ФЕДЕРАЛЬНОЕ ГОСУДАРСТВЕННОЕ БЮДЖЕТНОЕ УЧРЕЖДЕНИЕ НАУКИ
ИНСТИТУТ ЯЗЫКОЗНАНИЯ РОССИЙСКОЙ АКАДЕМИИ НАУК

На правах рукописи

Куликов Сергей Юрьевич

Автоматическое извлечение мнений: лингвистический аспект

Специальность 10.02.21 – Прикладная и математическая лингвистика

Диссертация на соискание ученой степени
кандидата филологических наук

Научный руководитель:
доктор филологических наук
Рябцева Надежда Константиновна

Москва 2016

Оглавление

Введение

Современное информационное общество характеризуется большими объемами быстро меняющейся информации. Условия конкуренции и рыночной экономики диктуют новые требования к информационному обеспечению человека. Скорость доступа к информации зачастую является определяющим фактором при принятии решений. В связи с тем, что б*о*льшая часть информации в мире функционирует в языковой форме, перед лингвистикой (особенно прикладной) встают новые задачи.

Современная лингвистика характеризуется значительным интересом к субъективным факторам естественных языков, что проявляется в господстве когнитивно-дискурсивной или антропоцентрической парадигмы в современном языкознании.

Информационные технологии (и особенно технологии в сети Интернет) бурно развиваются. Ежедневно более 80% американцев и жителей развитых стран пользуются Интернетом, в СНГ доля пользователей составляет около 60% [33: 4]. При этом значительный интерес вызывают различного рода ресурсы, содержащие в себе оценочную информацию — форумы, чаты, Интернет-дневники и т.д. Поэтому на фоне бурного роста информационных технологий и усиленного внимания к человеку зародилось и успешно развивается автоматическое извлечение мнений.

Автоматическое извлечение мнений занимается проблемами определения отношения автора текста к описываемой в тексте проблеме, событию или продукту. По состоянию на 2010 год в мире насчитывалось более 100 действующих систем. Самыми известными из компаний, применяющих подобную технологию, являются Майкрософт, социальные сети Фейсбук и Твиттер, Интернет-магазин Амазон. В настоящее время в России действует более десяти коммерческих систем. Они применяются для задач оптимизации информационного поиска, бизнес-разведки, анализа причин роста или падения интереса к продукту или явлению, совершенствования систем автоматической проверки стиля и систем фильтрации спама. Отдельным бурно развивающимся

направлением является государственный мониторинг Интернета, направленный на предотвращение противоправной деятельности и анализ политической активности населения. На современном этапе наблюдается оторванность разработок в области автоматического извлечения мнений от лингвистических исследований по сходной тематике.

Актуальность исследования обуславливается важностью изучения общественного мнения и разного рода оценочных компонентов в сфере Интернет-коммуникации и необходимостью разработки принципов их автоматического извлечения из различных типов и видов текста.

Научная новизна исследования заключается в разработке лингвистических принципов автоматического извлечения мнений, призванных значительно повысить качество существующих технических средств идентификации субъективных компонентов контента. Что объясняется практической необходимостью совершенствования действующих систем, построенных на основе статистических моделей без учета собственно лингвистических факторов. Также в рамках диссертационного исследования уточнена классификация оценочной лексики, в частности введено понятие однореферентных оценочных слов.

Теоретическая значимость исследования состоит в систематизации лингвистической информации, необходимой для задач автоматического извлечения мнений, а также в развитии понятийного аппарата рассматриваемой предметной области. Практическая значимость заключается в возможности повышения качества действующих систем автоматического извлечения мнений за счет совершенствования лингвистического обеспечения. Материалы диссертации могут быть использованы при разработке таких ресурсов по автоматическому извлечению мнений как словари оценочной лексики, базы данных и размеченные корпусы текстов, а также при разработке комплексной системы автоматического извлечения мнений.

В качестве материала исследования выбраны тексты сети Интернет (блоги, сообщения информационных агентств и пользовательские отзывы на продукты и

события) на русском языке: корпус ruTenTen (15,8 млрд. словоупотреблений); корпус русскоязычных СМИ и блогов (свыше 25,5 млн. словоупотреблений) и ряд других источников (около 150 тыс. словоупотреблений). Объектом исследования выбраны языковые и внеязыковые способы выражения оценок в электронных текстах. Предметом исследования являются способы формализации оценочных суждений.

Методология настоящего исследования сложилась в первую очередь на базе работ отечественных и зарубежных специалистов в области теории оценки (Н.Д. Арутюнова, Е.М. Вольф, Г.Г. Шпет, В.Н. Телия, В.А. Маслова, И.Г. Жирова и др.), психолингвистики (И.Н. Горелов, А.А. Леонтьев, Ю.А. Сорокин и др.), теории автоматической обработки текста (Н.Д. Андреев, Г.Г. Белоногов, Л.Н. Беляева, Ю.Н. Марчук, И.А. Мельчук, Л.Л. Нелюбин, А.И. Новиков, Н.В. Лукашевич, R. Schank, R. Mitkov, Y. Wilks, W. Daelemans и др.), теоретической семантики и синтаксиса (Ю.Д. Апресян, Е.В. Падучева, Л.М. Васильев, и др.) и практики автоматического извлечения мнений (А.Н. Соловьев, Т.Е. Загибалов, И.И. Четверкин, J.M. Wiebe, B. Liu, L. Lee, M. Klenner, S. Pulman, M.-T. Taboada и др.).

Специфика объекта изучения обусловила применение следующих методов исследования: классификации, корпусного, контекстуального и дискурсивного анализа, метод «черного ящика», различные статистические методы и метод моделирования. На некоторых этапах работы применялся компонентный анализ.

Целью диссертационного исследования является разработка принципов создания лингвистического обеспечения системы автоматического извлечения мнений для анализа текстов на русском языке.

Для достижения поставленной цели были поставлены следующие задачи:

1. Изучить особенности существующих систем автоматического извлечения мнений;
2. Проанализировать типы оценочной информации, моделируемые в существующих системах;

3. Определить классы оценочной лексики, необходимые для повышения качества автоматического извлечения мнений;
4. Уточнить определение понятия «мнение», принятого в практике автоматического извлечения мнений;
5. Разработать методы фильтрации априорно нейтрального контента;
6. Определить фрагменты этапов автоматического извлечения мнений, зависящие от аспекта задачи;
7. Разработать принципы автоматического и автоматизированного создания оценочных ресурсов для русского языка.

Степень разработанности проблемы. Автоматическое извлечение мнений является одной из наиболее динамично развивающихся областей компьютерной лингвистики. В последнее время особую актуальность приобретает разработка новых лингвистических ресурсов, принципов их создания, а также теоретическая обоснованность данных принципов. Необходимо отметить, что для анализа текстов на русском языке в настоящее время не выработано универсальных, общепринятых критериев для автоматического определения той или иной оценки.

На защиту выносятся следующие положения:

1. Современное автоматическое извлечение мнений основывается преимущественно на методах машинного обучения. Из методов машинного обучения наибольшее значение получили методы обучения «с учителем», которые не позволяют оперативно исправлять ошибки классификации текстов на оценочные классы (позитивные, негативные, нейтральные).
2. Оценочная классификация текстов на уровне объектов наиболее полно отражает языковые свойства текстов на естественном языке. Для отдельных типов текста целесообразна иерархическая структура представления объекта, опирающаяся на его свойства. Для иерархической структуры объекта оптимальным представляется использование специализированных тезаурусов или онтологий.
3. При выделении объектов оценки из текстов необходимо проводить разграничение субъект-объектных и причинно-следственных отношений.

4. Для разных задач автоматического извлечения мнений требуется различная организация лингвистического обеспечения. Эти отличия заключаются в наличии или отсутствии дополнительных этапов автоматизированного анализа текстов. Наиболее сложной организацией лингвистического обеспечения обладают системы автоматической идентификации противоправного контента в сети Интернет.
5. При разработке лингвистических ресурсов для автоматического извлечения мнений требуется учитывать механизмы текстовой референции. К данным механизмам относятся деривационные модели имен прилагательных и принципы оценочного словосложения. Ключевым понятием оценочной референции также является ограничение на количество объектов-референтов у оценочных слов.

Апробация работы. Материалы диссертации обсуждались на заседаниях сектора прикладного языкознания Института языкознания РАН в 2011-2014 гг.. Основные положения диссертации были изложены на следующих международных конференциях: Всероссийская студенческая научно-практическая конференция «Проблемы современной лингвистики и методики преподавания иностранных языков» (Коломна, 2010-2011), «II Межвузовский студенческий форум по прикладной лингвистике» (Жуковский, 2011), «Международная конференция студентов-филологов» (СПб, 2010-2011), «Актуальные задачи лингвистики, лингводидактики и межкультурной коммуникации» (Ульяновск, 2010, 2012), «Translation and Technology» (TRALOGY-2011) (Paris, 2011), Международная конференция «Язык. Культура. Общество» (Москва, 2011), Международная конференция студентов, аспирантов и молодых учёных «Ломоносов» (Москва, 2011-2014), школа-конференция молодых ученых ИЯз РАН (Москва, 2013-2015), V Международный конгресс исследователей русского языка «Русский язык: исторические судьбы и современность» (Москва, 2014), Computational Linguistics in the Netherlands and the Flanders (CLIN 24, CLIN 25) (Leiden, 2014, Antwerpen, 2015). Содержание работы

отражено в 24 публикациях, из которых 3 опубликовано в изданиях, рекомендованных ВАК при Минобрнауки России.

Некоторые решения, предложенные в работе, нашли применение в модуле лингвистической обработки текста системы «Аналитический Курьер».

Структура диссертации. Диссертация состоит из введения, трех глав, заключения, списка использованной литературы и приложений, содержащих список мотивационных целей, методику расширения словаря первичной оценочной лексики при помощи морфемного синтеза, примеры оценочно-размеченных корпусов текстов и фрагмент базы данных по этнофобонимам.

Глава 1. История и современное состояние автоматического извлечения мнений

Проблема субъективности в языке заинтересовала лингвистов-исследователей достаточно давно. Пионером в этой области, скорее всего, следует признать Э. Бенвениста [75: 295]. Однако, первые упоминания об анализе субъективности содержатся уже у Шпета [177] и у Фосслера [64: 470]. Несколько позже, в середине 1970-х гг., в среде заказчиков систем искусственного интеллекта появилась идея об интеллектуальном анализе данных. Таким образом, появились два разных направления — одно сугубо лингвистическое (теория субъективности в семантике и психолингвистика), другое сугубо инженерное. В ходе развития этих двух направлений наметились некоторые тенденции к объединению подходов, в основном за счет использования достижений другого подхода.

Выделение исследований категории оценки из общих исследований по субъективности объясняется, с одной стороны, требованиями заказчиков систем искусственного интеллекта, а с другой стороны, интересом лингвистов к конкретным проявлениям субъективности в языке. Мы рассмотрим теоретические и практические исследования по отдельности, что связано с относительной независимостью ранних прикладных разработок от лингвистических теорий языка. В Главе также рассмотрены современные отечественные системы автоматического извлечения мнений. В заключительном разделе обосновывается выбор термина «автоматическое извлечение мнений» в качестве базового для обозначения области знаний в целом.

1.1 Ранние системы интеллектуального анализа данных на основе теории субъективности (1976 – 1980)

К началу указанного периода уже сформировались некоторые теории, использующие субъективные и оценочные характеристики. К лингвистическим теориям относится, в первую очередь, лингвистическая теория «Смысл⇔Текст» и психолингвистическая модель невербальной коммуникации. В рамках теории «Смысл⇔Текст» были выделены оценочные лексические функции, а в рамках

психолингвистической модели невербальной коммуникации — показатели субъективности. В рамках теории искусственного интеллекта Роджер Шенк разработал модель концептуального анализа. Значительное место в ней уделялось целям, которыми руководствуется человек при выполнении того или иного действия.

Примерно в 1976 году в рамках проекта военно-морского министерства США группа под научным руководством Шенка (основной разработчик Хайме Карбонелл) начала разработку системы компьютерного понимания POLITICS. Основными задачами проекта были: 1) выявить модели поведения руководства США при определенных действиях со стороны СССР, 2) выявить различия в поведении представителей различных политических партий США, 3) разработать прикладную систему, основанную на данных анализа новостных сводок о внешнеполитических событиях по всему земному шару.

К концу 1978 года проект был реализован на ЭВМ. В дальнейшем исследования Хайме Карбонелла повлияли на формирование модели оценки поведения человека. Концепция анализа личностных качеств человека через ее текстовое проявление повлияла на формирование начальных идей автоматического извлечения мнений.[1] Рассмотрим более подробно основные принципы этой концепции.

Базовая гипотеза формулируется следующим образом: «эмоции подтверждают стереотип поведения». В качестве ее доказательства приводится следующий пример: 1) John is very ambitious. 2) John is very compassionate. Задача определить какое из предложений верно. Для определения используется следующий контекст: He abandoned his invalid mother, worked very hard at his job and badmouthed his coworkers. John was elated when the boss promoted him. [7: 50] Исходя из контекста верно только первое предложение. Для точного анализа используется принцип соответствия действия личностным целям (список целей был составлен Р. Шенком для задач концептуального анализа данных).

[1] В то же время интерес к проекту POLITICS на протяжении периода с середины 1980-х до конца 2010 года был практически равен нулю [Х. Карбонелл, личное сообщение].

Приведем список целей и некоторые стереотипы поведения.

Цели:

- приобретения (acquision goals)
- сохранения (preservation goals)
- наслаждения (enjoyment goals)

Полный список целей приведен в Приложении 1. Порядковый номер цели соответствует их рангу относительно характера человека и может варьироваться.

Для каждого человека строится следующая схема модели поведения, связанная с анализом контекста. Приведем фрагмент схемы поведения Джона (для представленной выше ситуации):

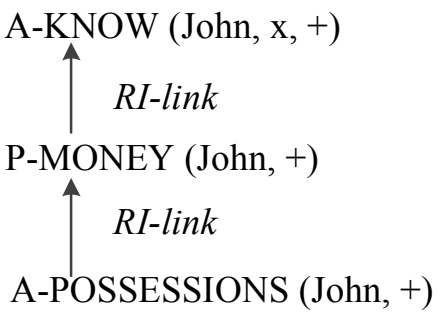

Рисунок 1. Фрагмент дерева целей (с учетом относительной важности) Джона[2]

Согласно рангу целей, с точки зрения относительной значимости, цели наслаждения менее важны, так как они могут меняться со временем. Процедура работы системы позволяет проводить поступательную коррекцию, т.е. видоизменение изначальной характеристики путем пополнения системы новыми знаниями. При этом возможно как сужение, так и уточнение задачи. В итоге анализа результатов разработки системы появилась теория планирования. В лингвистическом аспекте важными находками явились новая модель нахождения анафорических отношений [7: 51], указание на значимость контекста (в том числе глобального) и важность учета сочетаемости слов.

[2] A-KNOW – цели приобретения знаний, P-MONEY – цели сохранения денег, A-POSSESSIONS – цели приобретения собственности, RI-link – связь относительной значимости.

1.2 Непосредственные предшественники теории субъективности в компьютерной лингвистике (1980 – 1993)

В работе Йорика Уилкса и Биена (1983) [58] рассматривается возможность автоматического пополнения базы знаний диалоговой системы на основе анализа входного предложения. Рассматривая проблему поведения системы в приведенной ниже ситуации, авторы пытаются понять, почему система считает, что Фрэнк плохо думает о системе:

USER: Frank is coming tomorrow, I think

SYSTEM: Perhaps I should leave (I)

USER: Why?

SYSTEM: Coming from you that is a warning

USER: Does Frank dislike you?

SYSTEM: I don't know, but you think he does and that is what is important now. (11)[3]

Проблема заключается в объеме информации, которая занесена в память программы. Уилкс и Биен считают, что логические отношения, в случаях подобных данному, формируются по принципу логической триады. Этот принцип можно свести к следующей формуле — Важно не то, что я знаю об объекте, а то, что знает об объекте собеседник.

Существует два основных подхода к построению систем, которые могут решать сложные задачи общения в режиме диалога. Первый заключается в занесении в память программы всех возможных вариантов решения. Но такой подход представляет значительные трудности, как в алгоритмическом аспекте (проблема выбора нужного решения из списка возможных), так и в ресурсном аспекте (какое количество задач может решать система в условиях существующей компьютерной техники). Исходя из этих соображений, авторы предлагают строить диалоговые системы, способные решать задачи на основе анализа входного предложения (что Пользователь думает об отношении Фрэнка к системе и как ей нужно на это реагировать).

---

[3] [58: 96].

Несмотря на отсутствие четко прописанного в работе алгоритма, идеи, заложенные Уилксом и Биеном, в значительной мере сформировали изначальные взгляды основательницы автоматического извлечения мнений Джэнис Виби.

В своих ранних работах по анализу субъективности (до 1994 года) Дж. Виби рассматривала теоретический аспект представления субъективности в текстах различных жанров и пришла к выводу, что человеку свойственно вкладывать эмоции в любое описываемое событие. Были построены модели метафоры, точки зрения, эвиденциальности и т.д. Ключевым стало мнение о том, что указанные выше явления можно подвергнуть формализации.

Теоретическим итогом стала особая модель языковой субъективности, лежащая в основе текущих и прошлых исследований в этой области Виби и ее коллег. Она наиболее полно представлена в Виби и Рапапорт (1986, 1988, 1991) и Виби (1990, 1994). Данная модель была разработана для поддержки исследований по автоматической обработке естественно-языковых текстов и комбинирует теории из нескольких источников из областей достаточно далеких компьютерной лингвистики, таких как общее языкознание и теория литературы. Наиболее прямое влияние на модель оказали следующие работы: Долезел (1973) (типы субъективных показателей), Успенский (1970) (типы точек зрения), Курода (1973, 1976) (прагматика точки зрения), Чатмэн (1978) (рассказ в противопоставлении с дискурсом), Кон (1978) (лингвистические стили для представления сознания), Фодор (1979) (лингвистическое описание трудных контекстов) и особенно Банфилд (1982) (теория субъективности в противопоставлении с коммуникацией) [по 55].

1.3 Начальный этап развития систем автоматического извлечения мнений (1994 – 1996)

В 1994 году в журнале Computational linguistics появилась статья Виби [56], в которой она предложила алгоритм поиска точки зрения в повествовательных текстах и обосновала свою идею о том, что точка зрения тесно связана с субъективностью. Исходя из того, что основные эмоциональные проявления

состояния человека проявляются в предложениях, содержащих субъективность, Виби описала алгоритм, который учитывает внешний контекст ситуации. В работе, ставшей обобщением диссертационного исследования 1990 г., учитывались достижения модели Уилкса и Биена (1983) и, в меньшей степени, прослеживаются идеи Шенка-Карбонелла. Из-за отсутствия четкого понимания того, как можно формализовать внешний контекст, алгоритм не получил практической реализации.

Важнейшим достоинством работы стоит признать построение списка субъективных единиц текста. К их числу были отнесены прилагательные (poor), ряд именных групп (old bag), наречия (incredibly), модальные глаголы, фрагменты текста со сравнительными формами, неформальные обращения (Dad), условные предложения и некоторые оценочные глаголы.

Оценка работы алгоритма была проведена вручную на массиве более 700 страниц текста (17500 предложений), основой корпуса стали 7 художественных произведений различных авторов. Экспериментальная проверка на информантах подтвердила правильность гипотезы для человеческого восприятия.

Последующие работы автора и ее многочисленных учеников велись в направлении уточнения отдельных положений данной работы Виби.

1.4 Теоретические разработки в области субъективной оценки (1976 – 1996)

Прежде чем перейти к обзору раннего этапа разработок в области автоматического извлечения мнений обратимся к работам по общему языкознанию, в которых содержатся идеи и суждения о структуре оценочных отношений в языке.

В теоретических работах указанного периода наметились разнонаправленные тенденции. С одной стороны, развивалась теория «Смысл⇔Текст». В ее рамках велись работы по описанию слов разных языков (русский, английский, французский, испанский). Стоит отметить, что, несмотря на наличие единого подход к описанию семантических явлений, единоформатных описаний слов получено не было [166]. В то же время в рамках данной

общеязыковой теории были сформулированы некоторые значимые для теории и практики автоматического извлечения мнений положения. Так, авторы теории лексических функций, Ю.Д. Апресян [68] и И.А. Мельчук [144] впервые указали на количественные изменения в семантической структуре слова, входящего в коллокацию. Чуть позже И.И. Убин отмечал, что «не все слова языка могут быть усилены или ослаблены» [169: 164]. Указанный факт чрезвычайно важен для современного состояния разработок в области автоматического извлечения мнений, которые все больше ориентируются на сочетаемостные модели. Значительным недостатком (помимо отсутствия единого формата описания слов) мы считаем наличие дублирующих лексических функций (к таковым, в первую очередь, необходимо отнести лексические функции Bon и Pos, непосредственно связанные с формализацией описания оценок). С другой стороны, в области психолингвистики продолжались работы по созданию модели невербальной коммуникации. Наиболее полно исследования первого этапа представлены в работе И.Н. Горелова [88]. В дальнейшем, интерес к невербальным компонентам текста (и особенно к их формальным проявлениям) отошел на второй план.

Значительный интерес представляют работы группы «Логический анализ языка» конца 1980-х годов. Исследователи этой группы (прежде всего, Н.Д. Арутюнова и Е.В. Падучева) предложили несколько классификаций оценочных глаголов. В работе «Типы языковых значений: Оценка. Событие. Факт» Н.Д. Арутюнова приводит детальный обзор философских подходов к определению оценочных предикатов, ведущих свое начало с «Этики» Аристотеля. Именно Аристотелю принадлежит разделение объектов оценки и оценочных предикатов. Т. Гоббс указывал на субъективность ключевых критериев оценки – «хорошо» vs. «плохо». Дж. Локк показал континуальную природу данных критериев. Б. Спиноза определил зависимость между оценкой и свойствами оцениваемого объекта. И. Кант добавил принципиально иное понимание категорий «хорошего» и «плохого», выделив общественное «хорошо», общее для всех людей, т.е. «прагматический инвариант». В дальнейшем, в результате т.н. «лингвистического поворота» в философии, категория оценки стала более дифференцированной

благодаря обращению к конкретным текстовым реализациям различных модальностей. В XX веке в рамках теории эмотивности Ч. Стивенсон сделал попытку обосновать коммуникативный, прагматический характер оценок, связанный с воздействием на адресата сообщения. В дальнейшем были выделены сравнительные (компаративные) оценки, которые варьируются в зависимости от свойств субъекта оценки. Мы вынуждены ограничить наш обзор только данными соображениями логико-философского характера, т.к. полностью поддерживаем точку зрения Н.Д. Арутюновой, которая справедливо замечает, что «литература по этому вопросу необозрима» [69: 57]. Исследователи группы «Логический анализ языка» провели четкое разграничение между ключевыми, часто пересекающимися, понятиями «знание» и «мнение» (например, работа М.А. Дмитровской [92]). Категория экспрессивности также изучалась Е.М. Вольф, которая рассматривала экспрессивность в функциональном ключе. В своей книге «Функциональная семантика оценки» [83] она описала несколько типов оценки, уделяя особое внимание синтаксическим явлениям. В.Н. Телия в ряде работ по экспрессивности и фразеологии [164; 165] выявила соотношения экспрессивности и субъективности, а также указала на значимость экспрессивного компонента в толковании фразеологизмов.

