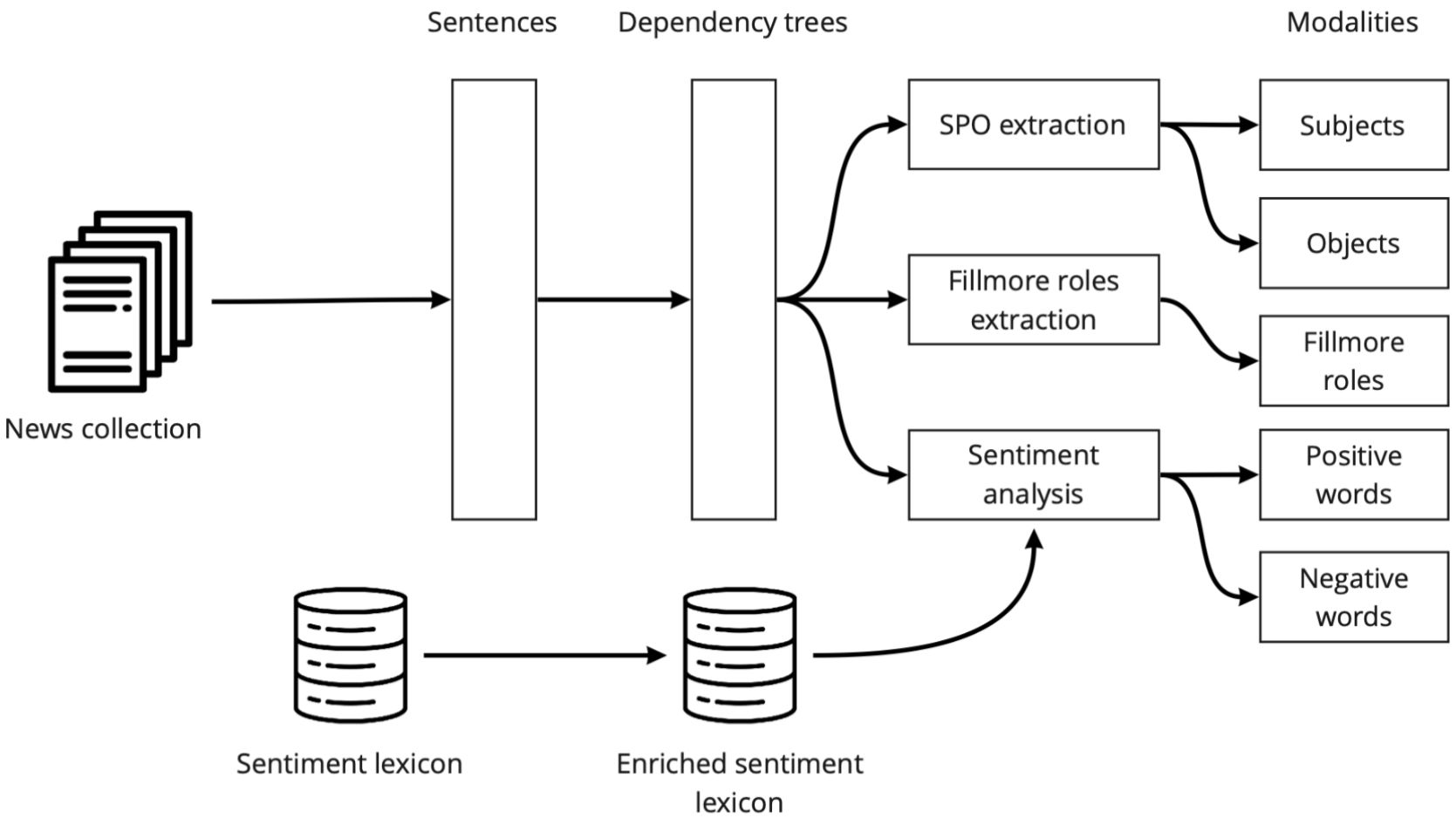
Данные

Количество документов в D порядка 10^2.
Каждый текст, как правило, выражает одно основное мнение.

