Министерство образования и науки Российской Федерации
«Московский физико-технический институт
(государственный университет)»
Физтех-школа прикладной математики и информатики
Факультет управления и прикладной математики
Кафедра «Интеллектуальные системы»

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

Выпускная квалификационная работа
(магистерская диссертация)

03.04.01 —- Прикладные математика и физика

Выполнил:
студент М05-874а группы _____ Фельдман Д.Г.
(подпись обучающегося)

Научный руководитель:
д.ф.-м.н., профессор _____ Воронцов К. В.
(подпись научного руководителя)

Москва
2020

Оглавление

Аннотация

Поиск мнений является популярной задачей, которая широко применяется для определения поляризации новостей, идентификации пропаганды, а также классификации отзывов о продуктах. Мы будем решать задачу кластеризации без учителя на текстах новостей о политических событиях, которые выражают некоторое мнение. Целью работы является определение того, какие языковые сущности и их комбинации определяют мнение: лексические, синтаксические и семантические. Более точно, мы проверяем гипотезу о том, что мнение можно формализовать, как комбинацию распределений фактов (SPO триплетов), семантических ролей и тональных слов. В данной работе мы поставим задачу поиска мнений и покажем, что использование композиции упомянутых признаков дает наилучшее качество при кластеризации текстов с мнениями. Для проверки гипотезы мы собрали и разметили два корпуса новостей о политических событий и предложили алгоритмы извлечения триплетов, ролей по Филлмору и тональных слов.

Ключевые слова: SPO триплеты, роли по Филлмору, тональные слова, opinion mining, ARTM.

1. Введение

Поляризация новостей. Каждое заметное политическое событие описывается в большом количестве новостных изданий. Про этом автор чаще всего не просто описывает произошедшие события, а выражает позицию одной из сторон. В результате этого, большая часть контента в новостном потоке поляризована, и читатель конкретной новости получает представление только об одной стороне проблемы. Мы хотим найти способ кластеризовать тексты в новостном потоке, которые покрывают некоторое событие, по мнениям. Такая задача и называется opinion mining. В более общем виде она предполагает одновременное нахождения числа мнений. Мы будем решать несколько упрощенную задачу, считая, что нам дано общее число мнений в корпусе.

Подходы к opinion mining. В данной работе мы исследуем различные методы кластеризации текстов с мнениями о некоторых политических событий. Основной вопрос, который мы ставим: какие вычислимые сущности в тексте определяют его мнение. Наиболее популярным является подход, когда считают, что мнение определяется лексическими признаками, то есть распределением слов. Мы будем рассматривать более сложную генеративную модель текста, которая использует семантические и синтаксические признаки. Целью работы будет понять, какие из них наилучшим образом позволяют формализовать понятие мнения.

Мы будем рассматривать три языковых сущности: факты (триплеты субъект-предикат-объект), семантические роли и тонально окрашенные слова. Эти признаки позволяют выделять различия в даже в тех текстах, которые имеют схожее распределение слов. Чтобы показать это, рассмотрим два примера реальных новостей о национализации предприятий в ЛНР и ДНР.

... 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 ...

... 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 для республик, так и для России ... Если им удастся сместить Порошенко и при этом не развалить Украину, то все вернется на свои места ...

В этих примерах достаточно схожие распределения слов, хотя первый выражает мнение Украины, а второе - мнение России. Слова *Порошенко*, *Россия*, *Украина* используются в обоих текстах с одинаковой частотой. При этом *Порошенко* используется в первом тексте, как субъект, и во втором, как объект. Слово *Россия* принимает разные семантические роли: агенс в первом тексте и локатив во втором. Если мы обратим внимание на тональные слова, то отрицательно окрашенные слова *оккупация*, *украсть*, *заложник*, *угроза* связаны с Россией в первом тексте и с Украиной во втором. Таким образом, в данном примере лексически, то есть на уровне распределения слов новости, выражающие различные мнения, неразличимы. Однако они существенно отличаются распределениями триплетов SPO, семантических ролей и тональных слов.

Выделение новостей из общего потока. Тематическое моделирование подновляет нам выделять новости, покрывающие, к примеру, политические события. Следующая задача - выделение мнений относительно данного события. При поиске мнений мы, по сути, строим второй уровень двухуровневой иерархической тематической модели. Вместо мнений могут выделиться сто-

роны или агенты. Мы предполагаем, что подход с использованием синтаксических и семантических признаков позволяет выделить именно мнения. Для задачи opinion mining мы будем рассматривать генеративную модель текста, аналагичную тематическому моделированию (ARTM).

Факты. Под фактами мы будем иметь в виду триплеты субъект-предикат-объект (SPO). При их исследовании текст представляют не в виде мешка слов (bag-of-words), а в виде мешка триплетов. Такой подход позволяет получить улучшение качества во многих задачах обработки текста, в том числе построении онтологий и определении поляризации. Использование их для поиска мнений основано на том предположении, что автор выражает мнение фактами, которые он называет и умалчивает, а также тем, как он рассказывает о действиях лиц, то есть структурой взаимодействия субъектов и объектов.

Семантические роли. Семантические роли позволяют выразить значение слов и показывают мысль, которую хотел выразить автор. Это способ показать множество отношений между словами ограниченным набором состояний. Существует множество различных наборов семантических ролей разного размера. На одном конце спектра роли, которые относятся к конкретной предметной области (domain-specific), такие как From_Airport, To_Airport, Depart_Date. Они относятся к конкретному семантическому фрейму (Полет), и задача их поиска близка к named entity recognition. На другом конце макро роли - прото-агенс и прото-пациенс (на практике часто совпадает с субъектом и объектом). Они общие для всех текстов, однако не показывают сложные зависимости в предложении. Посередине спектра множество наборов ролей, в том числе девять ролей по Филлмору ([13]): Agent, Experiencer, Instrument, Goal, Location, Object, Source, Time, Path. Мы остановили свой выбор на этом наборе, так как он общий для всех предметных областей, но при этом показывает сложные зависимости между словами.

Тонально окрашенные слова. Это достаточно популярный метод для поиска мнений. Он основан на том, что некоторым словам в тексте можно приписать некоторый сентимент, и распределения таких слов показывает мнение. В общем случае слова могут выражать большой набор тональностей, и для его определения необходимо учитывать не только само слово, но и контекст. Мы будем рассматривать только для противоположных сентимента: +1 и -1,

то есть положительный и отрицательный.

Цель работы. Формализовать понятие мнения и показать, что оно может быть определено, как комбинация распределений фактов, семантических ролей и тонально окрашенных слов.

1.1. Обзор литературы

Задача поиска мнений достаточно популярна, и она была широко изучена за последние годы. В работах [2] и [3] показан обзор существующих методов и проведено их сравнение. Более ранние работы ([1], [4]) концентрировались на поиске мнений в отзывах о товарах, но в более поздних исследованиях фокус сместился в сторону текстов о политических событиях и проблеме политической поляризации новостей ([5], [6], [12]). Большая часть работ опирается на вероятностные модели ([2]), а тематические модели, такие как LDA, использовались и в обучении с учителем [5], и в обучении без учителя [6]. В русскоязычных корпусах задача поиска мнений не была популярной, и большая часть описанных методов не была применена.

При анализе потока новостей supervised модели применять затруднительно, так как корпусы документов обычно недостаточного размера: речь идет о нескольких сотнях новостей о политическом событии. На практике собрать и разметить такой датасет за время, пока новость остается актуальной не вполне реализуемо, поэтому нас интересует обучение без учителя. В целом, генеративные модели наиболее предпочтительный и популярный метод в поиске мнений ([5], [4], [14]). Некоторые работы решают более общую задачу нахождения тем и мнений одновременно. Так, авторы работы [7] предложили модель поиска тем и аспектов, где последние могут быть интерпретированы как мнения. Мы же будем фокусировать на более узкой задачи поиска мнений в текстах о заданном политическом событии, так как наша цель - формализация понятия мнения.