В конце рассматриваемого периода появилась работа Л.М. Васильева по общему языкознанию, в которой автор рассмотрел типы оценочных предикатов для русского языка. В рамках создания «системного семантического словаря русского языка», фрагмент которого приводится в приложении к указанной работе [80: 125-174], была проведена их детальная классификация.

Следует заметить, что все описанные выше исследования носили скорее демонстрационный характер и были направлены не на полное и исчерпывающее описание того или иного лингвистического явления, а на выборочное описание одного объекта из исследуемого класса объектов. Именно такой подход отчасти объясняет некоторую оторванность прикладных разработок того времени от теоретических исследований.

Особо следует сказать о работе более раннего периода, которая оказала значительное влияние на работы в области автоматического извлечения мнений второй половины 2000-х гг. Это книга Б.А. Успенского «Поэтика композиции», изданная в 1970-м г. [170]. Особенно большое значение для т.н. «теории композициональности» имеют следующие главы: 1. «Точки зрения» в плане оценки, 4. «Точки зрения» в плане психологии и 5. Взаимоотношение точек зрения на разных уровнях в произведении. Сложная точка зрения. На основе идей Б.А. Успенского сформировалось представление об иерархической структуре мнений, отраженных в тексте.

1.5 Появление специальных ресурсов (размеченные корпусы текстов) (1997 – 2002)

Существенный прорыв в области компьютерных технологий в середине 1990-х годов произошел в связи с появлением сети Интернет. Глобальная сеть как источник данных стала причиной резкого увеличения объемов текста, привлекаемых к исследованиям. Большие ресурсы вычислительной техники и возможность удаленного доступа к коллекциям текстов привела к утверждению корпусной лингвистики в качестве одного из основных направлений в рамках компьютерной лингвистики.

Одно из первых применений корпусных технологий для задач автоматического извлечения мнений (в рамках анализа субъективности) пришлось на 1997 год [15]. В работе был использован Wall Street Journal corpus (21 млн. морфологически размеченных слов). На основе корпуса были проведены исследования семантической ориентации прилагательных. В их результате был составлен словарь, который лег в основу большого количества дальнейших разработок.

В последующих работах список слов расширялся, в него вводились новые типы прилагательных, наречия и глаголы. Значительно усложнилась схема разметки корпуса, помимо морфологических факторов стали применять

синтаксически размеченные корпусы. Итогом стали достаточно высокие результаты работы систем, которых на 2000 год было менее 10.

Примерно с 2000 года к разработке систем автоматического извлечения мнений подключились специалисты из других областей — автоматической обработки текста, информационного поиска, автоматического реферирования и др. Это привлекло в область большие финансовые потоки, а значительный рост доступности ресурсов и быстродействия компьютеров быстро привел к экономической эффективности разработок.

1.6 Внедрение методов машинного обучения в системы автоматического извлечения мнений (2002 – 2008)

Коренным переворотом в рассматриваемой нами области стало внедрение в практику разработчиков методов машинного обучения. Пионером в этой области следует признать Питера Тёрни [51]. В своей работе он использовал неразмеченный корпус.[4] Применив алгоритм машинного обучения без учителя и минимум лингвистических знаний, он добился лучших результатов, чем у [15]. Таким образом, была продемонстрирована принципиальная возможность применения машинного обучения к задачам автоматического извлечения мнений.

В последующих работах применялись другие методы машинного обучения (преимущественно методы обучения с учителем). Использование размеченных корпусов текстов и инкорпорация дискурсивной структуры позволили существенно повысить качество работы систем.

Помимо увеличения как объемов привлекаемой информации, так и глубины лингвистической разметки, исследователи обратили внимание на неоднородность оценки даже в рамках положительно трактуемых высказываний. Было замечено, что позитивные (также как и негативные) высказывания могут быть более позитивными или менее позитивными, не переставая выражать положительную оценку. Таким образом, значительным достижением описываемого периода следует признать более четкую дифференциацию оценки. Помимо оценочной

[4] В процессе первичной обработки текста к корпусу были применены процедуры морфологического анализа и извлечения биграмм с прилагательными.

шкалы в целом ряде систем были учтены уровни оценки, т.е. возможность определения оценки не только предложения, но и их последовательности, учитывая логическую структуру документа.

На этом этапе развития автоматического извлечения мнений более пристальное внимание было обращено на локальный контекст и дискурсивное поведение единиц в различных текстах. В настоящее время растет число работ по оценочным свойствам различных сочетаний слов, например, [37]; [22], а также пионерская работа Заенен [60]. К недостаткам существующих работ следует отнести то, что в них не учитываются степени изменения оценок сочетаний слов. В тоже время, часто отмечается, что «способы выражения категории *эмфатичность* разнообразны и многочисленны и затрагивают все уровни языка» [97: 10] и «эмотивность текста формируется за счет языковых средств всех уровней и работает на общую экспрессивность текста» [141: 185-186]. Впрочем, следует отметить, что четких правил определения оценки произвольных имен прилагательного и существительного в рамках одной именной группы с учетом их семантических признаков определено не было.

В рассматриваемый период произошел значительный рост числа систем автоматического извлечения мнений[5], что привело к расширению сферы применения технологии — именно в конце этого периода были проведены исследования политического дискурса, позволившие точно предсказать итоги президентских выборов в США [54].

1.7 Современный этап развития систем

Существенным фактором, послужившим прорывом в области автоматического извлечения мнений, стало появление в 2008 году первого подробного обзора по данной проблематике [40]. Появление работы в открытом доступе в начале 2009 года сделало ее одной из самых цитируемых в своем роде. По нашему мнению, этот обзор уделяет внимание в основном задачам, связанным с машинным обучением (проблема описания оценок и мнений в корпусе,

[5] к концу 2008 года только в США было более 20 фирм, специализировавшихся на задачах автоматического извлечения мнений.

значимые факторы для того или иного классификатора и др.), практически не касаясь подходов, основанных на правилах, которые при более низкой полноте имеют более высокую точность.

Основными достоинствами работы являются критический обзор работ, повлиявших на становление автоматического извлечения мнений, и описание направлений исследований. Ещё одним плюсом работы следует признать список ресурсов по автоматическому извлечению мнений, который приводится в Главе 7.

Авторы констатируют неудачность терминов рассматриваемой исследовательской области, однако за исключением узуального разграничения терминов *sentiment analysis* и *opinion mining* приводят лишь определения конкурирующих терминов в общем толковом словаре американского варианта английского языка Merriam-Webster's Online Dictionary [40: 8-10]. В тексте обзора указанные термины употребляются в качестве полных синонимов. Подобный подход позволил авторам дать относительно полную картину разработанности проблематики. С другой стороны, отсутствие четких определений терминов и их взаимосвязей не способствовали унификации терминологии данной предметной области.

Иную структуру имеет обзорная работа Б. Лю [26]. Она состоит из 11 глав, введения, заключения и библиографии. Признавая наличие других обзорных работ, Лю констатирует, что обзор Панг и Ли появился «сравнительно рано, в эпоху становления научной области» [26: 7]. Основное внимание автор уделяет различным аспектам анализа мнений. Так он выделяет 3 основных уровня анализа — уровень документа, уровень предложения и уровень объекта. На уровне объектов выделяется подуровень свойств объекта. Затем приводятся примеры того, какие методы применяются на каждом из указанных уровней. В первых двух главах Лю критикует ранний подход к оценочной классификации текстов на уровне документов. Там же приводится формальное определение мнения, о котором речь пойдет в Главе 2. Следует признать первенство Лю и в описании технологий определения оценочного спама. Интересным является и один из

заключительных параграфов, в котором отмечается, что исследователи «слишком сильно доверяли машинному обучению» [26: 134].

Критиками [8] уже отмечались некоторые недостатки обзора Лю. В первую очередь, это ориентация обзора на сферы обработки отзывов и некоторый перекос в сторону аспектной модели автоматического извлечения мнений. Отметим, что глава, посвященная аспектной модели, в два раза превышает по объему любую другую главу. Другим недостатком Карди считает излишнюю ориентированность на конкретные приложения информационного поиска.

По нашему мнению, к указанным недостаткам можно добавить отсутствие обзора собственно лингвистических подходов, например, подходов С. Пульмана и М. Кленнера, речь о которых пойдет ниже (Глава 2, разделы 2.4.1.5 и 2.4.1.6). К этому следует прибавить отсутствие ссылок на доступные оценочные ресурсы.

Восполнить отсутствие описания лингвистических подходов к решению задач автоматического извлечения мнений призвана обзорная статья М. Табоада, которая была опубликована в начале 2016 года [50]. Детальный анализ данной работы приводится в Главе 2, в разделе 2.4.1.5, посвященном данной тематике.

Первой отечественной работой, косвенно относящейся к рассматриваемой проблематике, по-видимому, является статья [77]. В ней описывается система, приписывающая словам эмоциональные значения на основе словаря. К сожалению, данная система, скорее всего, не получила дальнейшего развития. К плюсам описанного подхода следует отнести достаточно детальную классификацию классов оценки. Недостатком подхода авторов является сугубо словарный подход, опирающийся на словарь эмоциональной лексики русского языка начала 1990-х гг., без учета каких-либо правил сочетания оценочных слов.

В 2005 году на международной конференции «Диалог» был представлен первый в России доклад, непосредственно посвященный проблемам автоматического извлечения мнений. Основные идеи доклада нашли отражение в совместной публикации А.Е. Ермакова и С.Л. Киселева [93]. Несмотря на господство на том этапе развития автоматического извлечения мнений машинного обучения, авторы констатируют, что «в общем случае никакими

машинными методами невозможно разделить объективное и субъективное содержание текста — объективную констатацию фактов, пускай даже тонально окрашенных, и намеренное искажение действительности, в том числе сознательное выведение в фокус внимания определенных ее сторон на фоне замалчивания других» [93: 2]. По сути, описанная в статье система была одной из немногих систем в России, опиравшихся на сугубо словарные методы для определения оценки. Появлению данной системы способствовал государственный заказ на совершенствование систем Интернет-мониторинга. К 2007 году относятся первые масштабные работы по автоматическому извлечению мнений в России.[6] Эти работы были выполнены в русле применения методов машинного обучения для анализа текстов и повторяли (по методу) зарубежные исследования, но с привлечением нового материала, а именно текстов на русском языке (см. например, труды конференции «Диалог-2012»). Мы не рассматриваем данные работы, так как с лингвистической точки зрения они не представляют особого интереса.[7]

Русский язык привлекался в качестве вспомогательного материала диссертации Т.Е. Загибалова [61]. Автор указывает на несколько проблемных мест при проведении исследований по автоматическому анализу оценочной лексики в текстах на русском языке. Первым из них является отсутствие размеченных оценочных корпусов текстов. Данная задача была частично решена Т.Е. Загибаловым, а также организаторами соревнований по автоматическому извлечению мнений в рамках соревнования РОМИП. Другой проблемой является флективный характер русского языка. Это связано, в первую очередь, с традиционной организацией словарей оценочной лексики в исследованиях, посвященных английскому языку. Эта проблема решается Т.Е. Загибаловым при помощи стемминга. Принципиальным является его наблюдение о различном морфологическом составе оценочного инструментария языка [61: 112]. Так

---

[6] Первая обзорная работа была представлена А.В. Дегтевой на семинаре по автоматической обработке текста в 2006 году в Санкт-Петербурге [91].

[7] Исключение составляет система компании RCO [156], технология которой описана ниже в параграфе, посвященном системе Kribrum, которая является логическим продолжением данной системы.

подчеркивается значимость оценочных глаголов и существительных, а также больший набор оценочных конструкций в рассматриваемом исследователем корпусе отзывов по сравнению с аналогичным англоязычным корпусом.

К[8] собственно лингвистическим системам в России следует, в первую очередь, отнести систему, разрабатываемую коллективом под руководством А.Н. Соловьева [46]. В основе системы лежат принципы сочетаемости оценочных слов, сходные с предложенными А. Заенен в работе 2004 г. [60]. В ходе развития этих принципов сочетания оценочных слов стали делиться на усилительные, уменьшительные и нейтральные (см. также [128]). В системе имеется развитый лексический компонент и система автоматического расширения словаря за счет семантических классов. К недостаткам системы А.Н. Соловьева следует отнести отсутствие уменьшителей оценки[9]. Другим недостатком является отсутствие разграничения субъекта и объекта действия, а также единый класс усилительных слов.

К другим лингвистическим системам, работающим с русским языком, можно отнести систему Kribrum (компании «Ашманов и партнеры», с начала 2012 руководитель проекта — А.Е. Ермаков), разработки «АББИ» в рамках проекта Compreno (руководитель — А.С. Старостин) и систему компании «Эр-Си-О».

Особое внимание следует уделить системе Kribrum. Эта система является логическим продолжением разработок А.Е. Ермакова в компании RCO. Kribrum имеет развитый синтаксический компонент, использующий синтаксический анализатор компании «Диктум», а также компонент оценочных правил и онтологию[10], содержащую объекты оценки и их свойства. В основе правил лежат понятия семантических классов слов (прежде всего, глаголов) и типов допустимых синтаксических отношений. Правила имеют жесткую структуру и, чаще всего, сильно лексикализованы.

---

[88] Параграф опирается на работу автора [112]

[9] Уменьшители оценки были внедрены в 2012 году при непосредственном участии автора, принципы учета оценочных уменьшителей рассматриваются во второй главе.

[10] Под *онтологией* понимается инструмент формального описания фрагмента реального мира. В отличие от тезауруса онтология также содержит и экстралингвистические данные.

В период с 2000 по 2008 год быстро росло количество теоретических диссертационных исследований, посвященных языку сети Интернет. Эти работы позволили лучше понять особенности представления информации и выявить модели поведения человека в гиперпространстве.

Современный этап развития автоматического извлечения мнений, начавшийся примерно в 2009 году, характеризуется следующими основными тенденциями. Значительно увеличилось число ресурсов — появились в свободном доступе размеченные корпусы текстов, появились демонстрационные версии программ по автоматическому извлечению мнений. Резко выросло число языков, для которых имеются системы автоматического извлечения мнений — к английскому добавились испанский, французский, немецкий, нидерландский, итальянский, русский и ряд других. Продолжается совершенствование методов машинного обучения. Началось переосмысление самой задачи автоматического извлечения мнений (например, [26]).

1.7.1 Краткий обзор некоторых программных продуктов

В данном подразделе мы рассмотрим несколько систем автоматического извлечения мнений, которые либо представляют собой наиболее стандартный подход к построению систем (система И.И. Четверкина), либо носят узкоспециализированный характер (системы ФСО и Роскомнадзора, системы предсказания поведения фондового рынка).

1.7.1.1 Система Opiner[11]

Система Opiner, появившаяся в конце 2013 года, реализует идеи И.И. Четверкина [174]. В системе реализован традиционный классификационный подход на уровне документов. В качестве метода машинного обучения при классификации используется метод опорных векторов. Недостатками данной системы являются 1) отсутствие уровня объекта при классификации текста, 2) гипотеза о том, что не только оценка документа есть сумма оценок каждого

[11] https://opiner.ru/about

объекта, но и оценка каждого объекта соответствует общей оценке документа. Таким образом, данный подход имеет значительную ограниченность в своем применении. В первую очередь, применяемый метод позволяет точно оценивать «однообъектные» тексты, например, краткие отзывы о продуктах.

1.7.1.2 Система ФСО

В Федеральной службе охраны в 2011—2014 гг. разрабатывалась система, позволяющая автоматически определять отношение граждан России к законодательным инициативам [167]. С теоретической стороны представляет интерес структурная модель мнения, используемая в данной системе:

$$M = \langle W, F, V, O, \{A_i\}, D, C, S, T, Q, E, U \rangle, \quad (1)$$

где W — текст высказывания; F — факт, относительно которого высказывают мнение; V — оценка факта; O — отношение к факту; $\{A_i\}$ — автор(ы) мнения; D — степень доверия мнению (обоснованность мнения; C — цель высказывания мнения; S — оценка стиля высказывания; T — время; Q — аспект, по которому высказывают мнение; E — интенсивность проявления мнения; U — оценка полезности мнения [167: 5].

Данная модель достаточно точно описывает потенциальную оценку пользователем текста законопроекта. С другой стороны, различия между «оценкой факта» и «отношением к факту» представляются настолько несущественными, что ими можно пренебречь. К сожалению, в тексте автореферата не приводятся расшифровки некоторых интуитивно непонятных терминов, например, «оценка полезности мнения».

В целом, описанная система учитывает структурные особенности анализируемого объекта и, по-видимому, позволяет решать поставленные перед ней задачи.

1.7.1.3 Система Роскомнадзора

Система онлайн-мониторинга СМИ Роскомнадзора предназначена для автоматической идентификации в тексте противоправного контента и его классификации в соответствии с положениями Уголовного кодекса Российской

Федерации [79]. Данная система разрабатывается в НИВЦ Администрации Президента РФ. Более подробно принципы построения подобных систем приводятся в Главе 2.

1.7.1.4 Системы предсказания поведения фондового рынка

Системы автоматического анализа рыночной ситуации в настоящее время находятся в экспериментальной разработке. Они рассматриваются как частный случай экспертных систем, в которых ключевым фактором выбора того или иного решения является анализ внешнего фона публикаций об определенном объекте (публичные персоны компаний, аналитические отчеты о макроэкономических показателях и т.п.). Словари оценочной лексики здесь играют вспомогательную роль при модуле сравнения числовых значений.

1.7.2 Концептуальный подход к автоматическому извлечению мнений

Принципиально новый подход к автоматическому извлечению мнений был предложен в книге Э. Кэмбриа 2012 года [6]. Данный подход получил название концептуального подхода.

Подход заключается в следующем: единицами оценки являются концепты (соотносимые с семантическими классами). При этом оценки концептов могут быть связаны только с определенным набором концептов и не влиять на оценку других. Ключевым компонентом подхода является база эмоциональных концептов SenticNet, в которой каждый концепт рассматривается с точки зрения нескольких эмоциональных шкал, при этом нейтральные отношения в данной базе знаний отсутствуют. Необходимо отметить, что синтаксический анализ применяется в качестве вспомогательного компонента при анализе концептов.

Недостатком данного подхода следует признать сложность в расширении базы знаний данными других языков.

1.7.3 Обзор наиболее значимых существующих ресурсов

В настоящее время одним из базовых направлений исследований в области извлечения мнений является создание специализированных ресурсов: это словари

оценочной лексики и оценочные корпусы текстов. Из словарей необходимо отметить разработки в рамках проекта OpeNER[12], метод автоматического построения словарей И.И. Четверкина и разработки в области оценочной лексики в Антверпене. Более подробно об этих словарях говорится в Главе 3. Особое место среди оценочных словарей занимает лексическая база данных SentiWordNet. Другие ресурсы, опирающиеся на лексическую базу WordNet, оперируют понятием «эмоция». Так как не существует общепринятого списка эмоций, а также принимая во внимание различия в объемах понятий «эмоция» и «оценка» мы не будем рассматривать данные ресурсы. В последнее время появилось несколько размеченных оценочных корпусов текстов на уровне объектов[13] в рамках проекта OpeNER и разработок в TU Darmstadt. Ключевыми отличиями между данными семействами корпусов являются стандарты разметок. В Дармштадтском проекте применяется технология stand-off разметки, которая является машинно-ориентированной и на современном этапе развития технологий извлечения данных из текста представляется устаревшей.

1.8 Изменения по сравнению с предыдущими этапами развития систем автоматического извлечения мнений

Существенным изменением по сравнению с предыдущим этапом развития автоматического извлечения мнений можно считать больший интерес к собственно лингвистическим аспектам проблемы. Это легче всего проследить по росту числа публикаций на темы модификации принципов разметки оценочных явлений в корпусах текстов. На данном этапе задача автоматического извлечения мнений получила дальнейшее развитие за счет включения в тематику исследований новых аспектов проблемы. В первую очередь, необходимо упомянуть активное использование технологий машинного перевода для решения как частных задач (например, автоматическое пополнение словаря системы), так и более общих (таких как многоязычное извлечение мнений). Подобное

[12] http://www.opener-project.eu/

[13] Оценочных корпусов на уровне документов существует достаточно много, но мы не считаем их пригодными для использования в системах, ориентированных на объекты.

проникновение автоматического извлечения мнений наблюдается и в другие сферы компьютерной лингвистики — информационный поиск, проверку стилистики, борьбу со спамом, автоматическое построение онтологий предметных областей и др.

С другой стороны, продолжает иметь место оторванность прикладных разработок в области автоматического извлечения мнений от теоретических работ по психолингвистике, лингвистике текста, когнитивной лингвистике и ряда других дисциплин, которые, по нашему мнению, могли бы помочь улучшить качество работы систем автоматического извлечения мнений.

Одной из наиболее актуальных задач, стоящих перед автоматическим извлечением мнений сегодня, следует признать формирование принципов обработки языка Интернет-коммуникации. Здесь традиционные методы нормализации входных данных приводят к потере значительной доли важной лингвистической информации. Попытки использовать в качестве обучающих корпусов текстов искаженные тексты пока не привели к успеху. Это можно объяснить спецификой искажений написания текстов в Интернете [70], где на большом количестве площадок общения существуют гласные или негласные правила искаженного написания, которые, чаще всего, являются уникальными.

1.9 Теоретические основания оценки

В данном разделе мы рассмотрим теоретические основания оценки в аспекте их формализованного представления в системе автоматического извлечения мнений.

1.9.1 Субъективная оценка и ее основания (ценности и традиции)

Субъективная оценка, как уже писалось выше, неоднократно становилась объектом различных исследований. Оценка человеком явлений и предметов внешнего мира может осуществляться либо через систему ценностей (например, иерархия ценностей Р. Шенка), либо через традиции, которые в рамках определенной культуры рассматриваются в качестве норм [94; 95]. Подобная

дробная классификация способствует детальному описанию каждого из оценочных явлений. Но в настоящее время учет всех возможных иерархий ценностей и/или норм не представляется возможным как по причине наличия пересечений между классификациями, так и в силу значительного замедления скорости обработки текста в подобном компьютерном программном комплексе.