Кластеризация

- Unsupervised - данные используется только для оценки
- Soft - вероятности отнесения к кластерам
- Использует факты, семантические роли и тональные слова

План решения

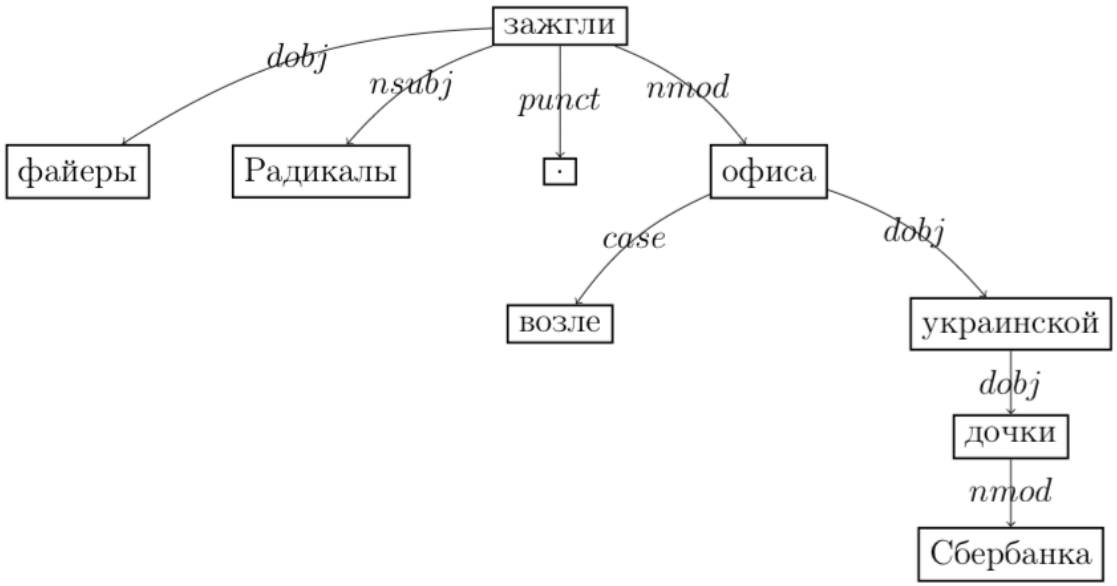


Построение деревьев зависимости

Синтаксический анализатор Google SyntaxNet

SyntaxNet — предобученная нейросеть поверх TensorFlow
Выход, для каждого слова в каждом предложении:

- id слова и id родительского слова
- часть речи: `NOUN`, `VERB`, `ADJ`, `ADV`, ...
- отношения с другими словами: `nsubj`, `dobj`, `conj`, `cc`, `nmod`, ...