Триплеты субъект-придикат-объект использовались для поиска мнений, а также они помогли улучшить качество в других лингвистических задачах. Например, в работе [8] их использовали в схожей генеративной модели для построения онтологий, используя иерархическую тематическую модель. Эта работа и дала предположение, что факты позволят улучшить качество в нашей

задаче. Тем не менее, мы не нашли алгоритмов по поиску триплетов в русскоязычном тексте. Тонально окрашенные слова широко использовали для поиска мнений. Многие исследования использовали поляризованные слова, чтобы классифицировать отзывы о товарах и политические тексты. Самый базовый подход для такой модели - использовать словари тональных слов, однако он не берет во внимание контекст, поэтому оказывается неэффективным. Необходимым улучшением здесь является метод обогащения словаря тональных слов с использованием правил ([16]), которые основаны на части речи, роли слова и соседних словах. Такие работы, как [17] предлагают использовать систему, основанную на большом количестве эвристических правил, и наша модель для извлечения тональных слов использует схожий подход. Семантические роли также использовались в задаче opinion mining, например, в [15] они позволили существенно улучшить качество. Существует несколько подходов к поиску ролей, но большая часть использует некоторую архитектуру нейронных сетей ([19], [20], [21]) и базу семантических фреймов, размеченную вручную, такую как FrameNet или VerbNet ([20], [21]). Модель для извлечения семантических ролей на русском языке была предложена Шелмановым и Девяткиным в [18]. Она также использует нейросеть в качестве модели и базу фреймов на русском FrameBank.

Новизна исследования. Факты и семантические роли были использованы для в задаче поиска мнений, но только в моделях обучения с учителем. Тонально окрашенные слова использовались для поиска мнений при обучении без учителя, но ни один из трех методов не рассматривали на русскоязычных текстах. Главным образом, в существующих исследованиях для поиска мнений не комбинировали факты, семантические роли и тональные слова. В существующих работах модели с обучением без учителя показывали достаточно низкое качество и не позволяли формализовать понятие мнения.

2. Постановка задачи

Рассмотрим корпус документов D, состоящий из текстов новостей. Размер такого корпуса порядка 10^2. Каждый документ $d \in D$ мы будем рассматривать как множество последовательностей элементов различных типов, или

модальностей. В общем случае это могут быть слова, n-граммы, картинки, ссылки и прочие элементы. Множество модальностей обозначается M. В данной работе оно состоит из:

1. Субъекты из словаря W^s
2. Объекты из словаря W^o
3. Пары (слово, роль по Филлмору) из словаря W^r
4. Положительно окрашенные слова из словаря W^p
5. Негативно окрашенные слова из словаря W^n

$M = \{subjects, objects, pairs, positve, negative\}$. Можно заметить, что вместо триплетов SPO мы ввели две модальности субъектов и объектов. Это связано с тем, что распределения триплетов в тексте очень разрежены: большая часть из них встречается только один раз. С другой стороны, для всех девяти семантических ролей мы ввели одну модальность пар вместо девяти модальностей для каждой роли. Это связано с тем, что некоторые роли встречаются в тексте достаточно редко, и их распределения также были бы слишком разреженными. Общий словарь для всех модальностей обозначим $W = \bigcup_{m \in M} W^m$. Тогда документ d можно представить как последовательность термов $(w_1, ..., w_{n_d}; d) \in W$. Под термом мы понимаем элемент из W, относящийся к одной из модальностей.

Конечное множество мнений в D обозначим за O. Будет считать, что каждый терм, встречаемый в документе d связан с некоторым известным мнением $o \in O$. Таким образом, текст можно рассмотреть как последовательность троек $(w_i, o_i, d_i)_{i=1}^n \in W \times O \times D$. В такой модели текста мнения o_i являются скрытыми переменными, а термы w_i и тексты d_i - явными.

2.1. Вероятностная модель

Будем считать в этом пункте, что мы работаем с одной модальностью со словарем W. Тогда можно положить, что каждый терм сэмплируется из распределения $p(w|o, d)$. Для упрощения мы будем преполагать, что терм не зависит от документа d, и на него влияет только мнение, выражаемое

автором, то есть $p(w|o,d) = p(w|o)$. Мнения же, в свою очередь, выбираются автором для каждого документа из распределения $p(o|d)$. Объединив эти утверждения, мы можем получить распределение термов в документе как сумму произведений скрытых распределений:

$$p(w|d) = \sum_{o \in O} p(w|o)p(o|d) \tag{1}$$

Введем обозначения $\varphi_{wo} = p(w|o)$ и $\theta_{od} = p(o|d)$, а также матрицы $\Phi = \{\varphi_{wo}\}_{W \times O}$ и $\Theta = \{\theta_{od}\}_{O \times D}$. Распределения слов по документам также запишем в виде матрицы $F = \{p(w|d)\}_{W \times D}$. Тогда последнее выражение примет вид:

$$p(w|d) = \sum_{o \in O} \varphi_{wo}\theta_{od} \;\text{ или }\; F = \Phi \cdot \Theta \tag{2}$$

Таким образом, задача поиска распределения термов по документов принимает вид матричного разложения, а параметрами модели становятся Φ и Θ. Можем записать функцию правдоподобия, как зависимость вероятности последовательности термов $(w_i, d_i)_{i=1}^n$ от параметров:

$$L((w_i, d_i)_{i=1}^n; \Phi, \Theta) = \prod_{i=1}^{n} p(w_i, d_i) \to \max_{\Phi,\Theta} \tag{3}$$

Введем также обозначение $n_{dw} = \sum\limits_{w' \in d} [w' = w]$ - число термов w в документе. Тогда выражение (3) примет вид:

$$L((w_i, d_i)_{i=1}^n; \Phi, \Theta) = \prod_{d \in D} \prod_{w \in W} p(w|d)^{n_{dw}} p(d)^{n_{dw}} \to \max_{\Phi,\Theta} \tag{4}$$

2.2. Оптимизационная задача

Чтобы получить оптимизационную задачу, возьмем логарифм правдоподобия (4). Можно заметить, что член $p(d)$ является постоянным относительно параметров максимизации, и при взятии логарифма им можно пренебречь. Чтобы получить ограничения, необходимо учесть, что параметры φ_{wo} и θ_{od} должны быть корректными вероятностными распределениями:

$$\begin{aligned} &\max_{\Phi,\Theta} \sum_{d\in D}\sum_{w\in d} n_{dw} \sum_{o\in O} \varphi_{wo}\theta_{od} \\ &\text{s.t.} \sum_{w\in W} \varphi_{wo} = 1,\ \varphi_{wo} \geq 0 \\ &\sum_{o\in O} \theta_{od} = 1,\ \theta_{od} \geq 0 \end{aligned} \tag{5}$$

При построении генеративной модели текста мы будем следовать подходу аддитивной регуляризации тематической модели (ARTM, [9]). Задача матричного разложения (2) не является корректной, потому что она имеет бесконечное множество решений. Модель, которую описывает задача оптимизации (5), называется Probabilistic Latent Semantic Analysis (PLSA). Для того чтобы сделать задачу корректной, мы будем добавлять регуляризаторы к целевой функции, в результате она принимает вид:

$$\sum_{d\in D}\sum_{w\in d} n_{dw} \sum_{o\in O} \varphi_{wo}\theta_{od} + \sum R_i(\Phi,\Theta) \tag{6}$$

Здесь $R(\Phi,\Theta) = \sum R_i(\Phi,\Theta)$ - примененные регуляризаторы.