1.9.2 Оценка фактов

Оценка фактической информации составляет одну из наиболее спорных областей автоматического извлечения мнений. Традиционный подход игнорирует факты, т. к. они не соответствуют идеям субъективности, лежащим в основе теории автоматического извлечения мнений. В то же время в целом ряде программных продуктов, в первую очередь, ориентированных на обработку текстов СМИ или финансовых отчетов, именно оценка фактической информации играет ключевую роль. Например, с точки зрения субъективности предложение «*Вася убил Петю*» является фактом. С другой стороны, в христианской культуре убийства осуждаются, таким образом, в предложении содержится некоторый негатив по отношению к субъекту действия. По отношению к объекту действия также определяется негатив (по другой шкале).

1.9.3 Точка зрения (автор текста, объект в тексте, читатель)

В тексте оценка (особенно фактов) может рассматриваться с различных сторон. Хотя в психолингвистике существует мнение об оценке как процессе восприятия субъектом внешнего мира [например, 147: 3], этот субъект может отличаться от читателя текста.[14] В работах по автоматическому извлечению мнений можно выявить три типа точек зрения: автора текста (или субъекта высказывания), объекта в тексте и читателя.

Точка зрения автора (если автор текста сам не является разработчиком программы) является приближенной к действительности, т. к. сознание автора скрыто от непосредственного наблюдения. Традиционно автору приписывается

[14] Отметим, что у компьютерной программы нет «собственного мнения», таким образом, модель мнения должна задаваться заранее (в виде алгоритма).

максимально обобщенная модель мира, описывающая среднестатистического носителя языкового сознания.

Рассмотрение текста с точки зрения объекта тесным образом связано с набором ценностей, учитываемым разработчиками систем автоматического извлечения мнений. Так, «*увольнение работника*» с точки зрения работника должно являться негативным событием.

Иногда разработчики прибегают к иерархии оценок. Например, общечеловеческие ценности превалируют над личными. Таким образом, достигается моделирование восприятия текста читателем. Подобная иерархия может быть проиллюстрирована следующим примером: «*Террорист был убит*». С точки зрения объекта — это предложение негативно, но с точки зрения рядового читателя (не родственника и не единомышленника объекта) данное предложение будет анализироваться как позитивное.

1.9.4 Темпоральный аспект оценки (возраст, точка отсчета, изменение позиции автора)

Темпоральный аспект оценки представляет собой сложную иерархическую структуру, которая включает в себя следующие параметры: возраст, точку отсчета и изменение позиции автора.

Параметр возраста с его увеличением ведет к ухудшению восприятия объекта (за исключением некоторых алкогольных напитков и произведений искусства). В качестве примеров можно привести негативные выражения «древняя рухлядь» (о старой мебели), «старый автомобиль», «допотопный компьютер». Восприятие объекта разными возрастными категориями в настоящее время не учитывается в известных нам исследованиях.

Точка отсчета накладывает ограничения на границу между нормой и ненормой. Так для текста 1990 года компьютер с оперативной памятью в 1 Гб будет оцениваться позитивно, а в настоящее время — скорее негативно («слабенький комп»).

Изменение позиции автора сообщения тесно связано с логико-временными связями типа «раньше-теперь», «было-стало». При этом для систем автоматического извлечения мнений интерес представляет именно самый поздний вывод об объекте.

1.9.5 Культурологический аспект (религия, культура, норма)

Оценка сильно зависит от мировосприятия человека, на которое влияют, прежде всего, религия и культура, через которые формируется норма. Доминирующим понятием в сфере оценки здесь является этическая норма, которая превалирует при автоматической интерпретации оценочных фактов. В отличие от эстетической нормы, которая в тексте в большинстве случаев выражается оценочными прилагательными, этическая нормативность эксплицитно не выражена.

Наибольший интерес в этической оценке вызывает явление ассоциации человеком себя с определенной этнической, социальной или религиозной общностью. Данное явление нашло свое отражение, например, в восприятии православным населением т. н. «дела Pussy Riot», населением арабских стран появления фильма «Невинность мусульман» и т.п. Анализ текстов с подобными ситуациями показывает, что ассоциация может вызвать одну из следующих реакций: 1) дистанцирование от группы, 2) частичное дистанцирование от группы, 3) полная ассоциация себя с группой. Первый тип реакции отражен в примерах типа *«Всякая продажная криминальная шваль имеет или фонд поддержки идиоток из Pussy Riot или еще какую-нибудь липовую партию,типа «Свободной России»*[15]*.»*. Ко второму типу можно отнести предложение «*есть несколько человек, которые без оговорок или с теми или иными оговорками - но одобряют преследование Пусси Райот. <…> среди этих людей нет ни одного православного*[16]». Пример «*Все честные и свободные люди должны возвысить*

[15] http://irek-murtazin.livejournal.com/1297976.html?thread=54227768
[16] http://taki-net.livejournal.com/1283868.html

свой голос за свободу Pussy Riot[17].» демонстрирует проявление третьего типа реакции.

1.10 Терминологический аспект автоматического извлечения мнений

Автоматическое извлечение мнений развивается на стыке трех смежных областей знания (искусственный интеллект, информационный поиск и автоматическая обработка естественного языка).

Оно возникло в рамках исследований по искусственному интеллекту под названием «*анализ субъективности»*, и в конце 1990-х гг. стало рассматриваться как частный случай информационного поиска или как раздел автоматической обработки естественного языка. В информационном поиске это направление получило название «*автоматическое извлечение мнений»* (*opinion mining*), а в рамках автоматической обработки естественного языка — «*анализ чувств*» (*sentiment analysis*). Компонентный анализ терминологического словосочетания *sentiment analysis* показывает, что термин включает в себя слово со значительной оценочной составляющей, при этом происходит частичная «десемантизация» оценочного слова [158: 127-128], характерная для англоязычной терминологии.

«Автоматическое извлечение мнений» можно отнести к подразделу одного из базовых направлений исследования искусственного интеллекта — моделированию эмоций (*affective computing*). При этом в процессе автоматического извлечения мнений можно выделить несколько этапов анализа, и несколько относительно независимых подзадач. Среди этих этапов: 1) определение типа оценки (положительная, отрицательная, нейтральная), 2) определение объекта оценки, 3) определение субъекта оценки, 4) определение причины оценки, 5) определение силы (степени) оценки, 6) определение времени (даты) оценки. К подзадачам относятся: 1) анализ отзывов пользователей сети Интернет на продукты, события, явления, 2) анализ поведения в социальных сетях (направленное, например, на предотвращение террористической деятельности), 3) обнаружение в сети Интернет материалов, противоречащих определенным

[17] http://valery-kichin.livejournal.com/258940.html?thread=2769788&

законам, 4) бизнес-разведка, 5) автоматическое определение и автокоррекция стиля, 6) автоматическое определение пола автора (как один из компонентов).

Если рассматривать различия в терминологии этой предметной области в английском языке, то каждый из трех терминов (*subjectivity analysis*, *opinion mining*, *sentiment analysis*) полностью соответствует критериям однозначности, системности, мотивированности, понятийной ориентации, лингвистической правильности, точности, внедренности и языковой ориентации в рамках своей изначальной области знания [162: 129-133], и не несет оценочной нагрузки [87: 40-41]. Но при попытке выбрать наиболее общий термин для названия области в целом возникает несколько проблем. Термин *subjectivity analysis*, восходящий к позднему структурализму, полностью отражает свои задачи — разграничение объективной (неэмоциональной) и субъективной (эмоциональной) сторон речевой деятельности. Однако логико-лингвистический анализ данного термина показывает, что он не соответствует критериям внедренности и точности при расширении значения на предметную область в целом. Компонентный анализ термина *sentiment analysis* демонстрирует, что рассматриваемый термин соответствует своему назначению и полностью соответствует лингвистической структуре родственных ему терминов, таких как морфологический анализ, синтаксический анализ, анализ риторической структуры. Основанием для выбора данного термина в качестве гиперонима служит его внедренность в научный дискурс специалистов по компьютерной лингвистике. Существенным недостатком термина *sentiment analysis* следует признать неполное покрытие описываемой предметной области [26: 14]. Наиболее удачным на наш взгляд является термин *opinion mining*. По своей деривационной структуре он совпадает с другими общепризнанными терминами компьютерной лингвистики (*text mining*, *data mining*), что способствует указанию на точное место данного направления в рамках компьютерной лингвистики. Логико-лингвистический анализ позволяет выявить значимый недостаток термина — его более низкая употребительность по сравнению с *sentiment analysis* в одинаковых контекстах.

В отечественной компьютерной лингвистике у технологии автоматического извлечения мнений существует множество альтернативных названий. В работах начального этапа развития технологии [93] терминологическим проблемам не уделялось должного внимания. Один из наиболее частотных терминов — *анализ тональности* — возник из устойчивого оборота «в позитивной/негативной тональности», который отражает особенности звучащей речи. Также термин тональность в подобном значении употребляется в стилистике [142]. В пользу термина говорит его высокая употребительность[18]. Необходимо отметить, что даже авторы, употребляющие термин «анализ тональности», пытаются найти более адекватные эквиваленты. К подобным терминам можно отнести «сантимент-анализ» А.Н. Соловьева, «извлечение оценочных слов» и «анализ мнений» И.И. Четверкина [174: 1, 19-20], «анализ мнений» и «определение тональности» П.Ю. Полякова [156: 51]. Широкое распространение получил и калькированный термин «сентимент-анализ».

Рассмотрим указанные термины и выявим потенциальный гипероним указанной предметной области. Наиболее распространенный на текущий момент термин «анализ тональности» (так же, как и его вариант «определение тональности») не соответствует критериям однозначности и понятийной ориентации. Понятие «тональность» употребляется в сходном значении в стилистике, однако для компьютерной лингвистики, занимающейся в том числе анализом звучащей речи, такая омонимия между лингвистическим (в т.ч. грамматическим) понятием «тон»/«тональность» нежелательна, несмотря на высокую частотность данного термина в научном дискурсе. Термин «сентимент-анализ» и его вариант «сантимент-анализ», представляя собой транслитерацию терминов английского и французского языков соответственно, не могут претендовать на право стать гиперонимом предметной области вследствие значительных эмоциональных коннотаций, которые вскрываются посредством компонентного анализа данных терминов. Термин «извлечение оценочных слов» — одновременно обозначает один из этапов рассматриваемой задачи и,

[18] См., например, [108], где название одного из разделов содержит термин «анализ тональности».

следовательно, не может считаться наиболее общим. Логико-лингвистический анализ терминов *анализ мнений*[19] и *автоматическое извлечение мнений* показывает, что термин «анализ мнений» уступает термину «автоматическое извлечение мнений» по нескольким причинам. Во-первых, из названия «*анализ мнения*» неявна подзадача идентификации мнения в тексте. Во-вторых, теряется связь с исконным англоязычным наименованием (*opinion mining*). Кроме того, в традиции отечественной компьютерной лингвистики термином *анализ* принято обозначать отдельный этап обработки текста (например, *морфологический анализ*, *синтаксический анализ*), а не комплексную задачу.

Таким образом, в качестве наиболее общего термина для комплексной области нами используется термин «*автоматическое извлечение мнений*».

[19] К данному термину примыкает и термин Е.Г. Бруновой «*контент-анализ мнений*» [78].

Выводы к Главе 1

1. Автоматическое извлечение мнений зародилось в рамках работ по искусственному интеллекту. Первые разработки, подобно большинству работ в области искусственного интеллекта, носили модельный характер и не претендовали на полноту. Наиболее существенным фактором, повлиявшим на дальнейшее развитие автоматического извлечения мнений, стало появление Интернета и внедрение в практику компьютерной лингвистики корпусных методов.

2. В современной лингвистике проблемы текстовой прагматики занимают одно из ведущих мест. Среди них необходимо выделить работы в области изучения семантики и прагматики оценки. Существующие классификации оценочных отношений описывают данную проблематику достаточно детально, однако для того, чтобы использовать их в системах автоматической обработки естественного языка, им не хватает формализованности.

3. История автоматического извлечения мнений демонстрирует наличие тенденции к переходу от простых методов классификации текстов к сложным, учитывающим лингвистические знания. Наибольший интерес в последнее время вызывают способы лексикографического описания оценочных слов и формальные модели, описывающие отношения между различными классами слов.

4. Отечественные разработки в области автоматического извлечения мнений развиваются как в русле традиционного прикладного языкознания, так и в русле популярных зарубежных исследований. Ключевым компонентом в большинстве отечественных систем, в отличие от зарубежных корпусо-ориентированных, является словарь оценочной лексики.

5. Механический перенос существующих зарубежных технологий, разработанных на материале английского языка, оказывается слабо применимым к текстам на русском языке. Для текстов на русском языке большое значение приобретает предварительная обработка текста, прежде всего, морфологическая.

6. Значительный интерес со стороны бизнеса и государства к проблеме автоматического определения отношения населения к продуктам, персоналиям или событиям привели к значительной дифференциации задач автоматического извлечения мнений. Данные системы существенно отличаются друг от друга как по реализованной модели оценочных отношений, так и по критериям определения объектов.

7. Терминология автоматического извлечения мнений остается слабо разработанным вопросом. Используемые в работах по проблематике термины не составляют единой терминосистемы, что приводит к путанице в терминах. Российская терминология автоматического извлечения мнений формируется без участия специалистов в области терминоведения. Таким образом, в отечественных работах используются термины, противоречащие базовым критериям нормативного терминообразования.

Глава 2. Структура лингвистического обеспечения в системах автоматического извлечения мнений

В главе рассматриваются вопросы организации лингвистического обеспечения в системах автоматического извлечения мнений. Лингвистическое обеспечение понимается нами как «совокупность средств и правил для формализации естественного языка» [90: 3]. Таким образом, под лингвистическим обеспечением мы понимаем и этап предобработки текста, и собственно лингвистический анализ текста. Помимо традиционных методов автоматического анализа текста рассматриваются методы фильтрации текстового контента. В заключительном разделе главы дается описание лингвистического обеспечения в различных системах автоматического извлечения мнений.

2.1. Предобработка текста

Предобработка входного текста является неотъемлемым этапом анализа любого текста на естественном языке. В данном разделе рассмотрены формальные процедуры извлечения текстовой информации из гипертекстового представления, ориентированные на использование метаинформации. Другим значимым этапом предобработки текста считается его классификация, рассматриваемая в подразделе 2.1.2. При этом методы классификации текстов понимаются как основа для дальнейшего лингвистического анализа естественно-языкового текста.

2.1.1. Извлечение текста из html-кода веб-страниц. Компонент анализа метаинформации[20]

В последнее десятилетие в компьютерной лингвистике наметился переход от традиционных узкоспециализированных систем к гибридным системам, ориентированным на анализ текстов различных стилей и жанров. Для подобных систем характерны механизмы автоматической адаптации к типу текста (газетный, научный, личная онлайн-переписка), которая достигается, в основном,

[20] Параграф представляет собой расширенную версию публикаций автора [125, 126].

за счет использования средств автоматической классификации текстов при помощи машинного обучения.

Специалисты, работающие в сфере компьютерной лингвистики, обращают внимание на недостаточное использование теоретико-лингвистических знаний в современных, преимущественно статистических, системах анализа текста [26: 134]. Одной из областей автоматической обработки текста, где учет совокупности лингвистических и экстралингвистических факторов играет ключевую роль, является автоматическое извлечение мнений. В рамках данной задачи традиционные подходы, основанные на простом удалении тегов из тела веб-документов [например, 176], ведут к потере большого объема важной информации.

Рассмотрим одну из возможных методик извлечения значимой метаинформации из html-разметки веб-страницы на примере лексикализованных причинно-следственных связей (подробнее п. 2.4.1.4.1). В сфере отзывов (о фильмах, отелях, машинах, и т.д.) причинно-следственные связи часто бывают лексикализованы. Например, предложение «Мне не понравилось» становится «Недостатки». Две части предложения связываются через глаголы называния или существования, кроме того, наиболее частотным случаем является замена глагола на знак препинания, такой как двоеточие или тире. Например, *Достоинства: отель красивый,большая территория. можно весь отпуск не выходить за пределы. Недостатки: обслуживание.*[21] Подобная «формализованная» запись способствует более быстрому усваиванию информации человеком и нацелена на человеческое восприятие текстовой информации.

С точки зрения систем автоматической обработки текстов подобные предложения требуют извлечения значительного объема экстралингвистической информации (например, недостатки чего анализируются). В подобных случаях возникает сложность не обнаружения оценки (которая в данном случае легко решается с помощью простейшего словарного метода), а установления межобъектных связей, особенно причинно-следственного типа. Наиболее

[21] http://www.otzyv.ru/hotel/turkey/Club_Hotel_Sera/

простым способом установления объектов для оценок в текстах с лексикализованными причинно-следственными связями является извлечение метаинформации из html-кода веб-страниц. Рассмотрим процедуру анализа метаинформации для отзывов об отелях на сайте http://www.otzyv.ru/. На первом этапе проводится идентификация фрагмента веб-страницы, содержащего текст отзыва, в нашем случае это блок . Для извлечения названия объекта, к которому относится отзыв, можно использовать запрос на языке запросов XPath, например: */html/body/table/tbody/tr[110]/td[2]/text()[7]*. Для анализа формализованной части отзыва (представленной в виде таблицы) можно применить следующие действия. Поиск ведется внутри блока . При этом игнорируются служебные теги, например, <noindex>, которые предназначаются исключительно для визуализации результатов и не несут текстовой информации. Внутри данного тега может находиться служебная информация, например, пользовательская оценка формальных характеристик отеля (расположение, обслуживание и др.) в табличном представлении. Собственно резюме отзыва, содержащее лексикализованную оценку-следствие, находится внутри первого параграфа <p> указанного html-блока. Положительные (*Достоинства:*) и отрицательные (*Недостатки:*) фрагменты разделены двойным переносом строки
. Лексикализованная оценка-следствие отделена от причины однократным переносом строки. При генерации итогового результата анализа система автоматического извлечения мнений должна приписывать объекту, определяемому при помощи запроса XPath, положительные оценки для каждой из причин из второй части первой группы предложений (сами слова «достоинства» и «недостатки» не должны выводиться в конечный результат) и отрицательные для второй группы. Причинами в данном случае будут являться предложения, описывающие достоинства или недостатки.

Ключевым недостатком данного метода является учет особенностей представления информации на конкретном сайте. Также необходимо отметить, что для получения наибольшего эффекта от использования лингвистических знаний необходимо преобразовывать максимально возможный объем

экстралингвистической информации, содержащейся в html-коде веб-страниц в текстовый формат.

На этапе преобразования форматов представления целесообразно проводить орфографическую коррекцию текста. Существует несколько подходов к орфографической коррекции. Наиболее известный подход использует словарь и набор допустимых буквосочетаний, которые получаются на основе данного словаря. При проверке «допустимости» анализируемого слова используются т.н. «меры близости», которые опираются на расстояние Левенштейна. Реализации словарного способа орфокоррекции широко доступны в виде готовых программ, например *hunspell*. Недостатком этого подхода является невысокая полнота и сложность учета различных нововведений в языке.

В качестве альтернативного подхода, который позволяет учесть особенности Интернет-коммуникации, можно использовать словарь замен. В подобном словаре содержатся наиболее частые орфографические ошибки (пропуски букв в некоторых окончаниях, например, «орфографичский», использование похожих букв другого алфавита, например, латинские буквы в русских словах собака, русские буквы в английских словах desktop). При этом сленговые слова и выражения рассматриваются как нормальные русские слова. Это обусловлено неодинаковой оценкой слов «аффтар-афтар» и «автор».

2.1.2. Компонент классификации текста

В данном подразделе речь идет о требованиях к лингвистическому обеспечению в системах автоматического извлечения мнений общего назначения. Отличия данных систем от специализированных рассматриваются ниже в подразделе 2.4.1.

2.1.2.1. Теория подъязыков и ее прикладное значение[22]

Одной из ключевых задач корпусной лингвистики принято считать создание сбалансированных репрезентативных корпусов данных. В ходе многолетних

[22] Параграф представляет собой расширенную версию работы автора [127].

исследований была выработана методика создания сбалансированных корпусов текстов. В основу данной методики легла теория о функционально-стилевой дифференциации языка. Дальнейшим развитием теории функциональных стилей стала теория подъязыков, получившая распространение в машинном переводе [139], терминоведении и терминографии [85] и информационном поиске [103].

Сущность теории подъязыков заключается в том, что язык представляет собой суперсистему, включающую в себя ряд самодостаточных в пределах сферы своего применения подсистем. Данные подсистемы характеризуются особым лексическим составом, достаточно четко определенным набором синтаксических конструкций и др. Исключение составляет разговорная разновидность языка, в которой также можно выделить ряд слабо связанных между собой систем более низкого уровня, относительно стабильных в области своего функционирования и набора синтаксических средств для выражения тех или иных речевых ситуаций.

Принято считать, что пионером в применении теории подъязыков в отечественной компьютерной лингвистике является Н.Д. Андреев[23] [65: 118]. Наиболее подробное описание теоретической составляющей данной теории содержится в монографии Б.Ю. Городецкого и В.В. Раскина «Методы семантического исследования ограниченного подъязыка» [89]. В названной работе подъязык (в терминологии авторов — «языковая подсистема») понимается как «часть естественного языка <...>, которая и сама используется как язык определенной группой людей в определенных целях» [89: 13]. Авторы перечисляют компоненты подъязыка: множество фигур плана выражения, словарь, набор грамматических конструкций, идеальная совокупность текстов, структура и субстанция содержания [89: 18]. Мы не согласны с невключением в понятие «подъязык» искусственных языковых подсистем, используемых только в письменной коммуникации. Основным возражением является отсутствие явных противопоставлений между современными письменными языками электронной коммуникации (языки SMS и ICQ, чатов, форумов) и повседневным языком

[23] Подробное описание взглядов Н.Д. Андреева на теорию подъязыков содержится в монографии «Статистико-комбинаторные методы в теоретическом и прикладном языкознании» [66].

общения молодежи[24]. Дальнейшее развитие теория подъязыков получила в диссертационном исследовании В.М. Зелко [99]. В этой работе также анализируется ограниченный подъязык. Наиболее значимым, на наш взгляд, является предложенная В.М Зелко иерархическая структура подъязыка. В рамках подъязыка выделяются микроподъязыки, макроидиолект и идиолект [99: 10].

С появлением сети Интернет проблема обнаружения того или иного подъязыкового массива стала одной из наиболее актуальных для т.н. «семантического» поиска. Классические методы классификации [рубрикации] текстов, такие как метод tf*idf, классификаторы, основанные на методе Байеса или методе опорных векторов, дают приемлемый результат, однако требуют значительной коррекции с течением времени. Это, в первую очередь, объясняется нелинейным изменением соотношения актуальности тем. Другим, сугубо техническим, методом является метаклассификация информационного массива. При таком подходе анализируется не лексика документов, а их метаописания — вэб-адрес, набор тэгов (или хэш-тэгов для ресурсов, подобных Твиттеру). В последнее время в практике разработчиков информационных систем стали применяться комбинированные методы, повышающие как точность поисковой выдачи, так и ее полноту.