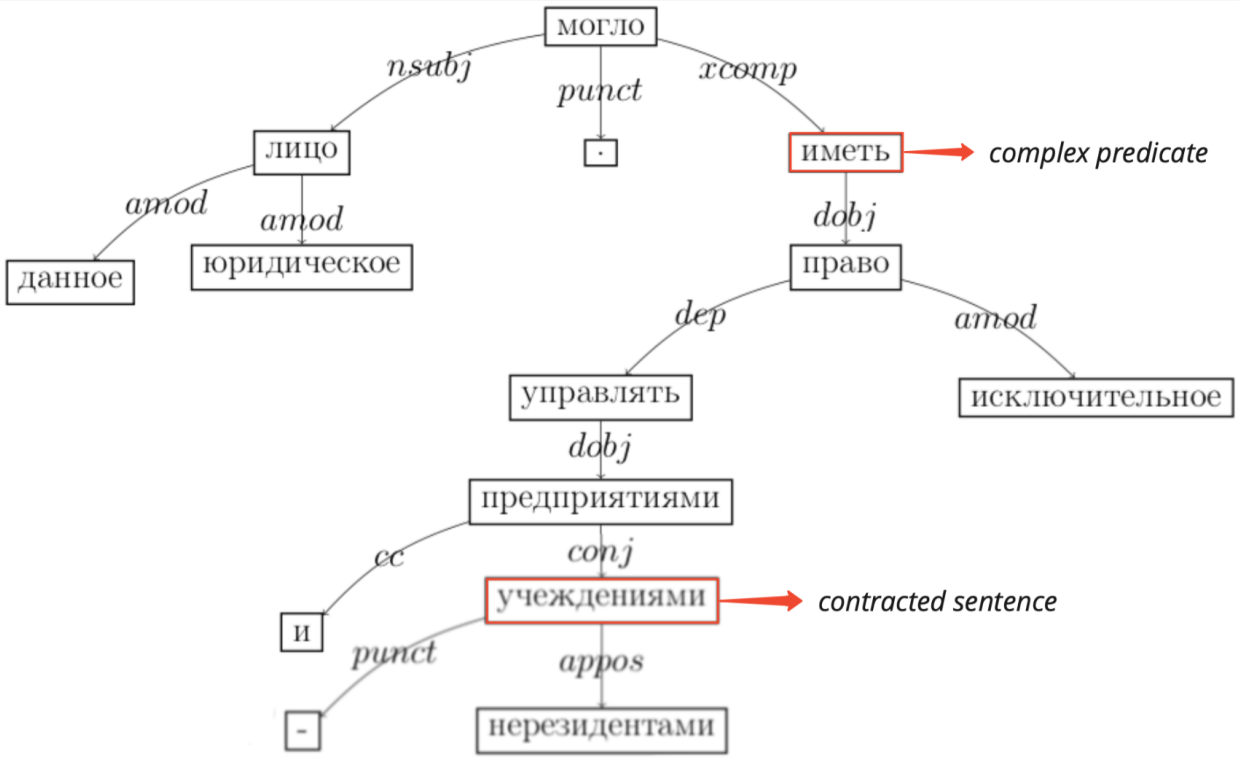


[0] *D.Andor, C.Alberti, D.dWeiss, A.Severyn, A.Presta, K.Ganchev, S.Petrov, M.Collins.* Globally Normalized Transition-Based Neural Networks. 2016.

Выделение фактов

Типы SPO триплетов

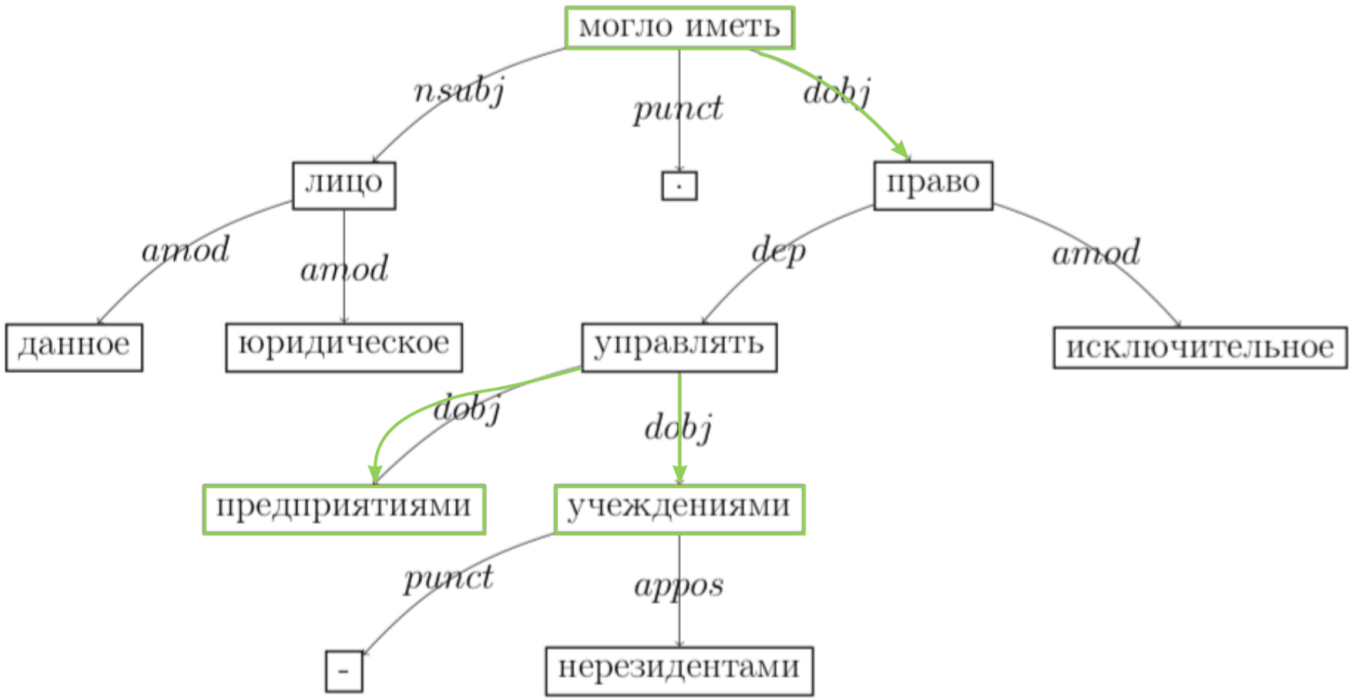
- Субъект-глагол-объект
- Субъект-причастие-объект
- Субъект-"есть"-объект
- Субъект-"есть"-прилагательное