До этих пор мы рассматривали текст с единственной модальностью. Обобщим вывод для случая множества модальностей M, которое мы ввели в начале раздела. Матрицу явных переменных для модальности m обозначим, как $F^m = \{p(w|d), w \in W^m\}$, а матрицу скрытых переменных $\Phi^m = \{\varphi_{wo}, w \in W^m\}$. Матрица Θ остается без изменений. Наконец, параметры модели F и Φ определим как конкатенацию соответствующих матриц модальностей: $F = \bigcup\limits_{m\in M} F^m$, $\Phi = \bigcup\limits_{m\in M} \Phi^m$. Также, введем веса модальностей $\{\tau_m,\ m \in M\}$, которые отвечают за вклад каждой из них в целевую функцию. Учитывая вышесказанное, из задачи (5) мы можем получить итоговую задачу оптимизации:

$$\begin{aligned} &\max_{\Phi,\Theta} \sum_{m\in M} \tau_m \sum_{d\in D} \sum_{w\in W^m} n_{dw} \sum_{o\in O} \varphi_{wo}\theta_{od} \\ \text{s.t.} \ &\sum_{w\in W^m} \varphi_{wo} = 1, \ \ m \in M \\ &\sum_{o\in O} \theta_{od} = 1, \ \ m \in M \\ &\varphi_{wo} \geq 0, \ \ \theta_{od} \geq 0 \end{aligned} \tag{7}$$

2.3. Кластеризация мнений

Решив задачу (7), мы получим оптимальные праметры Φ и Θ, из которых нас в первую очередь интересует последнее. Каждый столбец матрицы Θ содержит распределение мнений в документе d: $\theta_d = \{\theta_{od}, o \in O\}$. Таким образом, мы делаем мягкую кластеризацию текстов, где каждый кластер - это мнение, а модель показывает вероятность отнесения к этому кластеру. Так как в размеченных данных у каждой новости меткой является одно мнение, для оценки модели мы будем приписывать каждому документу наиболее вероятное мнение: $y_d = \max_{o\in O} \theta_{od}$.

Постановка задачи. Пусть дано множество новостей D и общее число мнений в корпусе $|O|$. Требуется построить мягкую кластеризацию документов на $|O|$ кластеров без учителя. Модель должна основываться на фактах, семантических ролях и тональных словах. Нам будет необходимо решить следующие подзадачи:

1. Построить алгоритмы поиска фактов, семантических ролей и тональных слов в русскоязычном тексте.

2. Получить модальности для оптимизационной задачи.

3. Решить оптимизационную задачу (7).

В следующем разделе рассмотрим последовательно решение каждой из подзадач.

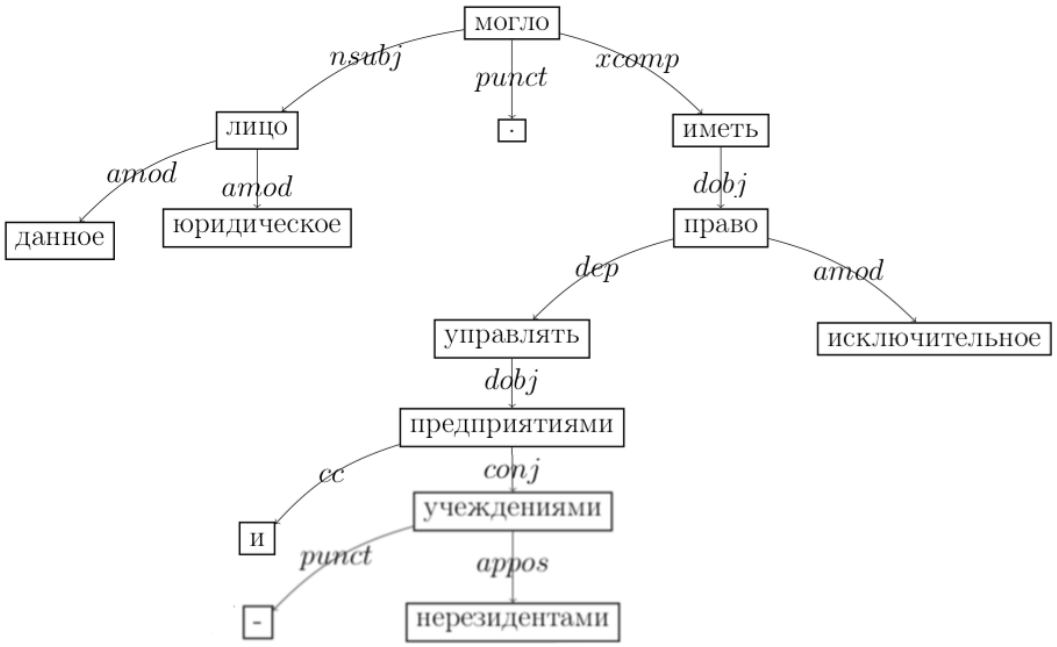

Рис. 1. Пример синтаксического дерева предложения

3. Решение

Перед поиском трех видов признаков в тексте, мы один раз предобрабатываем его. Для этого из каждого документа мы исключаем лишние символы, такие как пунктуацию, удаляем стоп слова и приводим остальные к нормальной форме (лемматизируем).

Для нахождения фактов, ролей и тональных слов мы будем использовать синтаксические деревья, или деревья зависимостей. Это структура, которая показывает связь слов в предложении: между каждой парой мы определяем направление зависимости и ее тип. Пример такого дерева для предложения *Данное юрическое лицо могло иметь исключительное право управлять предприятиями и учеряждениями-нерезидентами* показан на Рис. 1. Для их построения мы будем использовать синтаксический анализатор Google SyntaxNet - предобученную нейросеть, в том числе, на русском языке. Для каждого слова в каждом предложении она позволяем определить множество свойств, таких как часть речи, связи с родительскими словами и типы этих связей.

3.1. Поиск фактов

Мы будем искать несколько типов триплетов SPO:

- *Субъект-глагол-объект*
- *Субъект-причастие-объект*
- *Существительное-есть-существительное*. Например, словосочетание *президент Трамп* является триплетом *Трамп-есть-президент*
- *Существительное-есть-прилагательное*. Формально они не являются фактами, однако в итоге улучшают качество.

Для поиска триплетов первых двух типов мы сначала находим предикат, а затем исследуем связанные с ними слова: нас интересуют типы *nsubj* и *dobj*. Для второго типа мы исследуем связи между существительными типа *appos*. Изначальное дерево зависимостей не позволяет находить все факты корректно, так как структуры однородных членов, сложные сказуемые, местоимения и имена собственные вызывают отдельную трудность. Перед применением алгоритма поиска факторв нам необходимо перестроить дерево, чтобы учесть это. Сложные сказуемые мы объединяем в единый узел, добавляя к нему подходящие связи от дочерних узлов. Для однородных членов мы строим параллельную ветвь к их родителю для каждого члена, местоимения заменяем существительными. На Рис. 2 показан пример дерева до перестроения, а на Рис. 3 - после.