Для улучшения работы систем автоматического извлечения мнений представляется целесообразным использовать экстралингвистические данные о входном тексте. В первую очередь, к подобного рода данным следует отнести информацию о подъязыке документа. Другим значимым аспектом является дата создания документа.

Рассмотрим более подробно необходимость привлечения экстралингвистических знаний. Подъязык, являясь субсистемой языка, одновременно является суперсистемой для единиц более низкого уровня. Важным представляется и тот факт, что единицы низших уровней могут пересекаться, т.е. входить в промежуточные состояния между различными подъязыками. Особенно это явление характерно для Интернет-коммуникации, где в рамках одного

[24] Различие заключается лишь в использовании своеобразной письменной формы языка.

гипертекстового массива могут сосуществовать различные функциональные стили, различные взаимоотношения между участниками одного обсуждения, различные модели поведения, а также автоматически сгенерированная информация (чат-боты, спам-сообщения и т.п.). Важно учитывать, что слова одинаковой эмоциональной окраски в различных подъязыках могут иметь различную силу оценки, а также различные референты. Например, слово «спам» несет однозначно негативную оценку и на форуме программистов, и на форуме футбольных болельщиков. Но при этом на форуме футбольных болельщиков фраза «Как же мне надоел спам!» будет иметь однозначно негативную оценку автором сообщения по отношению к 1) футбольному клуба «Спартак (Москва)» 2) его болельщикам, а на форуме программистов — только по отношению разработчиков некачественной поисковой системы (или почтового сервера). Таким образом, указанная фраза в разных подъязыках потребует соотнесения с различными референтами, и, следовательно, фраза с форума болельщиков не должна попасть в выдачу по запросу относительно качества почтового сервера компании X, а фраза с программистского форума не должна попасть в выдачу по запросу о готовящихся потасовках футбольных болельщиков.

Таким образом, для каждого подъязыка целесообразно использовать отдельную модель, описывающую оценочные отношения внутри данного подъязыка. С другой стороны, классы лексических единиц и правила сочетаемости (например, правила усиления оценочных выражений) должны оставаться неизменными. Изменению должно подвергаться только лексическое наполнение данных классов, в т.ч. с учетом семантики лексических единиц в рамках данного контекста.

Информация о дате создания документа (или комментария) важна по причине того, что язык неформальной Интернет-коммуникации существенно обновляется примерно раз в полтора года. Здесь следует сделать отступление в связи с тем, что некоторые люди предпочитают не менять стиль общения, несмотря на изменение общественной «моды». Впрочем, изменения обычно касаются не всей лексики, а лишь некоторой (пусть даже значительной) ее части,

при этом оценочные отношения (включая вкусы и предпочтения аудитории) практически не подвергаются радикальным переменам. Особую сложность для обработки составляют случаи, когда временной интервал между связанными сообщениями составляет более года. В таких ситуациях следует использовать более новую модель, т.к. вероятность использования при порождении высказывания старой модели значительно ниже. Здесь важно сделать отступление относительно общих принципов работы большинства систем автоматического извлечения мнений. В них анализируется текст в режиме «на лету», т.е. каждый новый документ, попадающий на вход системе обрабатывается, самое позднее, в течение нескольких часов после появления публикации на сайте. Таким образом, проблема взаимодействия различных временны*х* оценочных моделей возникает только при подключении к системе нового Интернет-источника, для которого необходимо проанализировать весь архив сайта.

Учет типа подъязыка и его временного интервала способствуют динамическому подключению той или иной оценочной модели при обработке текстового массива. Привлечение теории подъязыков для совершенствования качества действующих систем автоматического извлечения мнений будет способствовать более глубокому пониманию процессов функционирования оценочных явлений в различных областях межличностного общения.

2.1.2.2. Типология текстов в задачах автоматического извлечения мнений[25]

Проблема отбора исходного материала исследований неоднократно обсуждалась в литературе по прикладной лингвистике. Исследователи подходили к решению данной задачи по-разному: с позиций компьютерной лингвистики — Р.Г. Пиотровский, А.В. Зубов [100], Л.Л. Нелюбин [150], с позиции квантитативной лингвистики — Ю.А. Тулдава [168], с позиции корпусной лингвистики — А.Н. Баранов, В.А. Плунгян, T. McEnery и др. Они выделяют следующие критерии — репрезентативности, объема исходных данных, содержания, выборки материала, структуры представления и ряд других.

[25] В основе параграфа статья автора [123].

Теория и практика исследований по автоматическому извлечению мнений показывает, что в данной области возникает еще одна базовая проблема. Ее можно назвать проблемой разноформатного представления исходных данных [57].

Суть проблемы сводится к следующему. Материал для создания исходного корпуса обычно берется из Интернет-источников. В отличие от текстов научно-технического характера, имеющих, как правило, жесткую структуру представления информации,[26] Интернет-тексты даже на одном сайте могут иметь различную структуру и объем. Так, например, отзывы о фильмах могут варьироваться от одного слова (иногда даже иконического знака) до развернутого подробного рассказа. Рассказ, в свою очередь, может иметь несколько структурных видов: «общие сведения – сюжет – игра актеров – что понравилось – что не понравилось – рекомендация», «общие сведения – сюжет – что понравилось – рекомендация», «название – что понравилось – сравнение с другими фильмами – рекомендация», «название – рекомендация» и т.п. Отзыв может содержать или не содержать иконических элементов. Количество подобных элементов также может быть различным.

Для облегчения задачи определения качества оценки события используются материалы с сайтов с заранее выставленными оценками. На практике подобный подход не всегда приводит к требуемому результату. Это связано, в первую очередь, с тем, что для ряда задач данный подход неприемлем. В качестве примера можно привести программу классификации пользовательских отзывов в компании Microsoft.[27] В пользовательских отзывах используется своеобразная терминология — названия программных продуктов компании, которая не будет отражена в обучающем корпусе программы, т.к. линейка продуктов находится в процессе постоянного обновления. Другим минусом подобного подхода следует

---

[26] А.С. Герд выделяет несколько основных разделов научных статей, число которых может меняться в зависимости от национальной научной традиции и конкретного подъязыка [84]. К сходным выводам приходит также Дуглас Байбер [4], который отмечает уменьшение жесткости структуры для гуманитарных дисциплин (особенно истории и археологии).

[27] Программа должна проводить классификацию отзывов по трем группам — технические отзывы, положительные отзывы и отрицательные отзывы.

признать использование разнообразных шкал оценок на различных сайтах (3-х бальная, 5-ти бальная, «шкала с полузвездочками», шкала с положительными и отрицательными частями, и др.). Третьим минусом (с нашей точки зрения самым существенным) является несовпадение языковых и паралингвистических средств даже у одного пользователя при выставлении определенной оценки продукту или явлению. Так оценка «5» по пятибалльной шкале может соотноситься по своему выражению с оценкой «4», вынесенной тем же пользователем сходному продукту по той же шкале.

В литературе [40; 57] описаны следующие методы решения этой задачи: 1) приведение в единый формат, 2) унифицированная разметка. Методы приведения исходного корпуса в единый формат разнообразны, но могут быть сведены к следующему. Создается таблица соответствий между шкалами, и, исходя из этих соответствий, каждому документу приписывается оценка. Против данного подхода имеются существенные возражения [57]: 1) какую оценку следует считать положительной/ отрицательной? 2) как выделить нейтральный уровень оценки? 3) требуется ли создавать отдельную таблицу соответствий между сайтами? 4) требуется ли создавать таблицу соответствий между пользователями? 5) как эти соответствия определить? Поэтому, в последнее время, чаще всего применяется унифицированная составленная по единым принципам разметка обучающего массива (корпуса). При таком подходе, исходный материал подвергается значительной предварительной обработке: удаляются все внешние оценочные компоненты, документы приводятся в единый формат (кодировка, шрифт, размер и т.д.). Но у подобного подхода тоже есть свои существенные недостатки. Основным из которых является неучет изначальной авторской оценки документа, так некоторые изначально положительные отзывы классифицируются как отрицательные и (крайне редко) наоборот. Именно поэтому в системах подобного типа существенно возрастает роль оценочного словаря.

Типология текстов традиционно рассматривается в стилистике. В компьютерной лингвистике традиционно рассматривались только научно-технические тексты узкой тематики. Попытки ввести типологию текстов в

компьютерной лингвистике проводятся сравнительно недавно [139]. Данная типология не связана с тематической рубрикацией в информационно-поисковых системах, где в основу классификации положены различные веса ключевых слов. В типологии текстов для общего случая машинной обработки основными факторами являются однозначность значения термина (в случае автоматического извлечения мнений — однозначность оценки) и сходство синтаксических конструкций. Так тексты, имеющие различные наборы синтаксических конструкций, но одинаковые наборы терминов (или других типов слов), будут относиться к разным типам текстов. В качестве примера подобных текстов можно привести текст доклада на конференции по химии полимеров и текст патента, описывающего процесс производства данного полимера.

В зарубежной компьютерной лингвистике в настоящее время ненаучно-технические тексты разделяются на три типа текста: блоги, новости, твиты. При этом тексты блогов и твитов часто объединяются под наименованием «текстов социальных медиа». Несмотря на формальное объединение этих типов текстов в один на практике применяемые для анализа методы значительно различаются. Данную классификацию можно назвать формальной, т.к. она основывается на внешних признаках текстов, а именно области коммуникации. Следует отметить, что формальная классификация является первичной и должна дополняться вторичной, семантической.

Семантическая или тематическая классификация чрезвычайно важна при моделировании прагматики текста. Тексты можно классифицировать по целому ряду семантических признаков. Первым из них является классификация по объектам, например, банки, политические деятели, нефтегазовая отрасль. Вторым основанием, дополняющим первый, мы считаем пересечение выделенных рубрик. Дальнейшее деление может вестись в сторону уточнения (дробления) рубрик или выделения главных и вторичных рубрик по заданному объекту. Например, рубрика «высказывания политиков о банках, работающих в сфере кредитования нефтегазовой отрасли» выделяется при детализации рубрик, а при разграничении главных и вторичных рубрик — «банки» (основная рубрика) и «политические

деятели» и «нефтегазовая отрасль» (вторичные рубрики). На практике наиболее часто применяется пользовательская классификация текстов.

2.1.2.3. Автоматический выбор модели анализа

Под моделью анализа мы понимаем набор правил сочетаемости оценочных слов и оценочный словарь, которые наиболее точно описывают оценочные отношения конкретной предметной области или набора объектов.

Ручной выбор модели применяется для узкоспециализированных систем. Для систем общего типа необходим модуль автоматического выбора модели. Автоматический выбор модели анализа осуществляется за счет учета следующих факторов: 1) фактор подъязыка, 2) фактор модели мира, 3) фактор типа объекта. При этом наиболее значимым фактором является фактор модели мира.

При анализе определенного подъязыка достигается снятие оценочной полисемии, характерной для разных предметных областей. Определение подъязыка проводится при помощи статистических процедур, например *tf*idf*.

Учет различных моделей мира позволяет анализировать одно и то же предложение с разных точек зрения. Набор моделей мира задается экспертно и имеет два варианта реализации. При первом варианте модели мира связываются с определенным подъязыком. При идентификации определенного подъязыка из возможных моделей мира также выбираются только те, которые связаны с данным подъязыком. При втором — модель мира связана не с подъязыком, а с определенным типом текста, например отзывом или новостью. Анализ ведется по каждой из релевантных моделей. Результаты, видимые конечному пользователю, задаются во внешней конфигурации. Например, при выборе визуализируемой модели «Продавец топлива» предложение *С сегодняшнего дня в Казахстане подорожал бензин*[28] будет рассматриваться как позитивное, а при выборе модели «Автовладелец» — как негативное.

Разные типы объектов зачастую ведут себя по-разному в одинаковых контекстах. Например, субъект глагола «ненавидеть», выраженный персоной,

[28] http://khabar.kz/ru/news/ekonomika/item/24696-v-kazakhstane-podorozhal-benzin

нейтрален, а выраженный организацией — негативен по отношению и к субъекту, и к объекту глагола. Например, предложение «*Путин ненавидит США*.»[29] нейтрально относительно лексемы «*Путин*», в то время как «*Наше государство ненавидит детей на подсознательном уровне.*»[30] негативно относительно лексемы «государство».

2.2. Лингвистический анализ текста[31]

Методы автоматического лингвистического анализа были разработаны для традиционных направлений компьютерной лингвистики, таких как машинный перевод, информационный поиск, автоматическое аннотирование и реферирование, компьютерная лексикография, вопросно-ответные системы, исправление искаженно написанного текста, распознавание и синтез речи [Carstensen et al., 2010: 553–554]. Однако, за последние 10–15 лет возник ряд новых направлений, например, обнаружение спама, автоматическое извлечение мнений, кластеризация вэб-данных, для эффективного функционирования которых в Интернете необходимо по-новому взглянуть на базовые процедуры первичной обработки текста.

К базовым методам автоматической обработки входного текста относятся: токенизация (графематический анализ [132: 56]), морфологический анализ, поверхностный синтаксический анализ и локальный семантический анализ.

Методы лингвистического анализа текстов можно разделить на поверхностные и глубинные. Поверхностные методы лингвистического анализа [132; 106] текста включают в себя предварительную обработку текста (токенизацию, морфологический анализ, постморфологический анализ) и модуль извлечения первичной семантической информации. Глубинные методы лингвистического анализа текста — это синтаксический (поверхностный и глубинный) и вторичный семантический анализ. Дискурсивный анализ текста не является обязательным на современном этапе развития компьютерной

---

[29] http://mikle1.livejournal.com/3385051.html

[30] http://vk.com/wall-40684908_255150

[31] Рассматриваются методы автоматического лингвистического анализа.

лингвистики, особенно при работе с текстами на русском языке. Он заключается в разрешении анафорических ссылок, установлении кореферентных последовательностей, а также восстановлении эллиптических конструкций, оставшихся после синтаксического анализа.

2.2.1. Графематический анализ и токенизация

На этапе токенизации (графематического анализа) проводится первичный анализ текста исходного документа: определяются границы заголовков, абзацев и отдельных предложений [132, 106]. Из предложений выделяются отдельные слова — *словоформы*. При этом должна учитываться возможность использования разделителей в аббревиатурах и названиях, например в предложении:

В ВТБ и "Э.ОН Россия" отказались от комментариев.[32]

Ключевую роль графематический анализ играет при обработке блогосферы и текстов социальных сетей, где традиционные методы, ориентированные на наличие знаков препинания, не помогают точно определять границы предложений. На ряду с графематическим анализом на этапе первичной лингвистической обработки текста проводится лексический анализ, направленный на идентификацию семантически или синтаксически значимых последовательностей. К этим последовательностям относятся, в первую очередь, адреса электронной почты (e-mail), адреса веб-сайтов (URL, URI), границы прямой речи, а также фразеологические единицы.

2.2.2. Морфологический анализ

Основные принципы автоматического морфологического анализа текстов на русском языке были сформулированы в начале 1960-х гг. в работах специалистов по машинному переводу и информационному поиску, особое место среди этих работ занимают исследования Г.Г. Белоногова [73, 72]. По определению О.В. Минтусовой «морфологический анализ <...> понимается как обработка словоформ с целью получения информации, отображающей в явном и

[32] http://www.kommersant.ru/doc/2799888

удобном виде те их свойства, которые необходимы для последующего синтаксического анализа» [148: 3].

При морфологическом анализе для каждой словоформы строится множество вероятных *лемм* с морфологическими атрибутами (к морфологическим атрибутам сущности относятся: падеж, число, род, одушевленность). Каждая лемма представляет собой *нормальную* форму слова (именительный падеж, единственное число). В случае отсутствия формы единственного числа (грамматический класс *Pluralia tantum*, к которому относятся, например, названия некоторых государств (Соединенные Штаты Америки, Багамские острова)) нормальной формой является именительный падеж множественного числа. Определение нормальной формы словоформы проводится, чаще всего, на основе заранее подготовленного морфологического словаря. Для русского языка наиболее продуктивным считается использование словаря основ [например, 74: 54; 154: 32-34[33]], формат которого представляет собой набор лексических входов (каждый лексический вход обозначает один словоизменительный тип), состоящий из основы (или квазиосновы) слова и списка окончаний с набором грамматических атрибутов. В случае отсутствия словоформы в словаре (в случае со словарем основ — отсутствие подходящего сочетания основы слова и окончания), либо отсутствия самого словаря, для определения постоянной части слова используется *стемминг* (англ. *stem* — основа*).* При наличии словаря основ задача стемминга сводится к нахождению окончания слова и приписыванию словоформе набора характеристик наиболее вероятного словоизменительного типа. Для этого, по правилам, определённым для данного языка, определяется изменяемая часть слова и заменяется усечением, допускающим любые символы. Преимуществом стемминга является отсутствие необходимости в полном морфологическом словаре, простота реализации и скорость работы, недостатком — невысокая точность. Все результаты морфологического анализа могут быть представлены в виде XML-разметки исходного текста.

---

[33] С другой стороны, ср. идею Г.Г. Белоногова о внесении в словарь морфологического анализа словоформ [154, 34-35] и целых словосочетаний [72].

В модуле морфологического анализа текста для нужд автоматического извлечения мнений необходимо использовать особую организацию частеречной информации, которая отличается от традиционной, принятой в системах машинного перевода или информационного поиска. Так для нужд информационно-поисковых систем целесообразно иметь минимальный набор исходных частей речи. Это обусловлено требованием полноты выдачи на поисковый запрос, который автоматически расширяется формами слов из запроса. В некоторых случаях это приводит к появлению 188 форм у глагола в русском языке [например, 74: 44]. Для задач автоматического извлечения мнений, напротив, требуется более дробная классификация, при которой видовые пары глаголов являются различными леммами, наряду с причастиями и деепричастиями. В качестве причин подобного разделения можно назвать различное поведение глаголов и причастий в тексте.

Теоретические проблемы автоматического морфологического анализа применительно к ряду подзадач автоматического извлечения мнений рассматриваются в п. 2.4.1.3.5.

Отличия в оптимальном представлении морфологической информации для разных подзадач в рамках системы семантического поиска[34] подтверждают тезис С.А. Коваля о сложности достижения «единства представлений морфологии для многофункциональных приложений» [107: 109-110].

На этапе *постморфологического анализа* снимается межчастеречная омонимия (например, словоформа «*мыла*» может разбираться как *глагол прошедшего времени женского рода* или как *существительное в родительном падеже единственного числа* или *именительном или винительном падежах множественного числа*), а также приписывается лексико-семантическая информация (при наличии в системе автоматической обработки текста словаря семантической информации).

2.2.3. Синтаксический анализ

[34] В которую входят компоненты морфологической индексации текста, расширения запроса формами слов и модуль автоматического извлечения мнений.

Целью синтаксического анализа текста является обнаружение синтаксической структуры текста. В большинстве задач современной компьютерной лингвистики синтаксический анализ состоит из двух этапов — поверхностного, направленного на определение границ именных и глагольных групп, и глубинного, ориентированного на установление связей между именными группами.

2.2.3.1. Поверхностный синтаксический анализ

Одной из наиболее простых процедур поверхностного синтаксического анализа, также называемого чанкингом (от chunking — сегментирование), считается определение предложных групп. В совокупности с результатами семантического анализа анализ предложных групп способствует сокращению числа потенциальных объектов, подлежащих дальнейшему анализу.

К этапу поверхностного синтаксического анализа относятся также процедуры определения числовых конструкций (например, *10 лет*) и вводных конструкций.

В большинстве случаев результатов поверхностного синтаксического анализа достаточно для эффективной работы системы автоматического извлечения мнений. При этом работа глубинно-синтаксического анализа заменяется набором правил, работающих с набором именных и глагольных групп.

2.2.3.2. Глубинный синтаксический анализ

Глубинный синтаксический анализ направлен на выявление связей между именными и глагольными группами, а также между клаузами и несвободными оборотами. Для русского языка наибольшее распространение получило представление результатов синтаксического анализа в виде деревьев зависимостей.

Ключевым компонентом автоматического синтаксического анализа является словарь моделей управления. Традиционно подобный словарь создается для глаголов. На практике возникает необходимость также и в словаре

управления для отглагольных имен, а также правила преобразования моделей управления для причастий.

Для автоматического извлечения мнений важным является не только идентификация зависимости элементов друг от друга, но и их семантическое наполнение. Учитывая текущее состояние синтаксического анализа текста на русском языке, восходящий синтаксический анализ лучше подходит для работы с нестандартными орфографическими и грамматическими вариантами. Несмотря на более низкую точность работы, восходящий анализ обеспечивает существенно большее покрытие синтаксических конструкций за счет связывания минимальных единиц.

2.2.4. Семантический анализ текста

Семантический анализ используется для выделения в тексте семантических единиц, т.н. *сущностей*. Сущности подразделяются на именованные и неименованные. Неименованными сущностями являются слова-идентификаторы классов, например, профессий (*депутат*, *учитель*) и продуктов питания (*йогурт*, *салат*). К именованным сущностям относятся имена собственные, которые могут включать в себя слова-идентификаторы класса, например, *йогурт Danone*, *Газпром*, *Игорь Петрович Волков*. Неименованные сущности присваиваются леммам на основе информации из семантических словарей систем автоматической обработки текста. Для идентификации именованных сущностей применяется ряд методов, наиболее распространенными среди которых следует признать машинное обучение с учителем, онтологии и шаблоны [например, 39: 26].

Первичный семантический анализ проводится при помощи семантического словаря и, реже, онтологий. Он проводится параллельно с морфологическим анализом. Его результат используется при вторичном семантическом анализе, который учитывает синтаксическую структуру предложения.

Вторичный семантический анализ проводится с целью идентификации в тексте комплексных именованных сущностей. К комплексным сущностям

относятся именованные сущности, обозначающие названия законов, предметов интеллектуальной собственности (например, книги и фильмы), названия некоторых событий (например, выставки, конференции и форумы) и ряд других.

Результатом семантического анализа текста чаще всего является семантическая сеть. Впрочем, для задач автоматического извлечения мнений достаточным является идентификация сущностей верхнего уровня (которые могут иметь вложенные сущности).

2.2.5. Дискурсивный анализ текста

Дискурсивный или суперсинтаксический анализ текста заключается в восстановлении эллиптических конструкций, разрешении анафорических ссылок и нахождении кореферентных цепей. Основной задачей дискурсивного анализа текста считается увеличение полноты выдачи в системе автоматической обработки естественного языка. С точки зрения автоматического извлечения мнений это значит, что сокращается количество конечных (видимых пользователем) объектов анализа, а оценка объекта производится в максимальном числе его вхождений в текст.