Перестроение деревьев зависимости

Поиск триплетов

1. Обработка однородных членов, имен и местоимений
2. Обработка сложных глаголов
3. Поиск основы триплета
4. Дополнение основ до триплетов



Выделение семантических роли

Роли по Филлмору

9 ролей: Agent, Patient, Theme, Experiencer, Goal, Benefactive, Source, Instrument, Locative (cross-domain)

Модель выделения ролей

1. FrameBank - база размеченных семантических фреймов на русском языке
2. Построение деревьев зависимостей и выделение из их структуры признаков
3. Block-FCNN - извлечение frame-specific ролей (разреженные распределения в новостях)
4. Кластеризация ролей (плотные распределения ролей в новостных текстах)

A. O. Shelmanov, D. A. Devyatkin *«Semantic role labeling with neural networks for texts in Russian»*. In: *Proc. of the International Conference «Dialogue 2017»*

Выделение тональных слов

Словарь тональных слов

Словарь: Linis Crowd - получен из текстов на политические и социальные темы
Обогащение: RuWordNet - синонимы, гипонимы и антонимы

2454 слов $\rightarrow$ 3419 слов

Алгоритм выделения тональных слов

1. Отмечаем слова из словаря
2. С помощью дерева зависимостей отмечаем связанные слова и изменяем тональность

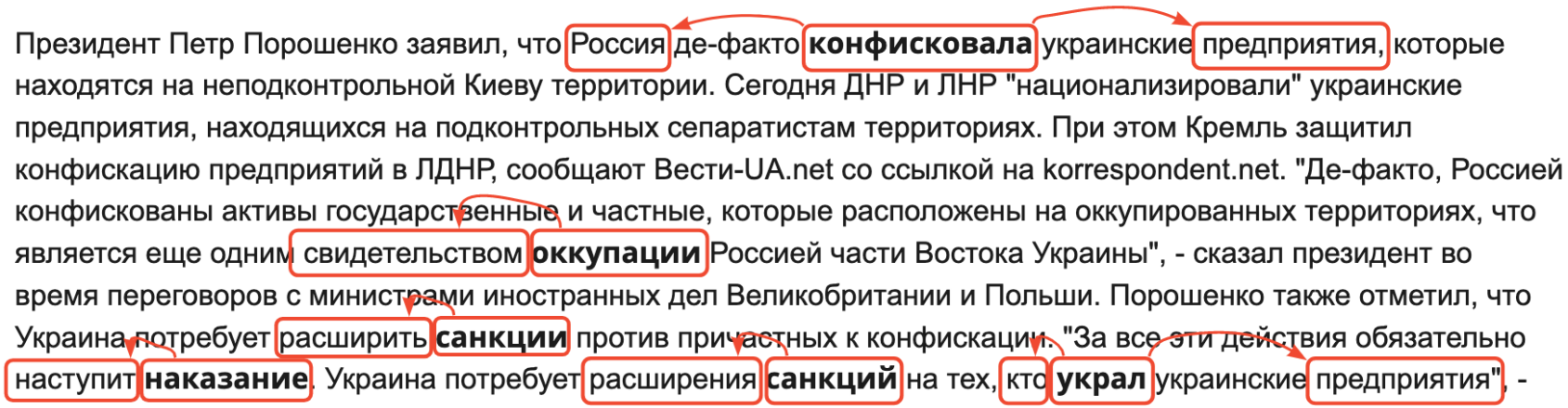
O. Koltsova et al., *"An opinion word lexicon and a training dataset for russian sentiment analysis of social media."*. In: *Proc. of the International Conference "Dialogue 2016"*