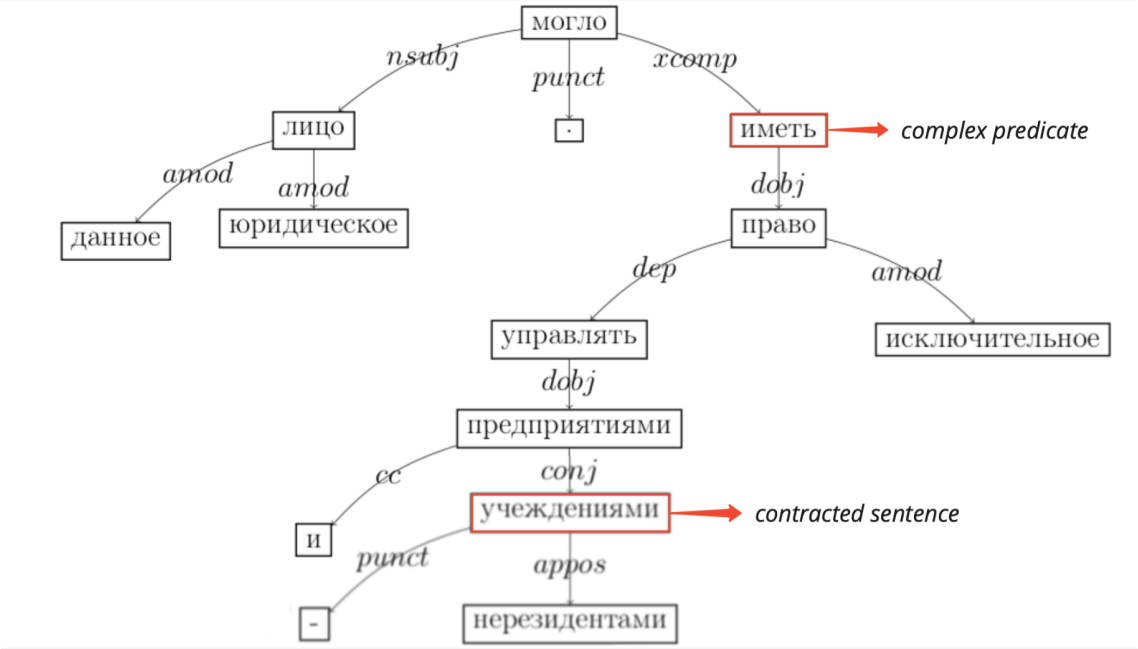

Рис. 2. Дерево зависимостей до перестроения

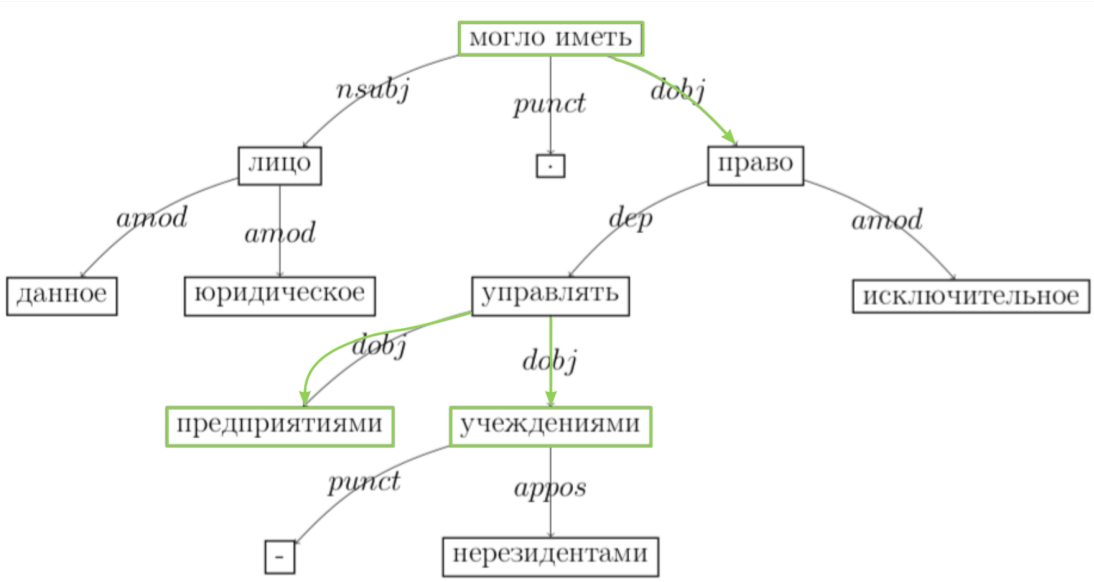

Рис. 3. Дерево зависимостей после перестроения

В данном примере мы можем выделить факты *лицо-может_ иметь-право*, *право-есть-исключительное* и *учереждение-есть-нерезидент*. Вышесказанное можно сформулировать в алгоритме 1, который принимает на вход документ d и выдает список фактов в нем.

3.2. Поиск ролей по Филлмору

Для извлечения ролей по Филлмору мы будем использовать подход, предложенный в 2018 году в [18]. Это блочно-полносвязная нейронная сеть с двумя слоями, которая использует признаки, полученные из дерева зависимостей - путь до слова, позиция, часть речи, типы связей. Для эмбедингов слов мы используем Rusvectors 2.0, которые также подаются в модель вместе с признаками. Сеть обучается на базе размеченных семантических фреймов на русском языке Framebank. Отдельной трудностью является то, что семантический фрейм относится к некоторой предметной области, а поэтому и роли, размеченные в нем. В результате модель выделяет в тексте порядка ста различных ролей, назовем их множество R', а множество ролей по Филлмору за R. Следующая задача - кластеризовать роли из R' в роли из R. Для этого мы разметили набор новостей так, что каждая роль по Филлмору встретилась в нем порядка 30 раз, обозначим его за D_{test}. Запустив модель выделения ролей

Algorithm 1 Поиск фактов в документе

```
1: procedure FACTS(d)
2:     preprocess(d)
3:     S ← sentences(d)
4:     for s in S do:
5:         dep_tree(s) ← SyntaxNet(s)
6:         predicates ← verbs and participles in s
7:         for p in predicates do:
8:             subj ← nsubj children of p
9:             obj ← dobj children of p
10:            for (w_1, w_2) in subj × obj do:
11:                facts.append(w1,p,w2)
12:        nouns ← all nouns in s
13:        for n in nouns do:
14:            appos← appos children of n
15:            for a in appos do:
16:                facts.append(n,есть,a)
17:            adj ← all adjectives in s
18:            for a in adj do:
19:                facts.append(parent n,есть,a)
```

на D_{test}, получим роли из R' в нем:

$$D'_{test} = \{(w, r'),\ \ w \in D,\ r' \in R'\},\ \ D_{test}\{(w, r),\ \ w \in D,\ r \in R\}$$

Тогда каждой роли из R' будет соответствовать наиболее вероятная роль по Филлмору:

$$F(r') = \arg\max_{r \in R} \sum_{(w,s') \in D'_{test}} [s' = r', (w, r) \in D_{test}]$$

Здесь под $F(r')$ подразумевается операция преобразования ролей из R' в роли по Филлмору. В итоге мы получаем пары вида слово-роль по Филлмору.

3.3. Поиск тональных слов

Задача их поиска тональных слов решается в два этапа:

1. Сбор словаря тональных слов

2. Тегирование слов в тексте тональностями

Для первого мы использовали словарь из ресурса Linis Crowd [10], который получен из политических и социальных текстов. Изначально он содержал 2454 слов, и первое, что нужно сделать - обогатиться его. Для этого мы для каждого слова добавили его синонимы и гипонимы с тем же сентименом и антонимы с противоположным, используя RuWordNet [11]. В результате получили словарь из 3419 слов, где у каждого слова оценка +1 или -1. После этого мы протегировали слова в тексте в соответствие с их оценками в словаре. Но таким образом мы отмечаем совсем небольшую часть слов и никак не учитывает контекст, который также влияет на сентимент слова.

Мы применили подход, когда отмечаются не только слова из словаря, но и связанные с ними. Это делается на основании набора эвристических правил:

- Если отмеченное слово - существительное, прилагательное или наречие, его родитель отмечается той же тональностью
- Если отмечен глагол, то связанные с ним субъекты и объекты отмечаются той же тональность
- Если со словом связана отрицательная частица, его тональность меняется на противоположную

Результат работы такого алгоритма показан на рисунке 4. Можно заметить, что с помощью такого подхода удалось отметить, к примеру, что в данном тексте слово *Россия* с отрицательным отношением.