На схеме (Рисунок 2) представлена упрощенная последовательность ключевых этапов подкомпонента дискурсивного анализа текста. Разрешение анафорических местоимений должно предшествовать этапу разрешения кореференции, что объясняется ограничениями на глубину поиска кореферентов в тексте. Указанный на схеме этап построения семантической сети не является обязательным для решения нашей задачи, поэтому он указан в скобках.

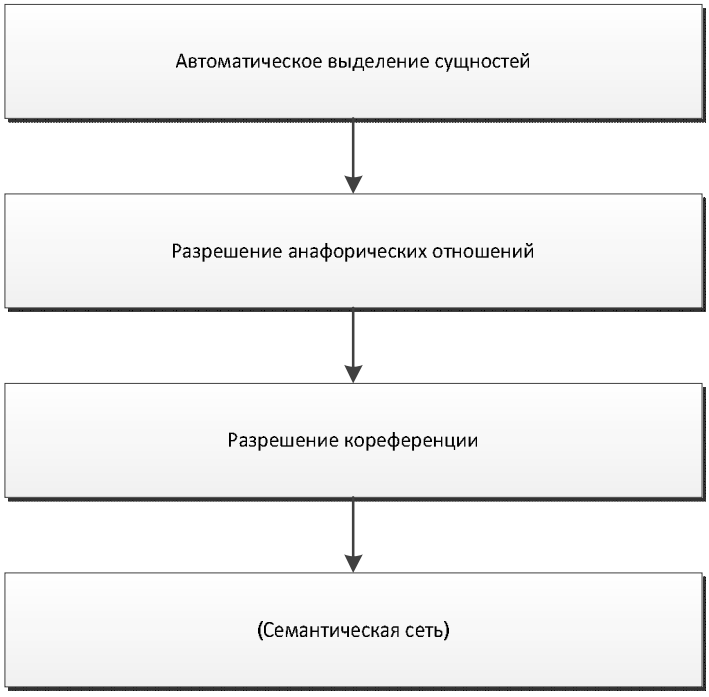

Рисунок 2. Упрощенная схема дискурсивного компонента

Практика исследований в области автоматического разрешения анафоры и кореференции [62] показывает, что каждую подзадачу целесообразно разбить на несколько этапов (Рисунок 3; Рисунок 4). Под разрешением скобочной катафоры (Рисунок 3) понимается нахождение антецедентов в предложениях вида *Anaphor-LeftBracket-Antecedent-RightBracket*, например *Если* ***он*** *(Путин) встает на сторону Асада, и Ирана, и "Хизбаллах", то у него будут серьезные проблемы с суннитскими странами региона…*[35]

[35] http://www.newsru.com/world/30sep2015/kerrysays.html

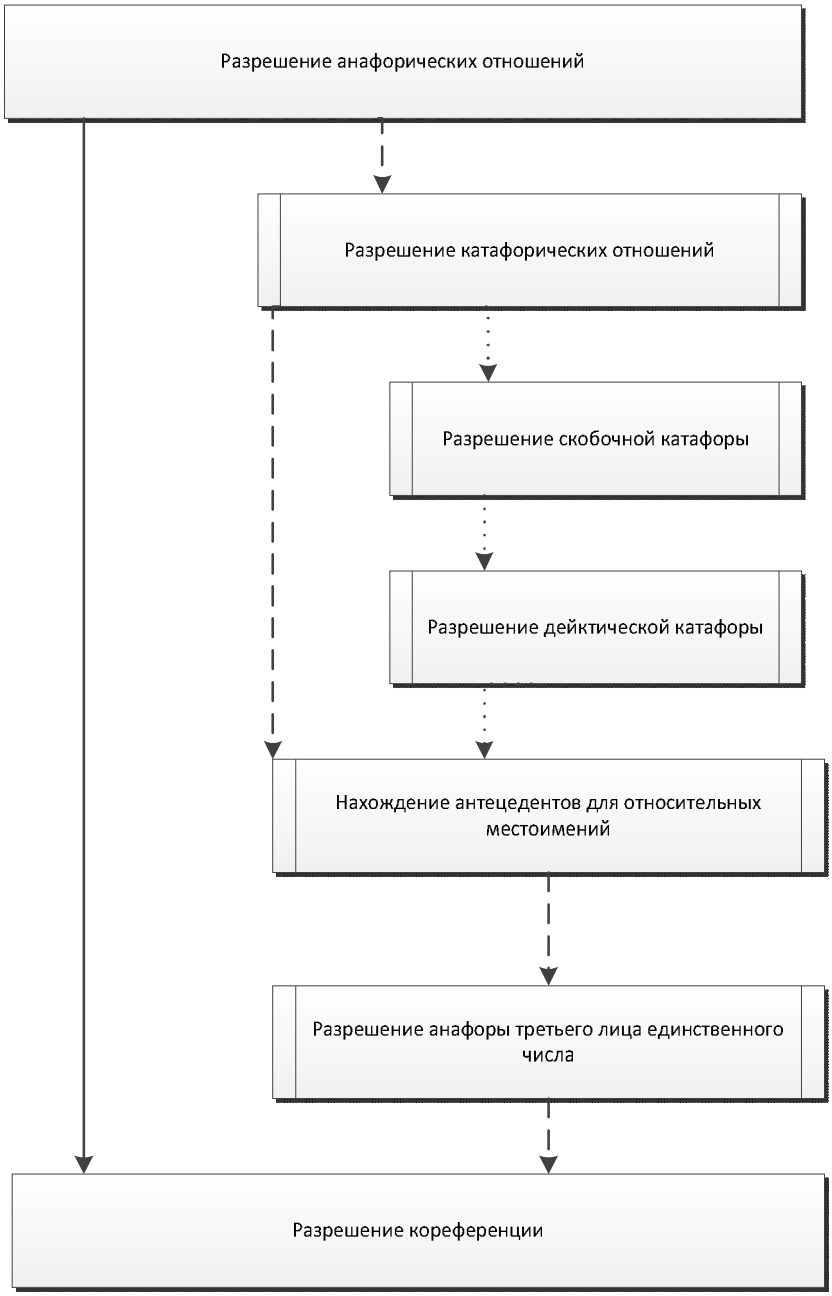

Рисунок 3. Этапы разрешения базовых анафорических отношений

Задача автоматического разрешения кореференции состоит из двух последовательных этапов (Рисунок 4). На первом этапе проводится разрешение *эндоклаузной кореференции*. Под разрешением эндоклаузных случаев кореференции мы понимаем две задачи: 1) связывание приложений (например, *Бывший начальник украинского Генштаба*, ***экс-первый замминистра обороны Украины Анатолий Лопата*** *заявил, что силы, воюющие на стороне ЛНР и ДНР,*

вскоре пойдут в наступление.[36]); 2) идентификация расшифровок аббревиатур (например, *ГПУ 26 мая начала уголовное производство по факту возможного участия должностных лиц Главного разведывательного управления* ***(ГРУ)*** *Генштаба России в "террористической деятельности" на востоке Украины.*[37]). На втором этапе проводится разрешение прочих случаев кореферентного употребления текстовых единиц. Т.к. кореферентные элементы в данных случаях находятся в разных клаузах, то подобное употребление можно назвать *экзоклаузным*.

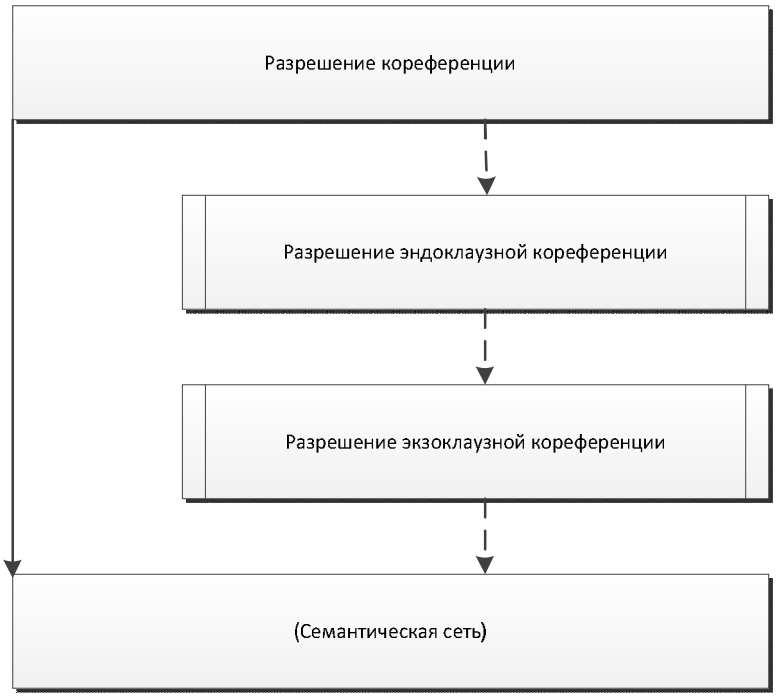

Рисунок 4. Этапы разрешения кореферентных отношений

В процессе разрешения анафоры для достижения наилучшего результата следует проводить отсев нереферентных употреблений анафоров и нереферентных слов [ср. 32]. А в качестве оптимального подхода к разрешению анафорических отношений нам представляется гибридный подход, в котором последовательно сочетаются высокоточные правила со статистическими методами, которые обладают большей полнотой.

[36] http://www.bbc.com/russian/russia/2015/08/150818_rus_press
[37] http://www.utro.ru/articles/2015/08/05/1251575.shtml

Выше описаны случаи, характерные для одного документа. Для того, чтобы построить целостную картину оценки объекта в рамках ветви обсуждения на каком-либо Интернет-ресурсе, часто требуется определять связь между оценочными элементами. Например, на форумах сообщения, несущие оценочную информацию, связаны одно с другим посредством как линейной связи (А согласен с Б, при том что сообщение Б следует за сообщением А), так и гипертекстовой (Ф согласен с А, при том что между данными сообщениями находится N других сообщений). Сообщения подобного рода не имеют четкой структуры в плане длины сообщения (если длина сообщения не ограничена правилами форума), количества паралингвистических компонентов, наличия или отсутствия ссылок на другие ресурсы.

Данный случай требует разработки особого модуля определения связей между участниками обсуждения на основе метаинформации в сочетании с моделью диалога, разработка которых выходит за рамки нашего диссертационного исследования.

2.3. Компонент фильтрации объектов

Фильтрация контента не является обязательным компонентом в любой системе автоматической обработки текста. Целью этого этапа обработки текста считается сокращение ложных срабатываний системы за счет исключения из дальнейшего анализа некоторых типов данных. Компонент фильтрации html-контента относится к этапу предобработки текста. Логическим продолжением данного модуля является вторичная фильтрация контента с целью сокращения количества потенциальных объектов последующего анализа. Отметим, что данный компонент фильтрации не является итоговым. В рамках определения субъектов высказывания также проводится сокращение количества объектов оценки. В данном разделе под «контентом» мы будем понимать именно лингвистическое наполнение документа, а не метаданные, которые к моменту начала работы рассматриваемого компонента уже преобразованы.

2.3.1. Роль компонента фильтрации контента в системе автоматического извлечения мнений

Основными задачами фильтрации контента следует признать повышение полноты и точности, а также ускорения обработки путем сокращения объема анализируемого материала. Данный этап тесно взаимосвязан с этапами выбора субъекта и объекта высказывания.

Наиболее частым случаем является задание субъектов и объектов мнения извне. При таком подходе достаточным оказывается идентифицировать все элементы текста и проводить последовательное сравнение множества элементов текста со множеством субъектов или объектов. Такой подход используется в целом ряде коммерческих продуктов, из наиболее известных — это продукты компании «Медиалогия» и система «Крибрум».

Ниже будут рассмотрены критерии, благодаря которым можно достичь качественной фильтрации «ненужного» контента, не требующие от пользователя предварительных знаний об интересующем явлении, событии или продукте.

2.3.2. Структурные критерии фильтрации контента

Структурные критерии являются наиболее продуктивными, т.к. не требуют детального анализа семантики и синтаксиса. Первый способ структурной фильтрации подразумевает сравнение текстовых элементов (единиц семантического анализа) с шаблонами допустимых синтаксических структур. Одной из них является одиночное слово. При идентификации одиночного слова происходит обращение к оценочному словарю системы, и, в случае отсутствия слова в словаре, фильтрация данного слова. При нахождении слова в словаре компонент структурной фильтрации не исключает слово[38] из списка потенциальных субъектов и объектов.

Условные придаточные предложения не содержат в себе оценку сверившегося или неизбежного действия, а описывают ситуацию, которая возможна лишь при одном из возможных сценариев развития ситуации. Таким

[38] Под словом здесь мы понимаем единицу семантического анализа в определенной позиции в тексте.

образом, данные контексты могут считаться априорно нейтральными, не требующими дальнейшего анализа.

Анализ вопросительных предложений с точки зрения выраженных в них оценок представляет собой сложную задачу. При структурном анализе текстов от вопросительных предложений отделяются сходные предложения, выражающие недоумение и негодование. Подобные предложения, чаще всего, содержат помимо вопросительного знака и другие символы. Предложения вида «*«Откуда он?» — спрашиваю я, указывая на флаг Новороссии.*[39]» или «*С. КОРЗУН: В чем предмет спора и разногласий был основной с Фишером?*[40]» не анализируются, в то время как предложения вида «*Какой отец?!*[41]» анализируются по особой модели.

2.3.3. Семантические критерии фильтрации контента

С точки зрения автоматического извлечения мнений каждая из именованных сущностей имеет единую оценку, которая не зависит от оценки каждого из компонентов, составляющих данную сущность. Так именованные сущности типов «Организации» и «Законы» включают в себя оценочные слова, например, Комитет по противодействию коррупции, Закон «О защите прав потребителей». Чаще всего оценка подобных сущностей нейтральна, и сами сущности выступают в роли объекта анализа. Здесь следует оговориться, что оценочно нейтральными являются только именованные части сущностей указанных типов. Таким образом, при наличии оценочных модификаторов оценка сущности изменится в соответствии с оценкой модификатора, например *Александр ЛИХАЧЕВ: «По-другому исполнить этот дурацкий* ***закон о капремонте*** *нельзя»*[42].

К другим семантическим типам, которые могут содержать в своем названии оценочные компоненты, но при этом быть нейтральными, относятся названия предметов интеллектуальной собственности (книги, фильмы, выставки) и события. Например, название фильма в следующем примере «*Особо*

---

39 http://www.novayagazeta.ru/politics/65753.html
40 http://echo.msk.ru/programs/korzun/1419892-echo/
41 http://www.novayagazeta.ru/politics/65753.html
42 http://www.omskinform.ru/news/88205

международное жюри отметило драму Юрия Быкова "Дурак", уже завоевавшую приз на кинофестивале в Локарно в августе.»[43] не несет никакой оценочной информации. Та же ситуация наблюдается в названии книги «Убийца», например, в предложении «*В новелле «Убийца» наш герой добирается до самого центра степей, в столицу Мараху, затем посещает диких горцев, чтобы заручиться поддержкой в случае нападения на его королевство воинственной империи Мардинан.*[44]». Неоценочным является следующее событие: *«Ответ СК РФ на моё приглашение: марш "Бабий бунт" 10.08.2013г. г. Москва.*[45]*»*. С другой стороны, при наличии оценочного модификатора перед названием сущности она так же получает оценку данного модификатора. В качестве примера можно привести следующий: «*Его главные произведения — это окопные дневники, а также, говоря по-простому, профашистская книга «Рабочий и гештальт*»[46]».

Фрагменты назывных предложений, выраженные оценочными словами, также подлежат фильтрации. Это связано с тем, что эти слова не имеют референтов в тексте. В качестве примера можно привести следующее предложение: «*Директор института Языкознания, Витя Виноградов, говорит, что Плунгян - гений*».[47] В данном предложении слово «*гений*» не имеет референта и поэтому не может являться объектом. При этом в препозиции к именованной части слово «гений» может быть референтным, например *Наш величайший гений Пушкин написал однажды: «Поэма никогда не стоит улыбки сладострастных губ»*[48]. В подобных случаях оценочные слова не подлежат фильтрации при определении объекта анализа.

2.4. Компонент автоматического извлечения мнений

Системы автоматического извлечения мнений можно подразделить на 3 основных класса систем: 1) общего назначения, 2) специализированные и 3) системы идентификации противоправного контента.

---

43 http://ria.ru/culture/20141018/1028968755.html#ixzz3GbppmAEj
44 http://books.imhonet.ru/element/1159123/
45 http://democrator.ru/complain/5726/5726/?page=2
46 http://www.odnako.org/blogs/chitat-3/
47 http://husainov.livejournal.com/292460.html?thread=2444652#t2444652
48 http://litrossia.ru/archive/112/writer/2632.php

2.4.1. Система автоматического извлечения мнений общего назначения

Системы общего назначения предназначены для анализа оценочной информации в любых текстах. Обработка сложных случаев оценки достигается при помощи подключения различных моделей анализа мнений.

2.4.1.1. Общая структура лингвистического обеспечения (ЛО) системы

Лингвистическое обеспечение систем автоматического извлечения мнений состоит из нескольких модулей (см. Рисунок 5). Модули работают в строго определенной последовательности, при этом результат работы предыдущего модуля становится входной информацией для последующего модуля.

Первым модулем обработки является **модуль фильтрации контента**. Его задачей является преобразование текста с целью получения первичной информации о тексте (из html-кода обрабатываемого документа). Данный модуль передает информацию и модулю лингвистической обработки, и модулю автоматического извлечения мнений (Рисунок 5).

Модуль лингвистической обработки. Модуль лингвистической обработки, описанный в предыдущем разделе, направлен на получение информации о структуре входного текста и решает задачу снятия оценочной полисемии. Данный модуль является универсальным для всех классов систем автоматического извлечения мнений.

Рисунок 5. Схема взаимодействия модулей предобработки и лингвистического анализа текста

Модуль автоматического извлечения мнений, в обобщенном виде представленный ниже (Рисунок 6), является центральным в системе автоматического извлечения мнений. Необходимо отметить, что в данном разделе под «объектами» и «субъектами» понимаются «объекты мнения» и «субъекты мнения», а не единицы синтаксического членения предложения.

Подмодуль фильтрации объектов. Данный компонент предназначен для сокращения числа объектов, подвергаемых анализу. На вход данному подмодулю поступают результаты синтактико-семантического анализа текста, на выходе к этим результатам добавляются метки для слов, не подлежащих обработке в модуле определения объекта. Строго говоря, подмодуль фильтрации контента на уровне объектов относится к промежуточному этапу анализа, находящемуся между этапом лингвистической обработки и этапом извлечения мнений. Наше решение включить подмодуль фильтрации объектов непосредственно в модуль

автоматического извлечения мнений базируется на следующем соображении. Автоматическая фильтрация объектов анализа является специфическим этапом предварительного анализа текста именно в рамках автоматического извлечения мнений. В рамках других задач автоматического анализа текстов под фильтрацией объектов часто понимается определенная конфигурация, в которой вручную указываются обрабатываемые объекты и/или их классы (ср. подходы компаний «Медиалогия» и «Крибрум», а также уже упоминавшееся диссертационное исследование А.А. Толкунова [167]).

Подмодуль определения субъекта. Данный компонент опирается на синтактико-семантическую структуру текста. Результатом работы подмодуля является простановка атрибутов субъекта мнения и источника мнения. При этом для разных моделей анализа субъекты мнения могут исключаться из списка объектов (вторичная фильтрация объектов). Именно данный фактор влияет на последовательность следования подмодулей определения субъекта и объекта.

Подмодуль определения объекта. Данный компонент приписывает сущностям семантического анализа атрибут объекта с учетом их семантического типа. При этом для разных конфигураций (моделей анализа) набор анализируемых сущностей может различаться. Стоит отметить, что в явном виде компонент определения объекта отсутствует (см. Рисунок 6). Данный подмодуль выполняет единственную функцию приписывания метки объекта сущности, а данные о том, включать ли сущность в список объектов, передаются из подмодулей фильтрации объектов, модели анализа и метаданных. Особым случаем определения объекта является его имплицитное задание. В данном случае идентификация объекта анализа производится при преобразовании специализированных оценочных словарей, которые детально рассматриваются в Главе 3.

Подмодуль определения оценки. Ключевым компонентом в модуле автоматического извлечения мнений является подмодуль определения оценки. На вход данного подмодуля поступает список объектов с их синтактико-

семантическими характеристиками. На выходе каждому объекту приписываются атрибуты оценки.

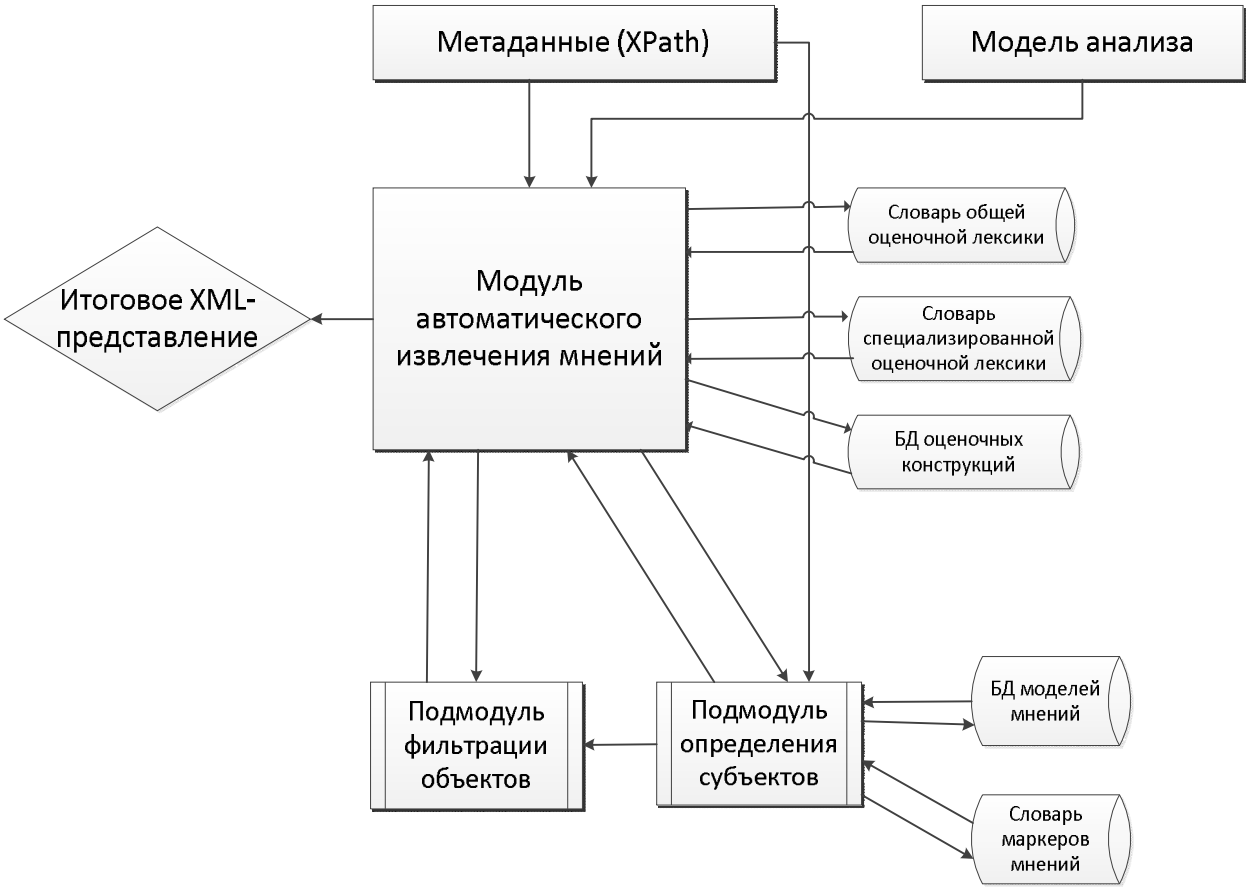

Рисунок 6. Обобщенная схема модуля автоматического извлечения мнений

2.4.1.2. Подмодуль определения объекта[49]

Как уже отмечалось выше, объект чаще всего задается настройками конфигурации, но для полноценного решения задачи определения объекта также необходимо использовать механизм расширенного поиска объекта при помощи следующих механизмов: расширение запроса при помощи онтологии; детализация объекта за счет его структуры (см. п. 2.4.1.2.1) и имплицитное задание объекта.