Lashevich G. et al. *"Creating Russian WordNet by Conversion."*. In: *Proc. of the International Conference "Dialogue 2016"*

Присвоение тональностей

Тегирование тональных слов и связанных

- Если отмечено существительное, прилагательное или наречение, его родители отмечаются той же тональностью
- Если отмечен глагол, его субъекты и объекты отмечаются той же тональностью
- Если слово связано с отрицательной частицей, его тональность меняется



Мультимодальная тематическая модель

Представление текста

Последовательность троек $(w_i, o_i, d_i)_{i=1}^n \in W \times O \times D$

$$p(w|d) = \sum_o p(w|o)p(o|d) = \sum_o \varphi_{wo}\theta_{od}$$

предполагаем независимость от документа: $p(w|o,d) = p(w|o)$

Матрицы $\Phi = \{\varphi_{wo}\}$, $\Theta = \{\theta_{od}\}$ - параметры модели

Функция правдоподобия

$$L(\Phi,\Theta) = \sum_m \tau_m \sum_d \sum_w n_{dw} \sum_o \varphi_{wo}\theta_{od} + R(\Phi,\Theta) \to \max_{\Phi,\Theta}$$

m - модальности, $R(\Phi,\Theta)$ - критерий регуляризации

Vorontsov K. et al. *«BigARTM: Open source library for regularized multimodal topic modeling of large collections»*. In: *International Conference on Analysis of Images, Social Networks and Texts.* (2015)

Определение модальностей

Факты

- n_{dw} - число раз, когда слово w встречается в триплете SPO в качестве субъекта
- n_{dw} - число раз, когда слово w встречается в триплете SPO в качестве объекта

Роли по Филлмору

- n_{dw} - число раз, когда слово w фигурировало в роли, то есть встретилась пара слово-роль

Тональные слова

- n_{dw} - число раз, когда слово w было положительно окрашено
- n_{dw} - число раз, когда слово w было отрицательно окрашено

Добавление регуляризаторов

Разреживающий регуляризатор

$$R(\Phi, \Theta) = -\beta_0 \sum_{o \in O} \sum_{w \in W} \beta_w \ln \varphi_{wo} - \alpha_0 \sum_{d \in D} \sum_{o \in O} \alpha_o \ln \theta_{od}$$

Делает распределения предметных мнений φ_o и θ_d более разреженными

Сглаживающий регуляризатор

$$R(\Phi, \Theta) = \beta_0 \sum_{o \in O} \sum_{w \in W} \beta_w \ln \varphi_{wo} + \alpha_0 \sum_{d \in D} \sum_{o \in O} \alpha_o \ln \theta_{od}$$

Делает распределения фоновых мнений φ_o и θ_d более сглаженными

Регуляризатор декоррелирования

$$R(\Phi, \Theta) = -\gamma \sum_{o \in O} \sum_{o' \in O \setminus o} \sum_{w \in W} \varphi_{wo} \varphi_{wo'}$$

Делает предметные мнения более различными

Описание эксперимента

Цель

Какая комбинация синтаксических и семантических методов позволяет формализовать понятие мнения?