3.4. Построение вероятностной модели

Возвращаясь к задаче оптимизации (7), можем заметить, что в целевой функции фигурируют частоты встречаемости термов в документе n_{dw}. В разделе 2 мы сказали, что будем работать с пятью модальностями: $M = \{subjects, objects, roles, positive, negative\}$, их словари $W = \{W^s, W^o, W^r, W^p, W^n\}$. Определим, как мы будем считать n_{dw} для них.

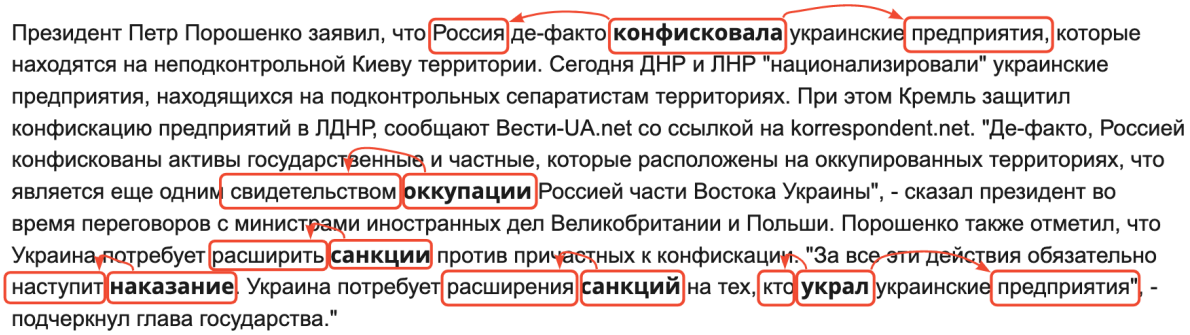

Рис. 4. Результат поиска тональных слов

Для модальности субъектов это то, как часто данный субъект встречается во множестве фактов, для объектов аналогично. Если обозначить множество триплетов документа за Tr_d:

$$n_{dw_s} = \sum_{(s,p,o) \in Tr_d} [s = w],\ w \in W^s, \quad n_{dw_o} = \sum_{(s,p,o) \in Tr_d} [o = w],\ w \in W^o$$

Для всех ролей мы ввели одну общую модальность, и частота терма - то, сколько раз встречалась пара слово-роль в документе:

$$n_{dw_r} = \sum_{(w,r) \in W \times R} [w = w, r = r],\ w \in W, r \in R$$

Для модальности положительно окрашенных слов мы считаем то, как часто слову приписывали оценку $+1$, для отрицательных - аналогично -1:

$$n_{dw_p} = \sum_{(w,ton) \in W \times \{-1,+1\}} [w = w, ton = p],\ w \in W, p \in \{-1, +1\}$$

$$n_{dw_n} = \sum_{(w,ton) \in W \times \{-1,+1\}} [w = w, ton = n],\ w \in W, p \in \{-1, +1\}$$

Найдем вначале решение мономодальной задачи оптимизации (5). Для этого нам необходимо получить частотные оценки распределений скрытых переменных $\varphi_{wo} = p(w|o)$ и $\theta_{od} = p(o|d)$. Введем обозначение n_{odw} - число раз, когда терм w относился к мнению o в документе d. Тогда $n_{od} = \sum\limits_{w \in W} n_{odw}$ - количество раз, когда некоторый терм из d относился к мнению o, а $n_{wo} = \sum\limits_{d \in D} n_{odw}$ - количество раз, когда терм w относился к мнению o в некотором документе. В таких обозначениях можем получить частотные оценки распределений:

$$\hat{p}(w|o) = \frac{n_{od}}{n_o}, \quad \hat{p}(o|d) = \frac{n_{wo}}{n_d}$$

Здесь $n_o = \sum\limits_{d \in D} \sum\limits_{w \in d} n_{odw} = \sum\limits_{w \in W} \sum\limits_{d \in D} n_{odw}$ - общее количество термов, относящихся ко мнению o, а $n_d = \sum\limits_{o \in O} \sum\limits_{w \in d} n_{odw}$ - число термов в d. Можно заметить, что в каждом выражении фигурирует n_{odw}. Условное распределение мнений можно оценить как $\hat{p}(o|d,w) = \frac{n_{odw}}{n_{dw}}$. С другой стороны, его можно выразить через φ_{wo} и θ_{od} с помощью формулы Байеса:

$$p(o|d,w) = \frac{p(o,w|d)}{p(w|d)} = \frac{p(w|o)p(o|d)}{p(w|d)} = \frac{\varphi_{wo}\theta_{od}}{\sum\limits_{o' \in O} \varphi_{wo'}\theta_{o'd}}$$

Таким образом, с помощью двух последних выражений, мы можем выразить параметры вероятностной системы φ_{wo} и θ_{od} через $p(o|d,w)$ и наоборот. Запишем эти выражения в виде системы:

$$\begin{cases} \hat{p}(o|d,w) = \frac{\varphi_{wo}\theta_{od}}{\sum\limits_{s \in O} \varphi_{ws}\theta_{sd}} \\ \varphi_{wo} = \frac{n_{od}}{\sum\limits_{w' \in W} n_{w'o}}; \quad n_{wo} = \sum\limits_{d \in D} n_{dw}\hat{p}(o|d,w) \\ \theta_{od} = \frac{n_{wo}}{\sum\limits_{o' \in O} n_{o'd}}; \quad n_{od} = \sum\limits_{w \in d} n_{dw}\hat{p}(o|d,w) \end{cases} \tag{8}$$

Мы получили систему уравнений для случая одной модальности без регуляризаторов. Решая ее методом простой итерации, мы поучим EM-алгоритм. Решение для общего случая дает теорема 1. Для ее формулировки введем оператор norm, который преобразует компоненты числового вектора $\mathbf{n}$ в корректное вероятностное распределение $\mathbf{p}$ по формуле:

$$p_i = \frac{(n_i)_+}{\sum\limits_j (n_j)_+}$$

Теорема 1. *Пусть функция $R(\Phi, \Theta)$ непрерывно дифференцируема. Тогда точка (Φ, Θ) локального экстремума задачи (7) удовлетворяет для всех*

невырожденных мнений o и документов d системе:

$$\begin{cases} p_{odw} = \underset{o\in O}{\text{norm}}(\varphi_{wo}\theta_{od}) \\ \varphi_{wo} = \underset{w\in W^m}{\text{norm}}(n_{wo} + \varphi_{wo}\frac{\partial R}{\partial \varphi_{wo}}); \quad n_{wo} = \sum\limits_{m\in M}\sum\limits_{d\in D}\tau_m n_{dw} p_{odw} \\ \theta_{od} = \underset{o\in O}{\text{norm}}(n_{od} + \theta_{od}\frac{\partial R}{\partial \theta_{od}}); \quad n_{od} = \sum\limits_{m\in M}\sum\limits_{w\in W^m}\tau_m n_{dw} p_{odw} \end{cases}$$

3.5. Добавление регуляризаторов

В этом разделе мы определим функциия регуляризаторов, которые мы ввели в целевой функции (6): $R(\Phi, \Theta)$.

В словарях модельностей W^m есть как предметные термы, так и фоновые. К первым относятся те, которые выражают мнение автора, ко вторым - общая лексика. Считая, что каждое мнение должно выражается небольшим ядром термов, получаем, что распределения термов по мнениям $\varphi_{wo} = p(w|o)$ должны быть разреженными для предметных мнений. Кроме того, на практике каждая новость чаще всего выражает только одно мнение, поэтому распределения мнений по документов $\theta_{od} = p(o|d)$ также должны быть разреженными.