Одним из последних трендов в автоматической обработке текстов последних десятилетий является активное привлечение экстралингвистических данных [39]. В рамках задачи автоматического извлечения мнений онтологии, в

[49] В основе параграфа работа автора [117].

основном, привлекаются для расширения списка объектов анализа. Рассмотрим два случая[50] возможного применения онтологий. В рамках первой задачи рассматриваются мнения пользователей форума о возможных победительницах конкурса красоты. В рамках второй задачи анализируется возможность покупки значительного пакета акций крупной компании. Условиями покупки являются высокая доходность актива, а также минимально возможные риски, связанные с активом.

Форумы являются традиционной областью применения технологии автоматического извлечения мнений. К их особенностям относятся: 1) наличие отдельных веток обсуждения, посвященных конкретным вопросам или их аспектам; 2) активное использование лозунгов и 3) обилие неполных предложений. Для решения первой задачи имеются следующие входные данные: имя объекта анализа (*Юля*), набор ее социальных характеристик. Таким образом, для решения задачи необходимо 1) выявить все оценочные высказывания относительно целевого объекта, 2) выявить список конкурентов, 3) выявить все оценочные высказывания относительно конкурентов и 4) сравнить результаты.

Задачи 1-3 часто невозможно разделить по сугубо лингвистическим причинам. Первым (и наиболее точным) из возможных действий является анализ сравнительных конструкций. По мнению ряда исследователей, анализ сравнительных конструкций [например, 26] считается одной из наиболее трудноподдающихся формализации задач в автоматическом извлечении мнений. Если анализ прямых сравнений вида *Маша красивее Юли* не представляет значительных проблем, то перефразировка *Маша красивее, чем Юля* приводит к усложнению модели связывания результатов выдачи синтаксического анализатора с результатами работы компонента анализа мнений. Значительно более сложный случай представляет скрытое сравнение, часто выражаемое превосходной степенью прилагательных. Рассмотрим данную ситуацию на примере предложения *Маша самая красивая девушка на свете!* Задача нашего

[50] материал носит сугубо иллюстративный характер, и использование названий реальных объектов действительности не несет рекламного характера и не относится к коммерческой тайне.

поиска мнений (как и в предыдущих случаях) заключается в поиске конкурентов девушки по имени *Юля* на конкурсе красоты. Традиционный подход заключается в поиске всех вхождений объекта анализа в анализируемый массив текстов. Запрос может быть расширен при помощи тезауруса (например, *харьковчанка*, *выпускница МГУ*, *лауреат конкурса «Молодежная весна-2008»*, и т.п.). В результате для дальнейшего анализа будут отобраны предложения, сходные с приведенными ниже: *Красивая харьковчанка пробилась в финал. Маша красивее Юли. Юля, по мнению жюри, была неотразима*. Другие предложения, не содержащие объект анализа, отобраны не будут как не соответствующие запросу. Между тем, подобные предложения могут включать релевантную задаче информацию, например, уже приводившееся предложение *Маша — самая красивая девушка на свете!* имплицитно включает в себя объект анализа *Юля*.

В подобных случаях мы имеем дело с онтологическими классами. Онтологии активно применяются для расширения поисковых запросов, но, во-первых, их объем существенно ограничен, во-вторых, составление онтологий достаточно трудоемкий процесс, и, наконец, включение всех личных имен в онтологию нецелесообразно как в техническом аспекте, так и в лингвистическом. В нашем случае даже зная о вхождении объекта анализа *Юля* в класс «*девушка*», мы столкнемся с противоречием нашего семантического анализатора, сопоставившего словоформу «девушка» из анализируемого предложения, с синтаксическим, который свяжет объект *Маша* с именным компонентом копулы *самая красивая девушка*. В результате мы получим онтологическое противоречие *«Маша=Юля»*.

Ещё более сложная ситуация возникает при анализе экономических текстов. Такие тексты имеют особую специфику, зачастую выражая оценочные суждения в числовой форме, например, *Газпром вошел в десятку крупнейших компаний мира*, или *CNN's profits rocketed by 35% in by the end of the financial year*. Даже если мы имеем список компаний-конкурентов, что довольно часто бывает в практическом использовании систем мониторинга, то необходима сложная модель мира для анализа аналогичных предложений, включающих вхождения

компаний конкурентов, например, *Казахгаз вошел в тройку мировых лидеров добычи газа в 2011 году*. Для правильной интерпретации данного предложения относительно объекта *Газпром* важно знать место *Газпрома* в данном рейтинге за такой же период. В случае с *Си-Эн-Эн* можно представить сходную ситуацию: *CBS's profits rocketed by 34% in by the end of the financial year*. В данном контексте модель мира должна содержать уже не только математические данные о разнице между величинами *0,35* и *0,34*, но и изначальные значения капиталов обеих компаний.

В силу значительных вычислительных сложностей и экспоненциального роста базы знаний при резком падении скорости обработки подобный подход представляется экономически неэффективным. По нашему мнению, выход из сложившейся ситуации возможен лишь в результате создания принципиально новой теории дискурсивного анализа, учитывающего прагматику языка во всех ее проявлениях. Таким образом, применение онтологий при определении объектов должно быть ограничено узкими предметными областями, список объектов в которых конечен.

Имплицитное задание объекта является наиболее сложной задачей рассматриваемого этапа автоматического извлечения мнений. Необходимо отметить, что среди исследователей не существует единого мнения относительно того, что следует включать в понятие имплицитного выражения мнений. Лю относит к имплицитным мнениям сравнительные конструкции [Liu, 2012], например, *Dictator Kim Jong-un is WORSE than his father, says North Korea defector*[51]. Ма и Ван [30] под этим понятием подразумевают отсутствие эксплицитно выраженного объекта в предложении (т.е. анафорические и эллиптические употребления). В другой работе Ван с коллегами [63] относят к имплицитным проявлениям объектов хеш-теги. Эйдельман считает, что имплицитное задание объектов также происходит при рассмотрении текстов с различных точек зрения (т.е. израильской или палестинской перспективы) [13]. Особая точка зрения приводится в упоминавшемся обзоре Б. Лю [26], который

[51] Диктатор Ким Чен Ын — хуже, чем его отец, — говорит северокорейский перебежчик. [Перевод мой — С.К.]

считает упоминание объекта через его свойства частным случаем имплицитных мнений.

По нашему мнению, отдельным случаем является определение имплицитных оценок, в которых объект оценки находится в том же слове, что и маркер оценки. Данный случай подробно рассмотрен в Главе 3.

2.4.1.2.1. Подмодуль определения свойств объекта

Свойства объектов играют важную роль при анализе ряда продуктов, товаров или услуг. Впервые на необходимость определения свойств продуктов указал Б. Лю [50]. Разные люди имеют разные представления о хороших или плохих свойствах и их значимости. Одним при выборе, например, мобильного телефона важно наличие фото- и видеокамеры с высоким разрешением, другим необходим легкий телефон с объемной батареей, третьим нужны все перечисленные свойства. Таким образом, учет свойств объектов при автоматическом извлечении мнений позволяет предоставить пользователю более комплексную картину об объекте. Наибольшее распространение модели, включающие информацию о свойствах объектов, получили в т. н. «рекомендательных системах».

Необходимо отметить, что модуль определения свойств объектов не требуется для большого количества классов объектов. К таковым, в первую очередь, относятся персоны и юридические лица[52]. Необходимость учета свойств объекта для конкретных семантических типов сущностей должна задаваться в конфигурации модели анализа.

Рассмотрим два наиболее популярных способа определения свойств объекта, а именно автоматическое определение свойств объекта (2.4.1.2.2) и применение онтологии (2.4.1.2.3).

2.4.1.2.2. Автоматическое определение свойств объекта

[52] Ср. подход Kribrum, в котором свойства для таких объектов, как юридические лица, определяются (http://www.ashmanov.com/services/kribrum).

Как отмечалось выше, учет свойств объекта необходим для решения достаточно специфических задач, а именно анализа специализированных ресурсов, например, сайтов с отзывами. С целью автоматического обнаружения свойств объектов можно использовать три техники: 1) использование метаинформации, 2) машинное обучение с учителем, 3) эвристический вывод свойств объектов на основе синтаксической структуры.

Рассмотрим возможности использования метаинформации на примере уже рассматривавшегося сайта http://www.otzyv.ru. Следующий XPath //*[@id="mainCol"]/div[3]/div[2]/noindex[1]/div дает возможность извлечь такие характеристики отеля как: *расположение*, *территория*, *обслуживание*, *питание*, *развлечения*, *пляж*, *для детей*. Непосредственно значения оценок для каждой из данных характеристик можно также извлечь из html-кода отзыва, но текстуальные обоснования для каждой оценки непосредственно из кода извлечь не получится. Для этого помимо упоминаний наименования отеля (с учетом его кореферентов) в тексте проводится поиск упоминаний полученных характеристик и их синонимов из тезауруса. Проблемой описанного подхода является отсутствие прямой связи между списком характеристик, которые задаются разработчиками сайтов, и характеристиками, которые указывают в теле отзыва пользователи. Таким образом, несмотря на высокую точность определения характеристик, полнота такого подхода оказывается очень низкой. При этом при подключении нового сайта возникает необходимость ручной настройки нужных XPath.

Машинное обучение с учителем требует большей предварительной работы, чем метод на основе метаинформации. Так, требуется вручную разметить объекты и их свойства на представительном массиве текстов. Кроме того, данные тексты должны быть репрезентативными, т.е. отражать все возможные комбинации объект–свойство. К таким комбинациям следует отнести сочетание свойства отеля, которое сравнивается с другим отелем, например, *Во втором корпусе- если вид на территорию, то анимация слышна весь день, если в противоположную, то шум этого странного здания или шум, который доносился от Макади Бич-*

там что-то тоже постоянно шумело.[53] Полнота подобной модели будет выше, чем у предыдущего метода, но ожидаемая точность будет ниже, т.к. в модель априорно не заложены все возможные контексты упоминаний характеристик отелей. Как и в предыдущем методе сохраняется необходимость переобучения модели для анализа нового сайта.

Эвристический вывод свойств объекта требует значительных трудозатрат на этапе создания правил. Преимуществом эвристического подхода является его большая адаптивность к материалу анализа. Так, если классы объектов, которые находятся в отношении «объект—свойство», совпадают, то их синтаксические взаимоотношения также будут тождественны. Практика показывает, что помимо чисто синтаксических свойств правила определения свойств объекта должны учитывать и дискурсивные элементы, такие как заголовки и маркеры переключения повествования. В качестве примера такой ситуации можно привести предложение *А тут 0 (ноль) рублей!*[54]

2.4.1.2.3. Определение свойств объекта при помощи онтологии

Данный метод получил наибольшее распространение в практике разработчиков коммерческих систем. Как известно, составление онтологий процесс длительный и трудоемкий. Для динамично меняющихся сфер товаров и услуг целесообразно создание онтологий совместно с создателями продуктов или лицами, предоставляющими услуги.

2.4.1.3. Подмодуль определения субъекта

Задача определения субъекта мнения примыкает к традиционным задачам компьютерной прагматики. Как отмечалось выше (Глава 1), ранние разработки в области построения диалоговых систем [например, 58] стали предшественниками задачи автоматического извлечения мнений. Кроме того, из четырех основных направлений языковой прагматики [178: 389-390] — 1) изучение субъекта речи, 2) изучение адресата, 3) изучение взаимодействия между субъектом и адресатом и 4)

[53] Анализируется отзыв об отеле *Iberotel Makadi Oasis & Family Resort 4**, http://www.otzyv.ru/read.php?id=188224
[54] http://www.banki.ru/services/responses/bank/response/8321342/ Оценивается стоимость обслуживания.

изучение ситуации общения, две имеют непосредственное отношение к задаче идентификации субъекта мнения.

Значимость модуля определения субъекта заключается в том, что он является дополнительным фильтром потенциальных объектов анализа. Необходимо отметить, что при наличии оценочных модификаторов фильтрации субъектов не происходит. Приведем пример такого случая.

Этот мудрый правитель понимал, что экономический прогресс возможен только тогда, когда общество живет по справедливым законам, а не по клановым понятиям.[55]

Прежде чем рассмотреть способы реализации субъекта в русскоязычном тексте, обратимся к структурной схеме мнения. Структура мнения редко рассматривается в исследованиях по автоматическому извлечению мнений. Наиболее часто заимствуется модель Б. Лю [например, 174], который представляет мнение в виде следующего множества *{объект, свойство объекта, оценка, субъект оценки, дата оценки}* [26: 12]. Лю указывает, что данная модель оптимальна для большинства сфер применения технологии, отмечая [Ibid.: 13], что слишком сложное определение ведет к чрезмерному усложнению задачи (ср. модель мнения А.А. Толкунова в Главе 1 [167]). У данных компонентов мнения значимость сильно различается в зависимости от задач системы, типа текста и модели представления мнения в системе.

Следует отметить, что тексты новостных лент практически не рассматриваются в работах по автоматическому извлечению мнений[56]. Таким образом, ни одна из существующих моделей мнений не моделирует непосредственно структуру новостного мнения. В работе [1: 9] под новостным мнением понимается триада *{автор, читатель, текст}*. При этом из текста извлекаются только случаи *цитат*, т.е. фрагментов прямой речи, в которых определяются объекты анализа. В новостном мнении под категорию *автор* попадают случаи авторской интерпретации фактов, категория *читатель*

[55] http://www.mk.ru/economics/2016/01/17/importozameshhenie-luch-nadezhdy-ili-doroga-v-propast.html
[56] Исключение составляет серия публикаций JRC EC (Joint Research Centre), начиная с 2009 года, а также работы А.Н. Соловьева.

определяет вариативную часть в возможной интерпретации фактов, а в категорию *текст* попадают эксплицитные оценки в тексте. В следующей публикации данного коллектива [2: 2217] обосновывается стратегия выбора прямой речи в качестве объекта анализа и описывается техника определения субъекта речи. В работах А.Н. Соловьева [например, 46] под субъектом мнения также понимается именованная или неименованная сущность, встречающаяся в словах автора (для прямой речью) или родительском предложении (для косвенной речи).

Между тем, указанные авторы игнорируют *модель передачи сообщения Шеннона-Уивера*. В данной шестиэлементной модели помимо субъекта информации также присутствует канал связи. Данный пункт представляется нам принципиальным, т.к. в текстах новостей часто присутствуют перепечатки и ссылки на другие источники. Более подробно данный вопрос рассматривается в п. 2.4.1.3.4.

С учетом канала передачи сообщения структурная модель мнения имеет вид:

$$Opinion = \{Obj;\ Sent; Subj; Chan^*; Source^*; Date^*; Feat^*\}, \quad (2)$$

где $Opinion$ — мнение, Obj — объект анализа, $Sent$ — оценка объекта, $Subj$ — субъект мнения, $Chan$ — канал передачи сообщения, $Source$ — источник мнения, $Date$ — дата сообщения, $Feat$ — свойство (характеристика объекта), * — маркер факультативного элемента.

Рассмотрим возможные способы кодирования субъекта в тексте на русском языке. Как уже отмечалось выше, ключевыми способами представления субъекта являются прямая речь (п. 2.4.1.3.1) и косвенная речь (п. 2.4.1.3.2). Кроме того, субъект может встречаться в конструкциях со свернутой косвенной речью (п. 2.4.3.3). Все конструкции, кодирующие субъект, представляются в виде особой базы данных, в которой в отдельном поле указывается тип данной конструкции. Задача разграничения субъекта мнения от канала передачи мнения и источника мнения решается после определения первичного (формального) субъекта мнения.

2.4.1.3.1. Определение субъекта в конструкциях с прямой речью

Задача определения субъекта в конструкциях с прямой речью решается при помощи шаблонов. Модели прямой речи в текстах на русском языке разнообразны и не исчерпываются классическими конструкциями с кавычками и жестко фиксированным набором знаков препинания. В качестве примеров различных способов данных конструкций приведем следующие предложения (формальные показатели прямой речи выделены подчеркиванием):

«Я не знаю ни одного хорошего примера денежно-кредитной политики, когда в условиях высокой инфляции страна бы смягчала денежно-кредитную политику», - сказала Эльвира Набиуллина, глава ЦБ России, выступая на Петербургском международном экономическом форуме.[57]

«Где моя охрана?!» – пропищал 79-летний швейцарец[58].

Он ответил: "Давайте как-то по-другому".[59]

<…> депутат заявил: “Когда через несколько лет наступит пик численности абитуриентов, то мест в вузах для них не будет”.[60]

– Оглянитесь вокруг, сюда ходят только парижане. Здесь нет ни одного туриста! – многозначительно сказал Антуан.[61]

«За три года до домашнего чемпионата мира Россия принимает решение идти разными путями с Капелло, – пишут шведские журналисты. – Работу опытного тренера подвергали резкой критике, особенно это касается результатов его команды. РФС пока не хочет называть преемника Капелло».[62]

Конструкции, представленные в приведенных примерах, просты для автоматической обработки, т.к. в них явно выражен субъект речи. Значительно сложнее определить субъект в эллиптических конструкциях, неопределенно-личных предложениях и конструкциях без явно выраженного субъекта. Конструкции без субъекта часто представляют собой обобщения («говорят»),

---

[57] http://expert.ru/2015/07/15/horoshij-primer-dlya-tsb/
[58] http://www.metronews.ru/kolumnisty/horoshij-plohoj-effektivnyj-jozef-blatter/Tpoogu---2M4DdY4OKMEDI/
[59] http://ru.tsn.ua/politika/klyuchevye-tezisy-brifinga-kolomoyskogo-horoshiy-muzhik-v-ukrnafte-i-sms-perepiska-s-ahmetovym-456620.html
[60] http://www.poisknews.ru/theme/edu/15256/
[61] http://lady.tut.by/news/food/457582.html
[62] http://www.championat.com/football/article-226401-otstavka-fabio-kapello--v-obzore-zarubezhnykh-smi.html

пассивные конструкции с речемыслительными глаголами. Приведем примеры подобных ситуаций.

Но, взяв себя в руки, указал на деньги и выдал новую порцию красноречия: «Тут надо прибраться. Это не имеет ничего общего с футболом».[63]

«Сказать, проверяются ли документы, подтверждающие вашу платежеспособность, мы не можем. Если возникают какие-то подозрения, этим может заниматься МИД Польши», — говорят в посольстве.[64]

"Начиная с 1997 года денежные потоки "Совэкс" были "захвачены и контролировались" двумя "хорошо известными российскими преступниками", - говорится в иске[65].

При определении субъекта в подобных случаях можно использовать информацию о семантическом типе объекта в клаузе, выражающей слова автора. Субъект мнения может быть выражен *персоной* или *организацией*.

Особую модель образуют заголовки новостей, содержащие прямую речь, а также реплики диалога в различных интервью. Процедура идентификации субъекта мнения в таких предложениях не отличается от традиционной модели со словами автора, что легко проследить на следующих примерах.

Тимощук: у «Кайрата» очень хороший подбор игроков.[66]

Кассиры говорят: ж/д билетов в продажу поступает мало и их разметают за час-полтора.[67]

2.4.1.3.2. Определение субъекта в конструкциях с косвенной речью

Конструкции с косвенной речью являются вторым типом конструкций, для которых характерно наличие субъекта. Определение субъекта мнения в таких случаях не представляет проблем. При этом в отличие от прямой речи для

[63] http://www.metronews.ru/kolumnisty/horoshij-plohoj-effektivnyj-jozef-blatter/Tpoogu---2M4DdY4OKMEDI/
[64] http://vesti-ukr.com/strana/108161-ukraincev-vystrojat-v-polutoramesjachnuju-ochered-za-shengenom
[65] http://www.inopressa.ru/article/20jul2015/times/miller.html
[66] http://www.championat.com/football/news-2194473-timoshhuk-u-kajrata-ochen-khoroshij-podbor-igrokov.html
[67] http://vesti-ukr.com/odessa/108129-chto-tormozit-razvitie-odessy-kak-turisticheskoj-stolicy

идентификации косвенной речи используются списки глаголов речемыслительной деятельности.[68]

Роман Скорый считает, что в Нижнем Тагиле нужно развивать индустриальный, военно-технический и патриотический туризм.[69]

Депутат Виктор Балога заявляет, что контрабанду на Закарпатье контролирует киевская власть, в частности премьер и министр внутренних дел.[70]

Два года назад состоялся совет при президенте РФ Владимире Путине, который активизировал работу по этому направлению, рассказал министр энергетики РФ Александр Новак.[71]

В относительно редких случаях прямая и косвенная речь могут соседствовать в одном предложении. При этом речемыслительный глагол, вводящий косвенную речь, стоит в форме деепричастия или (реже) причастия, например:

“Днепропетровская полиция выедет на улицы в ноябре. В целом будет сто постоянно действующих экипажей”, – отметил он, добавив, что в Днепропетровске будут работать 1100 новых патрульных полицейских.[72]

Аналогичным образом дело обстоит в случае с двумя предложениями, содержащими косвенную речь, например:

Говоря о ситуации внутри страны, Медведев сообщил, что ситуация с бюджетными расходами российских регионов сложная, но не безнадежная.[73]

В некоторых случаях однозначно идентифицировать субъект мнения в косвенной речи невозможно, что видно из примера ниже, поэтому в данных случаях в качестве потенциального субъекта выбираются все подходящие

[68] Данные о типе глагола также можно получить из семантического словаря, при наличии в нем соответствующей информации.
[69] http://www.nakanune.ru/news/2015/7/16/22407839
[70] http://www.pravda.com.ua/rus/news/2015/07/28/7075946/
[71] http://kazanfirst.ru/feed/50644
[72] http://news.finance.ua/ru/news/-/354953/yatsenyuk-skazal-kogda-v-dnepropetrovske-poyavitsya-politsiya-zarplatu-obeshhaet-takuyu-zhe-kak-i-v-kieve
[73] http://www.tvc.ru/news/show/id/73340

сущности. Непосредственное разрешение такой неоднозначности происходит на этапе разграничения субъекта мнения, его источника и канала связи.