Размеченные корпусы новостей

1. **LPR Business**: 100 новостей про национализацию бизнеса в ЛНР и ДНР. 2 мнения - России и Украины
2. **Trump Paris**: 220 новостей про решение Трампа выйти из Парижского соглашения. 2 мнения - Трампа и его противников

Метрики качества

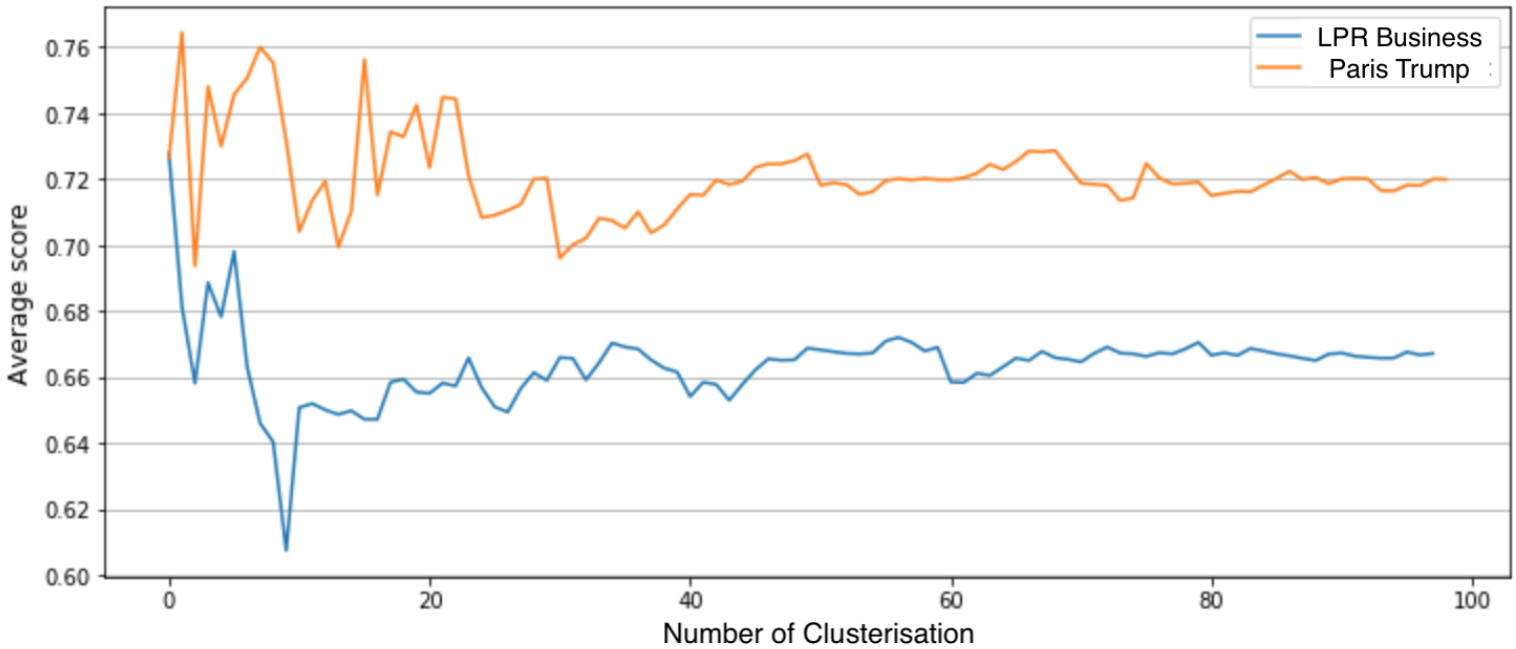
- Precision
- Recall
- F1-score

---

https://github.com/newfteddy/opinion_mining_features/tree/master

Лексический baseline

1. Векторизация: tf-idf
2. Кластеризация: K-menas
3. Мультизапуск по 100 кластеризациям



	ЛНР и ДНР	Парижское соглашение
Avg F1-score	0.67	0.72

Оптимизация гиперпараметров

Параметры: минимальная term frequency для словарей, коэффициенты регуляризации и веса модальностей

1. Оптимизируем term frequency для каждой модальности
2. Оптимизируем коэффициенты регуляризации для каждой модальности
3. Оптимизируем веса модальностей поиском по сетке