Этим эвристикам отвечает **разреживающий регуляризатор**. Он максиимизирует разницу между моделируемыми распределениями φ_o и θ_d и заданными распределениями $\beta = (\beta_w)_{w\in W}$, $\alpha = (\alpha_o)_{o\in O}$, которые полагают равномерными. Под разницей распределений мы понимаем дивергенцию Кульбака-Лейблера:

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

Мы будем выделять отдельные фоновые мнения с общими термами, которые не выражают никакое реальное мнение. Они нужны, чтобы избавить от общей лексики предметные мнения. Чтобы реализовать это, фоновые мнения должны быть близкими к равномерному для этого мы наклыдываем на них **сглаживающий регуляризатор**. Он получается, как разреживающий

с обратным знаком:

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

Помимо разделеиния предметных мнений от фоновых, в нашей задаче требуется, чтобы мнения различались существенно, то есть чтобы φ_o не были коррелированы. Это возможно реализовать при помощи **регуляризатора декоррелирования**:

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

Он увеличивает расстояния между распределениями термов по мнениям.

4. Вычислительный эксперимент

Для оценки качества модели мы собрали и разметили два корпуса документов:

1. **Бизнес ЛНР** - 82 новости про национализацию предприятий в ЛНР и ДНР из российских из украинских источников. Средний размер текста - 200 слов. Тексты относятся к двум мнениям: позиция России и позиция Украины.

2. **Трами Париж** - 218 новостей про решение Трампа выйти из Парижского соглашения. Средний размер текста также 200 слов. Тексты поровну разделены между двумя мнениями: позиция Трампа и его противников (таких, как Маск).

Корпусы размещены в свободном доступе для использования в репозитории[1]. Процесс разметки проходил следующим образом: вначале были собраны тексты по выбранному политическому событию до тех пор, пока они не начале повторяться. Затем был просмотрен весь датасет, чтобы установить общее количество мнений. После этого два независимых асессора поставили мнение

[1] https://github.com/newfteddy/opinion_mining_features/tree/master/data

у каждой новости, доля согласия была 91%. В спорных случаях голос третего асессора был решающим.

Инструменты разработки. Программная реализация модели и алгоритмов извлечения признаков была реализована на языке python. Для построения деревьев зависимостей использовался анализатор SyntaxNet. Для построения вероятностной модели использовалась библиотека BigARTM. Для кластеризации в лексической модели была применена реализация k-means в библиотеке skitit-learn.

Метрики качества. Для оценки кластеризации мы будем использовать три метрики: precision, recall и F1-score. Они определяются как:

$$Pr = \frac{tp}{tp + fp}, \quad Rec = \frac{tp}{tp + fn}, \quad F1 = 2\frac{Pr \cdot Rec}{Pr + Rec}$$

В нашем случае правильным классом мы будем считать больший, так как зачастую основная часть новостей выражает одно мнение. Верными примерами кластеризации будем считать новости, в которых корректно определено это мнение.

4.1. Лексический бейслайн

Один из основных вопросов, который появляется при рассмотрении нашего подхода заключается в целесообразности использования синтаксических и семантических признаков в целом. Для проверки этого мы будем сравнивать наши модели с лексической моделью. После предобработки, векторным представлением документа мы будет считать tf-idf всех слов в нем. Кластеризацию мы будем производить при помощи алгоритма k-means. Признаковый вектор документа d можем записать как $\text{tf-idf}(w, d, D) = tf(w, d) \times idf(w, D)$, где

$$tf(w, d) = \frac{n_{wd}}{\sum\limits_{w \in d} n_{wd}}; \qquad idf(w, D) = \log \frac{|D|}{|d \in D | w \in d|}$$

Важно принять во внимание, что результат кластеризации k-means зависит от начальной точки, что особенно существенно на небольших корпусах, таких как наш. Для репрезентативных результатов мы провели мультизапуск кластеризации и усреднили результат по 100 экспериментам. Результаты представлены на рисунке 5: график показывает усредненный результат

кластеризации в зависимости от количества экспериментов. На корпусе Бизнес ЛНР результат сошелся к значению 0.67, а на корпусе Трамп Париж - к 0.72.

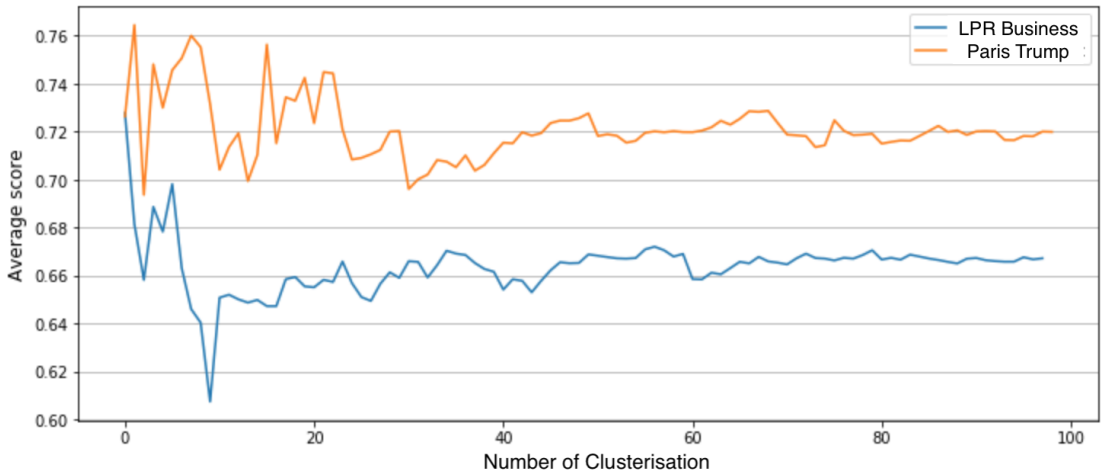

Рис. 5. Усредненая F1-мера для лексического бейслайна

4.2. Подбор параметров модели

Основные гиперпараметры, которые определяют нашу модель:

- Веса модальностей $\{\tau_m\},\ m \in \{s, o, r, p, n\},\ \sum_m \tau_m = 1$
  Это наиболее важный параметр - распределение весов модальной и задает определение мнения. Одной из целью работы является формализация мнения, то есть определение весов, с которыми факты, семантические роли и тональные слова оказывают вклад в оптимальное значение кластеризации.

- Коэффициенты регуляризации τ
  Эти параметры задают то, насколько сильно мы сглаживаем фоновые мнения и разреживаем предметные, в нашем случае мы положили его общим для этих двух регуляризаторов. Для каждой модальности эти коэффициенты определяются по-отдельности. У регяляризатора декорреляции также свой отдельный коэффициент.

- Минимальный TF модальностей

Мы фильтруем словари модальностей по некоторому порогу, чтобы включать только те термы, которые встречаются достаточно часто.

Оптимальный набор гиперпараметров мы подбирали в несколько этапов:

1. Фиксируем коэффициент регуляризации на значении $\tau = 1.0$ и подбираем минимальный TF для каждой модальности. Для этого мы строим модели, которые кластеризуют новости только по одному из признаков и оцениваем качество этих моделей. Данный параметр подбирается в первую очередь, так как он определяет словарь модальностей и принимает небольшое количество дискретных значений.

2. На отфильтрованных словарях оптимизируем коэффициент регуляризации для каждой модальности. τ подбирается во вторую очередь, так как его оптимизации производится для каждого признака в отдельности. Результат подбора коэффициентов для каждой модальности показан на рисунке 6 для корпуса Бизнес ЛНР и на рисунке 7 для Трамп Париж. Можно заметить, что этот параметр достаточно сильно влияет на качество моделей, но во всех случаях имеет выраженный максимум в районе 0.8. Можно также заметить, что на первом корпусе диапазон значений F1-меры значительно шире. Это связано со сравнительно небольшим размером датасета.