Как сообщили Накануне.RU в пресс-службе мэрии Нижнего Тагила, Сергей Носов встретил гостя на плотине тагильского пруда и показал ему завод-музей.[74]

Как и для прямой речи, для косвенной речи характерны бессубъектные конструкции, например, в предложениях:

Отмечается также, что такое заявление может служить подготовкой инвесторов к смене руководства или просто констатацией факта возможности функционирования бизнеса , в котором задействовано несколько сотен тысяч человек, без «ручного управления».[75]

Например, иногда у нас говорят, что у нас несовершенная Конституция.[76]

В итоге он не может и не будет делать это, говорится в статье.[77]

Определение субъекта в подобного рода предложениях осуществляется по единой с прямой речью модели.

2.4.1.3.3. Определение субъекта в конструкциях со свернутой косвенной речью

Свернутая косвенная речь представляет собой два случая свертки полноценного предложения: 1) свертка придаточного предложения, 2) свертка основного предложения. В первом случае субъект мнения и глагол, вводящий косвенную речь остаются в неизменном виде, например:

Глава Североатлантического альянса Йенс Столтенберг заявил об итогах экстренной встречи, на которой обсуждались угрозы Турции.[78]

При этом процедура определения субъекта остается неизменной.

Во втором случае происходит свертка и субъекта мнения и глагола, вводящего косвенную речь. К способам свертки относятся использование

[74] http://www.nakanune.ru/news/2015/7/16/22407839
[75] http://www.gazeta.ru/business/news/2015/07/24/n_7404617.shtml
[76] http://www.unian.net/politics/1102054-boris-filatov-snachala-lyudi-vyibirayut-verhovnuyu-radu-a-potom-govoryat-chto-j-za-idiotov-myi-tam-sobrali.html
[77] http://www.inopressa.ru/article/27Jul2015/welt/rus_gold.html
[78] http://nv.ua/world/countries/gensek-nato-podytozhil-rezultaty-ekstrennogo-zasedanija-po-turtsii-61344.html

отглагольных и предикативных существительных, чаще всего, с предлогами. К отглагольным конструкциям подобного рода относятся — конструкции со словами *мнение*, *слово*, *выражение*, *цитата* и ряд других.

По мнению экспертов, отход Галицкого от управления компанией сложно оценить как «позитивный», поскольку большие международные инвесторы считают очень важной персону генерального директора в компаниях, которые были созданы с нуля, а «Магнит» ассоциируется с личностью Галицкого.[79]

По его словам, подобные решения ставят под угрозу развитие демократии на Черном континенте, передает агентство France Presse.[80]

В данном случае для определения субъекта учитываются правила изменения модели глагольного управления при субстантивации речемыслительных глаголов.

2.4.1.3.4. Разграничение источника и субъекта мнения

Как уже отмечалось выше, для разрешения случаев неоднозначного определения субъекта необходимо разграничить субъект мнения от сходных явлений, таких как источник мнения и канал связи. Данные явления нехарактерны для блогов и иных социальных медиа и встречаются в основном в текстах СМИ. Например, в предложении из примера субъектом мнения является *Столтенберг*, а источником мнения — *Европейская правда*.

“Терроризм представляет прямую угрозу безопасности стран НАТО, международной стабильности и процветанию. Это глобальная угроза, которая не знает границ, национальности или религии – вызов, с которым международное сообщество должно бороться и бороться вместе”, - цитирует Столтенберга ***Европейская правда****.*[81]

К наиболее характерным различиям между СМИ и социальными медиа относятся, например, средства оформления цитат. В социальных медиа это, чаще всего, гиперссылка на источник, в то время как в текстах СМИ цитаты

[79] http://www.gazeta.ru/business/news/2015/07/24/n_7404617.shtml
[80] http://riafan.ru/354951-obama-ya-horoshiy-prezident-no-ya-ustal
[81] http://nv.ua/world/countries/gensek-nato-podytozhil-rezultaty-ekstrennogo-zasedanija-po-turtsii-61344.html

оформляются при помощи речемыслительных глаголов и отглагольных конструкций. Под цитатой мы понимаем повторение чужой речи со ссылкой на источник. Рассмотрим следующий пример, *«Все хотят, чтобы его (тело Тамерлана) отправили в Россию», - приводит слова руководителя похоронного бюро Питера Стефана NBC News.*[82] Исходным источником мнения здесь является руководитель похоронного бюро Питер Стефан. А источником, из которого мы узнали о данном мнении, — информационное агентство NBC News. Подобные случаи, когда субъект мнения отличается от конечного источника мнения, мы будем считать случаями двойного цитирования.

Двойное цитирование является достаточно частым явлением в текстах СМИ. Попробуем разграничить два источника мнения с целью выявления субъекта мнения. Для этого рассмотрим глубинно-синтаксическую структуру анализируемого предложения. В клаузе *приводит слова руководителя похоронного бюро Питера Стефана NBC News* на глубинном уровне содержится две клаузы, одна из которых в поверхностной структуре выражена глаголом и подлежащной именной группой, а другая представлена отглагольным дополнением с зависимым компонентом. Отглагольное дополнение представляет собой свернутое предложение с глаголом речи в качестве вершины. Согласно экстралингвистическим сведениям в функции информационного агентства не входит непосредственное порождение речевых единиц, в то время как подобное речепроизводство является коммуникативной потребностью индивида. Попробуем прийти к подобным выводам с лингвистической точки зрения.

Согласно гипотезе глубины В.Ингве свернутые конструкции представляют собой компактное представление уже известной информации. Таким образом, данная конструкция имеет больший вес в информационной структуре документа (в соответствии с теорией центрирования). В результате мы получаем рабочую гипотезу о том, что свернутая конструкция будет более достоверным субъектом мнения.

[82] http://forumkavkaz.com/index.php?topic=53.1770

Рассмотрим другие примеры двойного цитирования. *Газета New York Times сообщила в пятницу со ссылкой на источник в Белом доме, что Обама якобы потребовал от Пентагона увеличить количество целей для удара коалиции американских и французских войск по территории Сирии, а также изучить возможность задействования в этой операции сил ВВС*.[83] В данном предложении в качестве свернутой конструкции выступает клише *ссылка на*, которое указывает на то, что подлежащее предложения не является субъектом данного мнения. В то же время наличие отглагольных конструкций является обязательным условием превращения группы дополнения в субъект мнения. Так, в предложении *Сумма бюджетных средств, сэкономленных в случае отмены второго тура выборов мэра Новосибирска, составит не 8-9 миллионов рублей, как было сказано ранее, а 11,6 миллиона, сообщил журналистам в пятницу зампредседателя Совета депутатов Новосибирска Дмитрий Асанцев*[84] журналисты являются не исходным, а конечным источником мнения, и, следовательно, не могут претендовать на роль субъекта мнения.

Двойное цитирование является, по-видимому, частным случаем *n-арного цитирования*, где n≤9, в соответствии с упомянутой гипотезой В.Ингве. Так, в предложении *Активистки также приводят слова пресс-секретаря Путина Дмитрия Пескова, растиражированные российскими СМИ: «Это откровенное хулиганство.»*[85] можно выделить 3 источника мнения, которые отличны от его субъекта (Путин).

Наличие дополнительных источников мнения, которые не являются его субъектами, не вписывается в схему мнения, описанную Б.Лю. Таким образом, необходимо выяснить, чем являются подобные источники мнений.

Для решения данной задачи обратимся к обобщенной схеме сообщения (Рисунок 7).

[83] http://ria.ru/arab_riot/20130906/961375679.html
[84] http://news2world.net/politicheskie-novosti/novosibirsk-sekonomit-na-viborah-v-odin-tur-bolshe-chem-ozhidalos.html
[85] http://mr7.ru/articles/81301/

Рисунок 7. Обобщенная схема сообщения

Данная схема, учитывающая каналы связи и источник сообщения, на наш взгляд более адекватно описывает случай двойного цитирования, чем схема Б.Лиу, содержащая только субъект мнения. При этом следует учитывать, что источник сообщения может иметь и метаязыковое кодирование, как в примере ниже, где источник указан только в виде URL-адреса. *Пресс-секретарь президента России Владимира Путина Дмитрий Песков заявил, что возможная отставка ректора Российской экономической школы (РЭШ) Сергея Гуриева и его отъезд за границу никак не связаны с политикой. Об этом сообщает 29 мая сообщает «Коммерсант».*[86] (Сайт портала «KM.RU»).

2.4.1.3.5. Теоретические проблемы морфологии русского языка при автоматическом определении субъекта

2.4.1.3.5.1. Проблема одушевленности собирательных существительных

[86] http://www.km.ru/v-rossii/2013/05/29/ekonomika-i-finansy/711933-press-sekretar-putina-otkazalsya-svyazyvat-ukhod-guri

В ходе анализа возможных субъектов была выявлена проблема с подходом к выявлению субъектов при помощи семантических типов. Она связана с неполнотой семантических ресурсов, используемых при семантическом анализе текста. В качестве альтернативного решения было решено использовать данные о категории одушевленности сущности, к которым относится большинство проанализированных сущностей.

Между тем, некоторые релевантные субъекты мнений не подходят под категорию одушевленности. Данные случаи подпадают под две категории: собирательные существительные и организации[87]. В случае с семантическим типом «*организация*» мы имеем дело с метонимической заменой «работник организации => организация». Собирательные существительные в русском языке относятся к неодушевленным [105].

Между тем, собирательные существительные часто сочетаются с речемыслительными глаголами, что нехарактерно для неодушевленных существительных. Случаи-исключения типа «*Указанные факты говорят о значимости случайных ненаправленных воздействий на популяцию.*[88]» являются метафорическими и характерны практически исключительно для подъязыка науки. Существует также ряд фраз-клише типа «*факт/ситуация говорит*».

В семантической разметке НКРЯ слова типа *молодежь*, *народ*, *духовенство* имеют семантическую характеристику человека (sc:hum). При этом все три слова являются неодушевленными, что соответствует современной русистике, где признаками одушевленности/неодушевленности считаются чисто формальные критерии [например, 105]. Именно по формальным критериям (например, винительный падеж, сочетаемость с прилагательными во множественном числе) собирательные существительные в русском языке относятся к неодушевленным. С другой стороны, категория *одушевленности/неодушевленности* для некоторых языков (например, английского) является скрытой [81]. Если предположить, что для русского языка одушевленность собирательных существительных также

[87] В эту же категорию мы относим геополитические сущности.
[88] http://webground.su/rubric/2010/07/22/nauka_fizika/retro/

является скрытой категорией, то это не только поможет решить сугубо техническую задачу определения субъектов мнений, но также поможет объяснить сочетаемость семантического класса «организация» с речемыслительными глаголами.

Исходя из предположения, что собирательные существительные, некоторые из которых имеют неполные парадигмы[89], наследуют данную категорию от составляющих их элементов, можно также вывести правила их сочетаемости с некоторыми оценочными глаголами, например «*ненавидеть*». Сравним три предложения, в которых у глагола *ненавидеть* разные субъекты, а именно одушевленное существительное, собирательное существительное со значением «совокупность» людей и неодушевленное существительное.

Путин патологически ненавидит Лукашенко.[90]

Известный российский актер Владимир Епифанцев заявил, что российская молодежь ненавидит Русскую Православную церковь (РПЦ).[91]

Альфа-банк ненавидит надёжных клиентов[92].

В первых двух примерах субъекты глагола «*ненавидеть*» оценочно нейтральны, а в третьем — негативен. Таким образом, собирательные существительные, обозначающие совокупности людей, в оценочном плане ведут себя так же, как и другие одушевленные существительные.

Ещё одним свидетельством в пользу одушевленности собирательных существительных, обозначающих группы людей, является их способность выступать в качестве субъекта при наиболее типичных представителях речемыслительных глаголов, таких как *разговаривать*, *считать* и *думать*. Продемонстрируем это на основе следующих примеров:

Часто наблюдаю, как молодежь выходит на дорогу, разговаривая по телефону, и даже не смотрит по сторонам.[93]

---

[89] Что затрудняет формальную идентификацию одушевленности.
[90] http://obozrevatel.com/abroad/98567-patologicheskaya-tupost-belarusskij-politik-rasskazal-za-chto-putin-nenavidit-lukashenko.htm
[91] http://www.mv.org.ua/news/109186-izvestnyi_rossiiskii_akter_priznal_chto_rossija_-_strana_tretego_mira.html
[92] http://irubtsov.ru/2012/08/19/al-fa-bank-nenavidit-nadyozhny-h-klientov/ В примере заменен верхний регистр.
[93] http://ncrim.ru/articles/view/25-09-2015-na-ulicah-sevastopolya-opasno-perehodit-dorogu

Официальное духовенство в данном случае считает, что молитву можно перенести из-за выполнения чиновниками их законных обязанностей.[94]

Проснулось сознание народа, он думает не о своем закутке, а об обществе, не о роде, а о нации.[95]

2.4.1.3.5.2. Автоматическая обработка неадаптированной лексики

Другой проблемой морфологического уровня является формальное представление словоизменительных парадигм существительных-заимствований. Особенно данная проблема характерна для этнорелигиозной лексики. С точки зрения задач автоматического извлечения мнений проблема состоит в том, что не происходит отождествления одинаковых субъектов и/или объектов мнений.

Следует отметить, что компьютерная морфология зачастую оперирует собственными морфологическими интерпретациями тех или иных словоформ (ср., например, известное высказывание М.И. Стеблина-Каменского: «*Русское слово порося оказывается возвратным глаголом со значением существительного*» [Стеблин-Каменский, 1958: 4]). Поэтому прикладные решения в области автоматического морфологического анализа могут противоречить традиционным описаниям словоизменительных явлений [107].

Этнорелигиозная лексика зачастую бывает представлена двумя морфологическими вариантами: 1) изменяющимся по законам языка-реципиента, 2) продолжающим изменяться по законам языка-донора. Именно слова второго типа вызывают сложности при автоматическом морфологическом анализе текста. Это вызвано тем, что существующие алгоритмы морфологического анализа моделируют морфологическую систему языка-реципиента [73, 71; 98; 52].

Рассмотрим в этом аспекте следующие особенности этнорелигиозной лексики. Религиозная лексика в европейских языках заимствована (иногда через языки-посредники, например, греческий, латынь) с Ближнего Востока. Именно поэтому, на наш взгляд, данные слова имеют несколько форм образования

[94] http://www.news-asia.ru/view/8806
[95] http://www.gezitter.org/vybory/44101_kak_proshli_parlamentskie_vyiboryi/

множественного числа. В качестве примеров таких слов можно привести следующие — *серафим*, *сахават*, *муджахид*.

Рассмотрим примеры употребления слов данной группы с отклонениями в образовании множественного числа.

Соединенные Штаты демонстративно убрали иранскую оппозиционную группировку «Народные моджахеддины Ирана» из списка террористических организаций.[96]

Но, в общем, моджахеддины здесь были правы, и качество тогдашней армии Боснии и Герцеговины было весьма низкое.[97]

Этническая лексика, чаще всего, заимствуется из языка-источника. Это приводит к сосуществованию нескольких основ для одного и того же понятия. В качестве примеров таких слов можно привести следующие — *ашкеназ*, *фолькcдойч*.

Рассмотрим примеры употребления слов данной группы с отклонениями в образовании множественного числа.

Введенное с 1939 года нормированное распределение продуктов питания, одежды, обуви по карточкам позволило обеспечивать всем необходимым своих рейхсдойч почти до самых дней существования Третьего Рейха.[98]

Из них 965 (88%) сгруппировались по 55-ти ветвям, состоящим почти исключительно из ашкенази.[99]

Существование этнонимов типа *рейхсдойч*, *ашкеназ*, имеющих 3 и 2 словоизменительных модели (*рейхсдойч — рейхсдойчи|рейхсдойче|—*; *ашкеназ — ашкеназы|ашкеназим*), и религионимов типа моджахед(д) и муртад(д), имеющих по 2 словоизменительные парадигмы (традиционная -ы и заимствуемая -ин), требует объединения в единой единице анализа всех существующих парадигм. Чаще всего, в программах автоматического анализа текстов объединение таких

---

[96] http://mir-politika.ru/tags/%D0%BC%D0%BE%D0%B4%D0%B6%D0%B0%D1%85%D0%B5%D0%B4%D0%B4%D0%B8%D0%BD%D1%8B/
[97] http://www.saper.etel.ru/history/musulman-vs.html
[98] http://mreadz.com/new/index.php?id=346373&pages=98
[99] http://pereformat.ru/2014/01/dna-genealogy-jews/

единиц происходит при помощи тезауруса [136] (см. Рисунок 8). На рисунке непрерывные линии обозначают прямые переходы от одной словарной единицы к другой. Пунктирная линия обозначает допустимый переход (при помощи функции поиска «AnyDescendant»). Прерывистая линия обозначает непрямой переход.

Базовым недостатком этого подхода является лишний шаг при морфологическом синтезе словоформ, например, при расширении поискового запроса или увеличение количества синонимов в специализированных базах данных [119].

Другим подходом к объединению таких единиц может служить ссылочный прием, при котором в базе хранится не иерархическая структура, а горизонтальные связи между объектами (см. Рисунок 9). При этом связь может быть только однонаправлена, поэтому задать таксономические отношения невозможно.

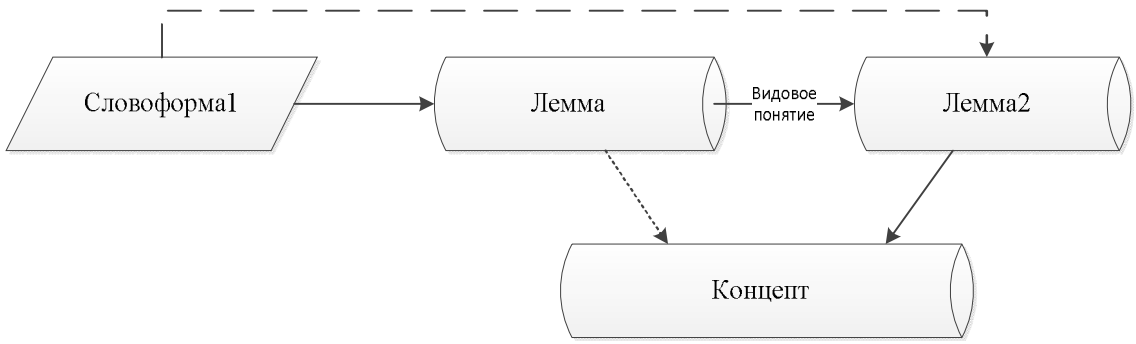

Рисунок 8. Идентификация концепта по словоформе при помощи тезауруса

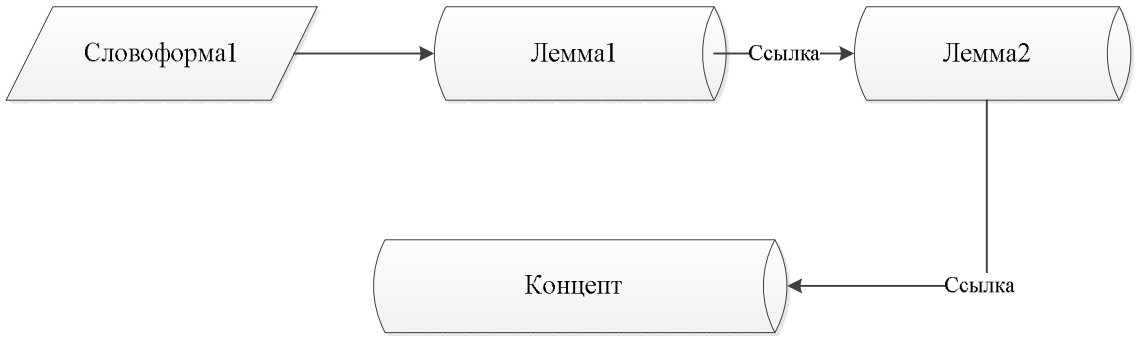

Рисунок 9. Идентификация концепта по словоформе при помощи ссылок

Приведем примеры использования неустоявшихся заимствований.

Вахман этот, пан Загурски, был фольксдойчем, но немецкий язык знал не хуже любого немца.[100]

Этот был рейхсдойче<…>, он мог даже позволить себе быть «справедливым».[101]

Хотя дивизия формировалась в теории из рейхсдойчей, в нее попало много фоль-ксдойчей, которые также носили руны СС в правых петлицах мундиров.[102]

Кого больше во 2 мировую войну погибло, ашкеназей или сефардов?[103]

Кроме того, сами дойчманы утверждают, что это существо из ашкеназей.[104]

Путин и Медведев, ашкеназы, проиграли сефардам Яценюку, Тимошенко-Капительман, Тягнибоку-Фротману.[105]

Ашкеназим были прадедушки, деды из ашкеназим перешли в советских людей.[106]

мамзик, я сефардов от русских ашкеназей тоже отличаю.[107]

Моджахеддины еще долго не сдавались[108]

Представь, серьезный дядя занимается своими серьезными делами, у правоверных куча проблем с кяфирами и муртаддинами, сплошная би'да и харам, а тут к нему подбегает его родственник (или соратник) и спрашивает "чё делать, муха в суп попала".[109]

Рассмотренные решения ведут к существенному увеличению ручной работы при составлении специализированных баз данных. При этом увеличивается количество полей в лемматизированных БД [119]. При этом качество

---

[100] http://alexander-apel.narod.ru/library/dohodyaga/part19.htm
[101] Там же.
[102] http://pirochem.net/index.php?id1=3&category=voennaya-istoriya&author=kiselev-vi&book=diviziiss2002&page=12 Сохранена орфография оригинала.
[103] https://otvet.mail.ru/question/169019485
[104] http://newsland.com/news/detail/id/576565/
[105] http://samlib.ru/i/iwanowmiljuhin_j_z/zzz-1.shtml
[106] http://old2.booknik.ru/colonnade/letters/lichnyyi-jeleznyyi-zanaves/
[107] http://www.gday.ru/forum/%CE%E1%F9%E8%E9-%F4%EE%F0%F3%EC/217638-%E8%F1%F2%EE%F0%E8%FF-%EB%FE%E1%E2%E8-3.html
[108] http://www.studfiles.ru/preview/1742725/page:25/
[109] http://www.nn.ru/community/velo/main/uskorit_velosiped_staroy_tetki.html

автоматического поиска при помощи таких БД будет существенно ниже запросов с использованием усечений (stem*).