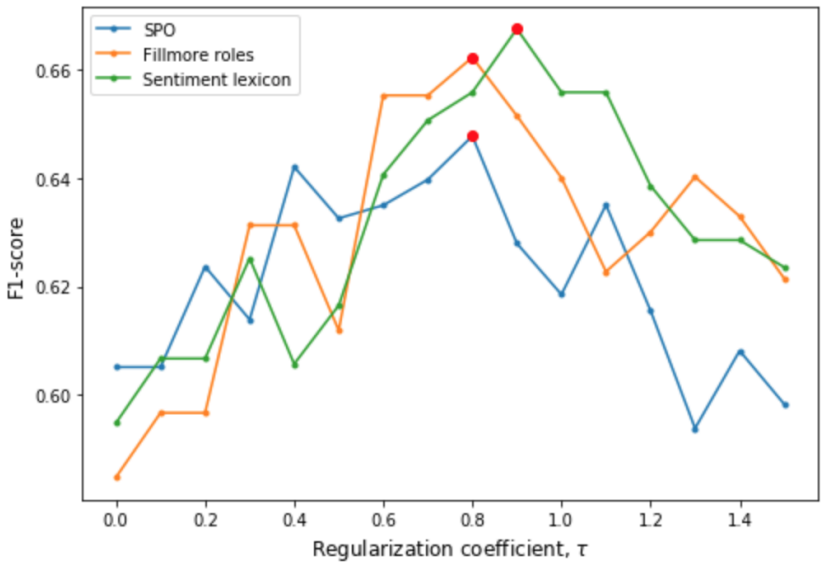


Figure: LPR Business

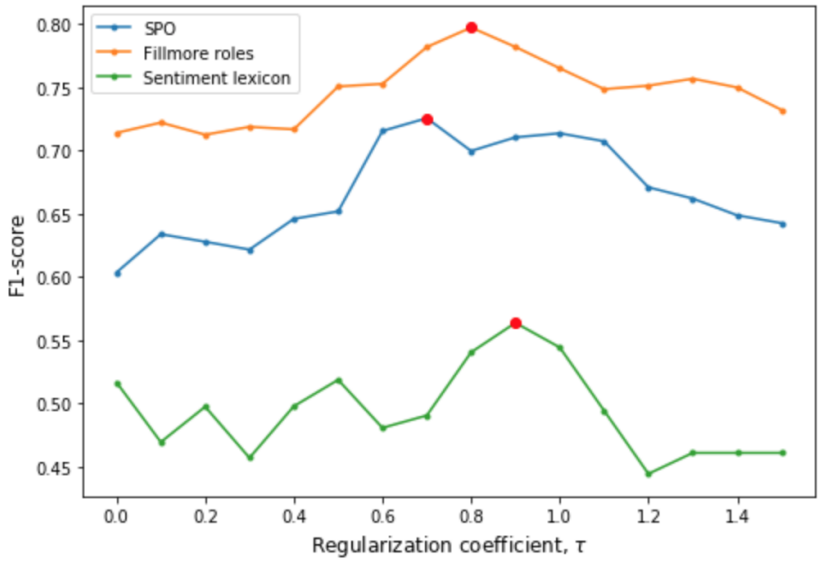


Figure: Trump Paris

Результаты экспериментов

Modalities	*Pr*	*Rec*	*F*1
TF-IDF	0.51	0.95	0.67
SPO	0.59	0.7	0.64
FR	0.86	0.49	0.65
Sent	0.69	0.57	0.66
SPO+FR	0.86	0.68	0.76
SPO+Sent	0.83	0.78	0.81
FR+Sent	0.9	0.52	0.67
All	**0.77**	**0.97**	**0.86**