3. Фиксируем коэффициенты регуляризации и на отфильтрованных словарях оптимизируем веса модальностей $\tau_m,\ m \in \{s, o, r, p, n\}$. Для этого мы делаем выборочный перебор по параметрической сетке: начинаем с пары признаков факты-роли и находим оптимальной значение в такой модели. Затем итеративно добавляем ненулевой вес тональным словам и двигаемся в сторону увеличения, проверяя соседние точки по остальным признакам. В районе полученного оптимума делаем более точную оценку по параметрической сетке с мелким шагом. Результат на обоих корпусах представлен в таблице 1.

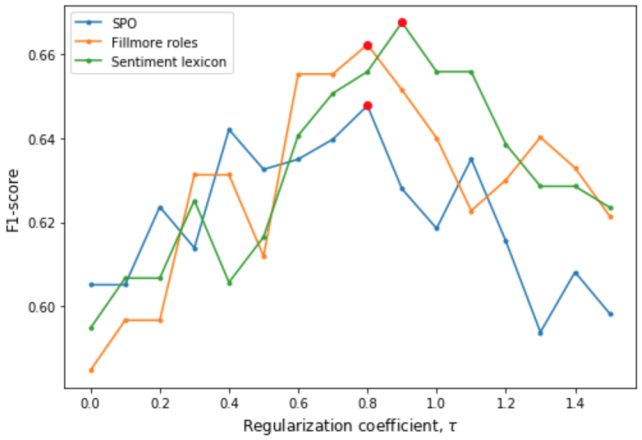

Рис. 6. Оптимальный τ на ЛНР Бизнес

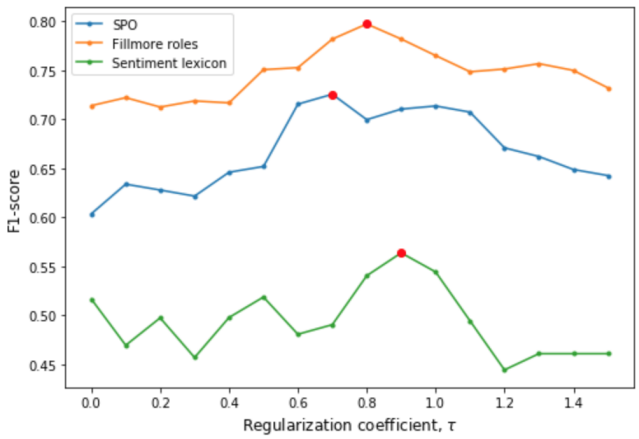

Рис. 7. Оптимальный τ на Трамп Париж

Subjects	*Objects*	*Pairs*	*Positive*	*Negative*
0.14	0.06	0.22	0.29	0.29
0.1	0.04	0.36	0.25	0.25

Таблица 1. Оптимальные веса модальностей

4.3. Результаты

Для проверки качества мы будем сравнивать precision, recall и F1-меру. Мы оцениваем кластеризацию на два класса, поэтому модель с константой дала бе результат 0,5. Цель эксперимента - установить, какие признаки определяют понятие мнения. Для этого мы будем строить вероятностную модель на всевозможных наборах модальностей. Мы будем использовать комбинации:

- Фактов (SPO)
- Семантических ролей по Филлмору (FR)
- Тональных слов (Sent)

Результаты представлены в таблицах 2 и 3 для двух корпусов документов. Первая строчка показывает лексический бейслайн. Вначале мы построили вероятностные модели с единственным признаком, им соответствуют следующие три строчки. Затем мы построили модели на попарных комбинациях призаролинаков: факты с семантическими ролями (SPO+FR), факты с

тональными словами (SPO+Sent) и семантические роли с тональными словами (FR+Sent). Последняя строчка показывает полную композитную модель, построенную на комбинации всех трех признаков.

Модель	Pr	Rec	$F1$
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
SPO+FR+Sent	**0.77**	**0.97**	**0.86**

Таблица 2. ЛНР Бизнес

Модель	Pr	Rec	$F1$
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
SPO+FR+Sent	**0.77**	**0.94**	**0.85**

Таблица 3. Трамп Париж

Первый вывод, который мы можем сделать: композитная модель из всех трех признаков показывает существенно лучшее качество, чем лексический бейслайн.

Модели с единственным признаком при этом показали достаточно низкое качество около 0.65 на первом корпусе и 0.72 на втором, того же порядка, что лексический бейслайн. Это говорит нам о том, что ни один из признаков сам по себе не разделяет тексты по мнения, и это не удивительно. Тем не менее, изучив результаты моделей, мы можем заметить, что они выдавали ошибочный результат на различных подмножествах новостей. Это также подтверждается эксперементально тем, что объединив любую пару признаков, мы сразу получаем заметный прирост качества во всех случаях - такие модели показали F1-меру порядка 0.8. Объединив все три признака, мы смогли еще сильнее улучшить модель и разделить тексты по мнения с достаточно хорошим качеством выше 0.85 в обоих случаях. При этом сложилась достаточно похожая картина на обоих корпусах.

Такой результат интерпретируем: каждый признак по-отдельности позволяет выделить один из аспектов, который автор использует для выражения своего мнения, причем эти аспекты различные. Композитная модель позво-

ляет найти их всех одновременно.

4.4. Устойчивость результатов

Результат, который мы получили справедлив для оптимальной точки весов модальностей τ_m. Появляется естественный вопрос: насколько устойчив такой результат? Возможно, это случайный результат для конкретной комбинации модальностей, и повторить его на практике без учителя будет невозможно. Чтобы ответить на этот вопрос, исследуем попарные модели и рассмотрим их результат в зависимости от весов модальностей. Например, для пары факты-роли мы перебрали вес последних в диапозоне $\tau_r \in \{0, 0.05, 0.1, ..., 1\}$, а вес ролей определили как $1 - \tau_r$. Подобные эксперименты проделали со всеми парами на обоих корпусах. Результаты представлены на на рисунках 8 и 9 для двух датасетов. Рассмотрим сначала 9. Каждый квадрат показывает F1-меру модели на паре признаков с некоторым весом. Для пары роли-факты (третья строка) слева показан результат модели, использующей исключительно семантические роли, а справа - использующей только факты, ее качество - 0.72. По мере движения влево, добавляя больше веса модальности ролей, качество начинает увеличиваться до момента распределение весов $\{\tau_r = 0.7, \tau_s + \tau_o = 0.3\}$ и постепенно уменьшается после этого. Похожее поведение можно заметить для других пар признаков, что позволяет сделать вывод об устойчивости результатов по отношению к весам модальностей. Переходя к 8, можем заметить похожую картину, однако результаты не настолько устойчивы. Можем объяснить это тем, что его размер несколько меньше. Для дальнейшего подтверждения такого предположения мы планируем проводить эксперименты на других корпусах документов большего размера.

5. Заключение

В данной работе была поставлена задача кластеризации новостей по мнениям без учителя. Мы предложили подход с использованием семантических и синтаксических признаков, который ранее не был применен для opinion mining, тем более на русскоязычных текстах. Были разработаны алгорит-

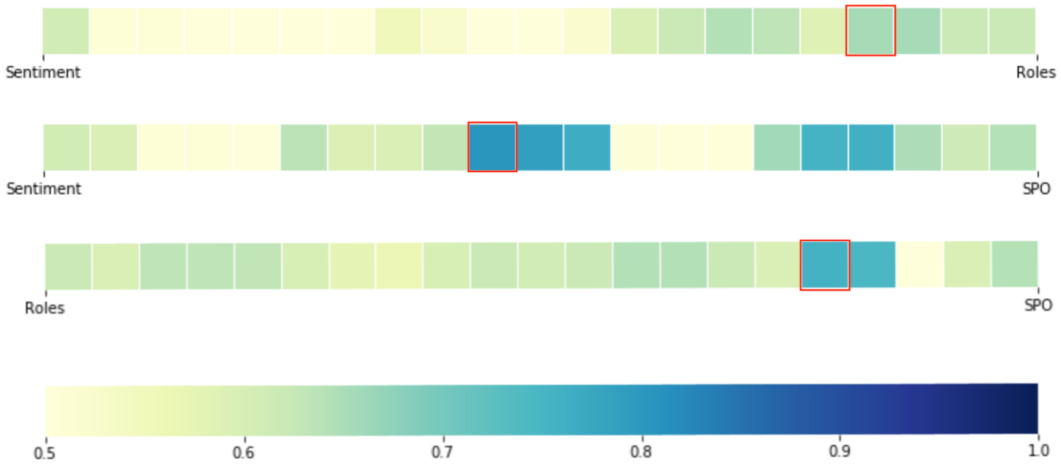

Рис. 8. F1-score distribution over modalities weights for Corpus 1

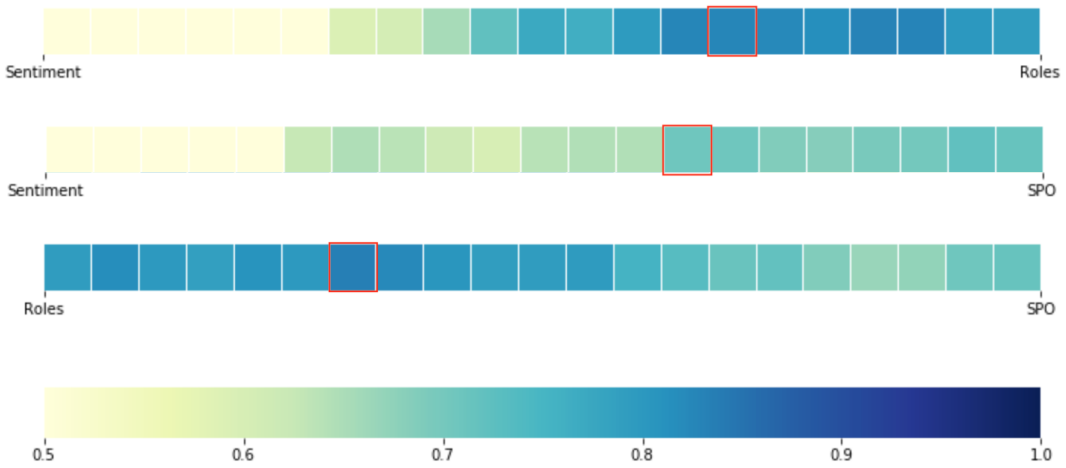

Рис. 9. F1-score distribution over modalities weights for Corpus 2

мы поиска фактов, семантических ролей и тональных слов в русскоязычных текстах и предложена генеративная модель текста на их основе. Для оценки качества мы собрали и разметили два корпуса документов, которые сделали доступными для использования. В результате экспериментов мы смогли сделать несколько выводов. Во-первых, лексические признаки не подходят для кластеризации текстов по мнения. Во-вторых, совместное использование фактов, семантических ролей и тональных слов дает существенный прирост качества, позволив нам кластеризовать новости на двух корпусах с качеством выше 0.85. Наконец, анализ устойчивости результата позволил формализовать понятие мнения как композицию фактов, семантических ролей и тональных слов.

Список литературы

1. Q. Mei, X. Ling, M. Wondra, H. Su, and C. Zhai. *"Topic sentiment mixture: Modeling facets and opinions in weblogs"*. In *Proceedings of the World Wide Conference* (2007), pp. 171-180
2. B. Pang, L. Lee. *"Opinion Mining and Sentiment Analysis. Foundations and Trends"*. In *Information Retrieval* (2008), pp. 1–135
3. M.S. Hajmohammadi, R. Ibrahim, Z.A. Othman *"Opinion Mining and Sentiment Analysis: A Survey"*. In *International Journal of Computers & Technology Vol. 2 No. 3* (2012)
4. M.J. Paul, R. Girju *"Cross-Cultural Analysis of Blogs and Forums with Mixed-Collection Topic Models"*. In *Proc. of EMNLP '09* (2009), pp. 1408–1417
5. Y. Fang, L. Si, N. Somasundaram, Z. Yu *"Mining Contrastive Opinions on Political Texts using Cross-Perspective Topic Model"*. In: *Proc. of WSDM '12* (2012), pp. 63-72
6. R. Balasubramanyan, W. W. Cohen, D. Pierce, D. P. Redlawsk *"Modeling Polarizing Topics: When Do Different Political Communities Respond Differently to the Same News?"*. In: *Proc. of the Sixth International AAAI Conference on Weblogs and Social Media* (2012), pp. 18-25
7. M.J. Paul, R. Girju *"A Two-Dimensional Topic-Aspect Model for Discovering Multi-Faceted Topics"*. In: *Proc. of AAAI '10* (2010), pp. 545-550
8. X. Zhu, D. Klabjan, P.N. Bless *"Unsupervised Terminological Ontology Learning based on Hierarchical Topic Modeling"*. In: *Proc. of ACL 17* (2017)
9. E.I.Bolshakova, K.V.Vorontsov and others *"Automatic word processing in natural language and data analysis"*. pp. 195-228
10. Koltsova, O.Yu, S.Alexeeva and S.Kolcov *"An opinion word lexicon and a training dataset for russian sentiment analysis of social media."*. In: *Proc. of the International Conference "Dialogue 2016"* (2016)
11. Lashevich G. et al. *"Creating Russian WordNet by Conversion."*. In: *Proc. of the International Conference "Dialogue 2016"* (2016)
12. P. Sobkowicz, M. Kaschesky, G. Bouchard *"Opinion mining in social media: Modeling, simulating, and forecasting political opinions in the web"*. In: *Government Information Quarterly vol. 29* (2012) pp. 470-479

13. Charles J. Fillmore *"Some problems for case grammar."*. In: *22nd Annual Round Table. Linguistics: Developments of the Sixties—Viewpoints of the Seventies. Volume 24 of Monograph Series on Language and Linguistics.* (1971) pp. 35–56
14. H. Wang and C. Zhai *"Generative Models for Sentiment Analysis and Opinion Mining"*. In: *Springer International Publishing AG* (2017)
15. S.M. Kim, E. Hovy *"Extracting opinions, opinion holders, and topics expressed in online news media text"*. In: *Proc. of the Workshop on Sentiment and Subjectivity in Text* (2006) pp. 1-8
16. K. Moilanen, S. Pulman *"Sentiment Composition"*. In: *Proc. of RANLP-2007* (2007) pp. 378–382
17. I. A. Karpov, M. V. Kozhevnikov, V. I. Kazorin , N. R. Nemov *"Entity Based sentiment analysis using syntax patterns and convolutional neural network"*. In: *Proc. of the International Conference «Dialogue 2016»* (2016ц) pp. 378–382
18. A. O. Shelmanov, D. A. Devyatkin *"Semantic role labeling with neural networks for texts in Russian"*. In: *Proceedings of the International Conference "Dialogue 2017"* (2017)
19. D. Marcheggiani, A. Frolov, and I. Titov *"A simple and accurate syntax-agnostic neural model for dependency-based semantic role labeling"*. In: *arXiv preprint arXiv:1701.0259* (2017)
20. C. A. Thompson, R. Levy, and C. D. Manning *"A Generative Model for Semantic Role Labeling"*. In: *Springer-Verlag Berlin Heidelberg 2003* (2003), pp. 397–408
21. A. Giuglea and A. Moschitti *"Semantic Role Labeling via FrameNet, VerbNet and PropBank"*. In: *Proc. of the 21st International Conference on Computational Linguistics and 44th Annual Meeting of the ACL* (2006), pp. 929–936