Table: LPR Business

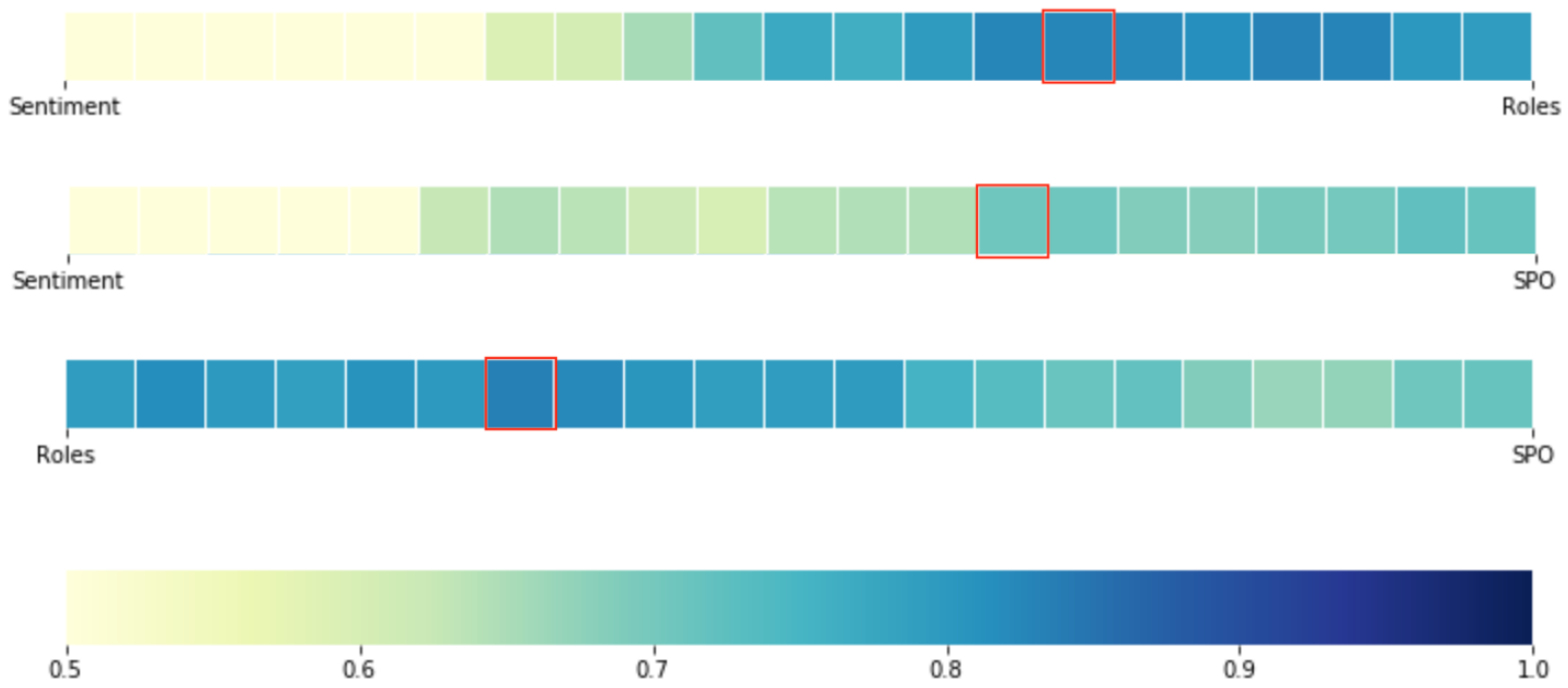
Modalities	*Pr*	*Rec*	*F*1
TF-IDF	0.57	0.97	0.72
SPO	0.56	0.99	0.72
FR	0.67	0.97	0.79
Sent	0.56	0.55	0.55
SPO+FR	0.72	0.99	0.83
SPO+Sent	0.57	0.99	0.72
FR+Sent	0.73	0.97	0.83
All	**0.77**	**0.94**	**0.85**

Table: Trump Paris

Вывод

На обоих датасетах композитная модель показала значительный прирост качества.

Устойчивость результата



Вывод

Результаты устойчивы относительно весов модальностей: по каждой паре признаков есть выраженный максимум, который получается при комбинации с некоторыми весами.

Результаты, выносимые на защиту

1. Формализована задача поиска мнений в новостном потоке
2. Разработаны алгоритмы поиска фактов, семантических ролей и тональных слов в русскоязычных текстах
3. Построена вероятностная модель на предложенных признаках
4. Экспериментально доказано, что мнение может быть формализовано, как комбинация фактов, ролей по Филлмору и тональных слов

Материалы

GitHub с размеченными датасетами: Github opinion_mining_features

Публикации

1. Feldman D.G., Sadekova T.R., Vorontsov K.V.
   "Combining Facts, Semantic Roles and Sentiment Lexicon in a Generative Model for Opinion Mining".
   In: *Proceedings of the International Conference Dialogue 2020*