В качестве выхода из данной ситуации можно использовать комплексные «понятийные» леммы, представленные ниже (Рисунок 10). Леммы при таком подходе могут содержать различные основы, что позволяет описать специфику этнонимов с разными основами. В отличие от тезаурусного представления «понятийная» лемма является более простой структурой, что позволяет быстрее оперировать ее данными.

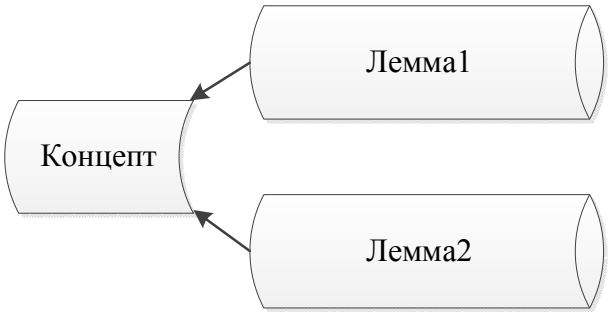

Рисунок 10. Упрощенная структура представления «понятийных» лемм

Необходимо отметить, что для эффективной автоматической обработки заимствований с неустоявшейся словоизменительной парадигмой требуется решение, направленное на максимальное покрытие словоформ, которые описывают одно понятие. В качестве подобного решения предлагается использовать «понятийные» леммы. Ещё одним дополнительным требованием может служить «двойной» морфологический анализ, при котором вначале анализируются слова с неустоявшейся словоизменительной парадигмой, а затем оставшиеся слова. Несмотря на то, что данное решение не является вычислительно эффективным, оно полностью соответствует требованию точности проводимого анализа.

Подводя итог, можно констатировать, что проблемы теоретического лингвистического описания существенно влияют на полноту и компактность представления информации в системе автоматического извлечения мнений.

2.4.1.4. Подмодуль определения причинно-следственных отношений[110]

2.4.1.4.1. Назначение модуля определения причинно-следственных отношений

Подмодуль определения причинно-следственных связей не является обязательным компонентом системы автоматического извлечения мнений. Среди формально (в основном, синтаксически) выраженных маркеров явлений, родственных оценке, некоторые авторы [например, 52: 57] подчеркивают особую значимость причинно-следственных связей. Для автоматического извлечения мнений определение причины, по которой пользователю нравится или не нравится продукт или услуга, в последнее время признается практически столь же значимым, как и само определение позитива/негатива в высказывании по отношению к продукту.

Как уже отмечалось выше (п. 2.3), условные предложения, формально выражающие вероятные причинно-следственные отношения, фильтруются на более раннем этапе обработки. Остальные способы выражения причинности можно разделить на следующие 4 типа: интрасентенциальные (п. 2.4.1.4.2), интерсентенциальные (п. 2.4.1.4.3), внутриклаузные (п. 2.4.1.4.4) и лексикализованные (п. 2.4.1.4.5). Последовательность обработки причинно-следственных связей приведена ниже (Рисунок 11)[111].

[110] В основе параграфа лежат работы автора [125; 126].
[111] Необходимо отметить, что в данной схеме не отмечен этап фильтрации условных предложений.

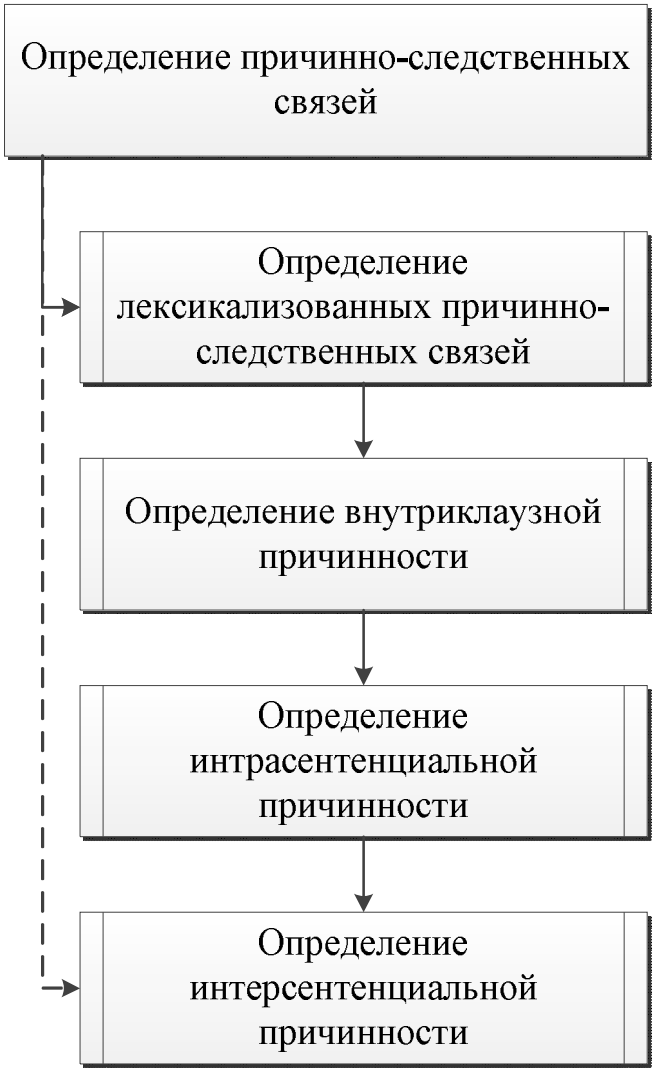

Рисунок 11. Схема подмодуля определения причинно-следственных связей

2.4.1.4.2. Определение интрасентенциальных связей

Для русского литературного языка группой исследователей под руководством академика Ю.Д. Апресяна был составлен исчерпывающий список союзов и предлогов, вводящих причины и следствия в предложения [67]. Наиболее частым случаем является нахождение союзов, выражающих причины и следствия, в одном и том же предложениях, например, *И если в начале XX в. этих животных в Австралии насчитывалось около 20 млн, то к середине столетия - уже 750 млн*[112]. В некоторых случаях союзы могут опускаться, например, *Но если*

[112] http://riatribuna.ru/news/2010/12/23/3688/

в Европе история покорения континента этим зверьком проходила более-менее мирно, в Австралии ситуация сложилась куда более драматично[113].

Процедура определения клауз причины и следствия проводится при помощи конечных списков союзов причины и следствия при поверхностном синтаксическом анализе. В случае эллипсиса одного из союзов информация о типе клаузы определяется на основе комбинаторной схемы вариантов взаимного расположения клауз причины и следствия. При использовании глубинного синтаксического анализа информация берется непосредственно из дерева синтаксического разбора.

2.4.1.4.3. Определение интерсентенциальных связей

В более редких случаях союзы причины и следствия расположены в различных предложениях, например, *Из Самарского областного кардиологического диспансера уходят детские кардиологи. Потому что их достало постоянное вранье начальства, потому что им надоело делать постоянно хорошую мину при плохой игре, потому что не видят возможности спасать детей и отвечать за результат в таких условиях, в которые они поставлены*[114]. Процедура идентификации предложений, описывающих причину и следствие, во многом сходна с определением интрасентенциальных связей. Различие заключается в невозможности использования только результатов глубинного синтаксического анализа.

2.4.1.4.4. Определение внутриклаузной причинности

Внутриклаузная причинность выражается двумя способами. Первым из них является использование предлогов, а вторым — лексикализация (см следующий пункт). Предлоги, обозначающие причинность, в отличие от аналогичных союзов, неоднозначны. Например, предложение *Из-за резкого торможения женщина упала.* выражает причинность, а предложение *Из-за угла показался полицейский.* — нет. В большинстве случаев многозначность предлогов можно разрешить за

[113] Ibid.
[114] http://m.echo.msk.ru/blogs/detail.php?ID=931642

счет их сочетаемостных свойств. Одним из ключевых факторов здесь является наличие зависимого отглагольного или предикативного существительного.

2.4.1.4.5. Определение лексикализованной причинности

Под лексикализованной причинностью мы понимаем такое языковое явление, при котором языковые средства, выражающие следствие, одновременно выражают оценку. В качестве примеров можно привести нижеследующие.

По подбору игроков, по нынешней форме «Зенит» — явный фаворит.[115]

Еще одна странность – нет ни одного доказательство присутствия сына политика в рядах Вооруженных сил.[116]

Как видно из указанных примеров, лексикализованность оценочно-следственных отношений выражена при помощи деадъективов или девербативов. Особый случай лексикализации следствия образуют перечисления и обобщения. Перечисления свойств и как результат их оценочная характеристика — один из наиболее частотных способов сжатого представления информации на сайтах отзывов. Подобного рода информация извлекается как при помощи анализа метаинформации (п. 2.1.1), так и при помощи лексических шаблонов, которые содержат наиболее частотные оценки-следствия. К данной группе слов, в первую очередь, относятся слова *плюс*, *минус*, *достоинство* и *недостаток*. Рассмотрим пример использования такого шаблона на фрагменте описания точилки из блога[117]:

Плюсы: - высокое качество исполнения. - отличные точильные свойства. - два типа зернистости. – компактность

Минусы: - в сложенном состоянии не фиксируется намертво. - не самая демократичная цена

Html-код данного фрагмента не содержит какого-либо особого метапредставления для автоматического определения при помощи технологии

[115] http://www.championat.com/football/news-1935884-radimov-zenit---javnyj-favorit-matcha-so-spartakom.html
[116] http://www.dni.ru/polit/2014/9/19/281226.html
[117] http://cheger.livejournal.com/478311.html

XPath. Поэтому к данному тексту применяется лексический шаблон, который имеет вид:

{EvWordType}{Colon}({AnyNonWordSeq}{SpacedText})+, (3)

где EvWordType — оценка-следствие с указанием полярности; Colon — двоеточие; AnyNonWordSeq — любая последовательность небуквенных символов, отличных от двоеточия[118]; SpacedText — буквенная последовательность с пробелами; проверка сгруппированного элемента выполняется минимум один раз.

Другим способом задания оценочного следствия является использование последовательности-разъяснения, например:

У этих разборов единственный недостаток: в следующий раз сценаристы Службы безопасности Украины могут попытаться избежать таких ляпов.[119]

Формальным показателем данного явления можно считать наличие двоеточия или тире непосредственно после оценочного слова, представленного деадъективом или девербативом. В качестве наиболее эффективного механизма определения лексикализованной причинности, по нашему мнению, выступают лексические шаблоны.

2.4.1.4.5.1. Вторичные тексты

Исследования в области вторичных текстов ведутся достаточно давно. В настоящее время существует несколько различных определений вторичного текста [см., например, 101: 18; 102: 11; 161: 3]. Их объединяет обязательное наличие первичного текста (исходного текста, текста-основы). Анализ литературы позволяет выявить следующие ключевые особенности вторичных текстов:

- Представляют собой результат аналитической обработки первичного текста;
- Характеризуются значительной речевой компрессией;

[118] В данном случае к небуквенным символам относятся неразрывный пробел и тире.
[119] http://awas1952.livejournal.com/3828452.html

– Разнообразны по жанрам и форме изложения: реферат, пересказ, аннотация, научно-популярная статья, пародия и др.

Новые формы коммуникации приводят к появлению новых текстовых жанров, например, твитов, Интернет-отзывов, демотиваторов и пр. Одними из наиболее значимых для разработок в области автоматического извлечения мнений вплоть до самого последнего времени являлись тексты Интернет-отзывов. К причинам выбора отзывов в качестве объекта анализа необходимо отнести, в первую очередь, явно выраженную оценку и социально-экономическую составляющую, которыми характеризуются подобные тексты.

Тексты Интернет-отзывов с лингвистической точки зрения обладают следующими отличительными чертами:

– Содержат оценку;
– Не всегда основаны на тексте (например, отзывы о песнях, отдыхе, кинофильме);
– Изобилуют свернутыми конструкциями;
– Характеризуются предшествованием следствия причине;
– Характеризуются доминированием повествования от первого лица.

Перечисленные черты Интернет-отзывов во многом пересекаются со свойствами вторичных текстов. Исходя из этого мы делаем вывод о том, что Интернет-отзывы являются частным случаем вторичных текстов.

Ретроспективная модель текстопорождения

Для подтверждения гипотезы о принадлежности Интернет-отзывов к вторичным текстам обратимся еще раз к базовым характеристикам вторичных текстов. Вторичные тексты[120] представляют собой осмысление некоторого первичного текста. Таким образом, в них содержится значительная доля субъективного восприятия прочитанного. Это подтверждают, например, эксперименты А.И. Новикова [151], в которых к наиболее частотным реакциям на текст относятся мнения и оценки.

[120] Рассматриваются только вторичные тексты, порожденные человеком, автоматически сгенерированные тексты не являются предметом нашего рассмотрения.

Интернет-отзывы не являются однородными по структуре, содержанию и описываемому объекту. Так отзывы на книги или статьи в Интернете являются вторичными текстами в полной мере. Их отличия от печатных аналогов заключаются в более эмоциональном содержании и большей резкости. Другие типы отзывов в качестве первичного источника реакции имеют не текст, а нечто иное — чаще всего, ситуацию или происшествие.

К общим чертам Интернет-отзывов со вторичными текстами относится, в первую очередь, строго заданная внешняя ситуация. В ее качестве может выступать как первичный текст, так и любое событие в прошлом. В этом заключается ещё одна общая особенность — ретроспективный характер изложения. Иными словами при построении результирующего текста используется общая ретроспективная модель текстопорождения. К чертам ретроспективного текстопорождения следует отнести отсылки к прошлому (времени или опыту) и высокую субъективность.[121]

К другим общим чертам можно отнести сжатость изложения, наличие логических возвратов, типа «*ещё вспомнил*». В Интернет-отзывах, целью которых является информирование других пользователей о собственном личном опыте, обязательно присутствует эксплицитно выраженная оценка. Последовательность аргументов при этом обычно «плюс => минус», реже присутствует только один знак. Сжатость изложения в Интернет-отзывах, иногда достигающая максимума, приводит к активному использованию лексикализации оценочно-следственных отношений, например:

Не понравилось![122]

Иногда сжатость изложения приводит к появлению ограничений на стиль отзыва, его длину или детализацию, например:

Только просьба - не надо писать просто в общем "отстой", "10 лет назад было бы круто" или наоборот "вау, нравится всё".[123]

---

[121] ср. признаки ретроспективного изложения в научном стиле [Ефремова, Н.В., 2015: 143]
[122] http://gorod1277.org/?q=content/ne-ponravilos, заголовок текста
[123] http://forum.allods.ru/showthread.php?t=119664

По нашему мнению, Интернет-отзывы являются частным случаем вторичных текстов, в которых на первый план выдвигается не информационный, а оценочный компонент, и которые характеризуются большим количеством лексикализованных оценочно-следственных отношений.

2.4.1.5. Словарь оценочной лексики, его содержание и структура

Как отмечалось в Главе 1, словарь оценочной лексики является центральным компонентом в системе автоматического извлечения мнений. В данном разделе мы проанализируем существующие лингвистические подходы к организации словарей оценочной лексики и их интерпретации (п. 2.4.1.5.1), опишем способы представления силы оценки (п. 2.4.1.5.2), рассмотрим оценочные классы лексики и их представление в словаре, а также в статистических системах (п. 2.4.1.5.3). Затем рассмотрим возможности учета частеречной (п. 2.1.5.4) и фразеологической информации (п. 2.1.5.5). Далее описываются правила сочетания оценочных слов (п. 2.1.5.6). В заключительной части раздела приводится обзор недостатков подходов на машинном обучении (п. 2.1.5.7).

2.4.1.5.1. Существующие лингвистические подходы

Как отмечалось выше (п. 1.8), в последнее время отмечается рост интереса к лингвистическим аспектам автоматического извлечения мнений. Косвенным свидетельством этого может служить появление обзорной статьи М. Табоада в журнале *Annual Review of Linguistics* [50]. В отличие от авторов других существующих обзорных статей [например, 18; 17; 5; 34] Табоада рассматривает не методологии машинного обучения, сферы применения или уровни анализа (документный, предложения и объектный), а роль лингвистического анализа в задаче извлечения мнений. В отличие от обзора [48], рассматривающего терминологические[124] и общелингвистические вопросы[125] автоматического извлечения мнений, в данной работе основной упор делается на уровни анализа (слово, фраза, клауза, предложение, дискурс) и модификаторы оценки (усилители,

[124] в части соотношения терминов *эмоция*, *чувство*, *оценка*.
[125] в части понятий интерсубъективность, эвиденциальность и др.

уменьшители, отрицание). Табоада, понимающая под словарным подходом комбинацию словаря оценочной лексики и набора правил оценочной сочетаемости, отмечает, что именно словарный подход к автоматическому извлечению мнений является наиболее многообещающим, т.к. он демонстрирует синергию между вычислительным и лингвистическим подходом [50: 328].

В работе описывается классификация оценочной лексики[126] по частям речи и силе оценки. Автор рассматривает структуру различных существующих словарей с различным диапазоном шкалы измерения силы, а также указывает на разницу в объеме словников [Ibid.: 329-330]. В части трактовки модификаторов оценки Табоада придерживается традиционного принципа полного разграничения между оценочными словами и их модификаторами (см. п. 2.4.1.5.3). Также она проводит анализ взаимосвязи оценочности с инверсией и маркерами субъективности. Описывая возможности вывода результирующей оценки для объекта по документу, автор констатирует, что возможности использования фактора когерентности существенно ограничены современным уровнем развития технологий автоматического определения дискурсивной структуры текста [Ibid.: 338].

В заключении своего обзора Табоада проводит мини-обзор новых направлений в рамках автоматического извлечения мнений. Из них она выделяет автоматическое определение суицидальных наклонностей, оценку эффективности учебного процесса, предсказание итогов политических кампаний, а также идентификацию сарказма и псевдо-рецензий. К рассматриваемым языкам, которые наряду с английским становятся объектом исследований в области извлечения мнений, автор относит арабский, китайский, французский, немецкий и испанский. Необходимо отметить, что данный список языков не является не только не исчерпывающим, но и уступает более ранним работам, таким как [34], где авторы дополнительно упоминают итальянский, нидерландский, японский и тайваньский. Примечательно, что ни в одном из известных нам обзоров по автоматическому извлечению мнений не упоминается русский язык.

[126] помимо одиночных слов также указывается возможность использования словосочетаний

Авторы приведенных выше обзоров Табоада и Штеде разработали модель SO-CAL[127] [49]. Основой данной модели является словарь оценочной лексики, учитывающий части речи и некоторые типы модификаторов оценки. В статье указывается, что оценочная сочетаемость представляет собой комбинацию из четырех вариантов (исходная оценка, ее отрицание, усиление и уменьшение) [Ibid.: 269]. В словаре представляемой авторами системы сочетаемостные свойства представлены усилителями, уменьшителями оценки, маркерами отрицания и ирреальности.

Структурно словарь состоит из отдельных словарей для различных частей речи (прилагательные, наречия, существительные и глаголы[128]). Оценочная сила представлена по шкале ±5 (без наличия нулевого класса). Интересным представляется способ автоматической генерации наречий из прилагательных (ср. метод морфемного синтеза, описанный в Главе 3). Процедура генерации заключалась в следующем. Авторы анализировали корпус на наличие наречий на -ly, при совпадении основы наречия с прилагательным из списка наречие добавлялось в словарь с оценкой, которая наследовалась от прилагательного. Получившийся словарь наречий проверялся вручную. Интересным решением представляется включение коллокаций в словарь, соответствующий части речи начального слова в коллокации.

Словарь модификаторов оценочности (усилителей и уменьшителей) может пересекаться по содержанию с полновесными оценочными словарями (например, прилагательное *pretty*, одновременно являющееся и положительным прилагательным и уменьшителем оценочности). Кроме словарного подхода при определении усиления оценки также применяются формальные эвристики, например верхний регистр, воклицательность предложения и дискурсивные коннекторы противопоставления [Ibid.: 276].

---

[127] Словарь доступен по адресу: https://github.com/DrOttensooser/BiblicalNLPworks/tree/master/SkyDrive/NLP/CommonWorks/Data/Opion-Lexicon-English/SO-CAL

[128] всего 5042 слова, из них прилагательные составляют 2252 вхождения (44,6%).

Среди других работ, посвященных лингвистическим аспектам автоматического извлечения мнений, мы выбрали работы двух исследовательских групп (Цюрихского и Оксфордского университетов), т.к. они, по нашему мнению, наиболее подробно описывают вопросы оценочной сочетаемости. Рассмотрим сначала работы Цюрихской группы компьютерной лингвистики[129].

Из многочисленных работ, выполненных под руководством проф. М. Кленнера, мы выделим следующие — [22; 42; 21]. На наш взгляд, эти работы отражают эволюцию в подходах данной группы к решению задачи оценочной сочетаемости.

В первой из перечисленных статей описываются сочетаемостные правила вывода общей оценки для позитивных, негативных и нейтральных слов. Предлагаемый подход является инженерным, т.е. эмпирически выведенным [22: 181]. Существенным аспектом работы мы считаем жесткость правил, которые в отличие от словарей являются языконезависимыми. Также усилители оценки, по мнению авторов, не могут нести собственно оценки. Общее количество сочетаемостных правил в системе равно 70. Авторы признают ключевым недостатком подхода «нейтрализующий эффект» комбинации «позитив+негатив», которая в рассматриваемом подходе ведет к априорно нейтральной оценке. В качестве результатов анализа ошибок в статье приводится список вопросов теоретического характера, решение которых должно способствовать дальнейшему повышению качества извлечения мнений. К данным вопросам относятся [Ibid.: 183—184]: модель мира (к какому оценочному классу относить фразы типа «идеальный шпион»); неоднозначное поведение некоторых слов;[130] фразеологическое и метафорическое употребление слов; ирония и имплицитные мнения; лексическая многозначность (например, «дешевое лечение» — недорогое или неэффективное).

---

[129] порядок рассмотрения данных групп связан с тем, что на взгляды автора оказала сильное влияние лекция руководителя группы автоматической обработки текстов Цюрихского университета М. Кленнера по проблемам оценочной сочетаемости лексики (2009 г.).

[130] речь идет о неоднозначных усилителях, которые в зависимости от присоединяемого слова либо усиливают оценку, либо меняют на противоположную.

В работе следующего периода [42] делается попытка создания теории оценочной композициональности именных групп. В работе рассматривается восходящий подход к определению оценки предложения через определение оценки именных групп. Авторы критикуют подход группы Оксфордского университета (см. ниже) за использование ручного составления правил и предобработки низкого качества. В части словаря от описанной в предыдущей работе системы PolArt существенных отличий нет, за исключением увеличения относительной доли прилагательных. Следует отметить, что целью данной работы является отказ от трудоемкого процесса создания всеобъемлющей теории композициональности вручную.

В последней на данный момент публикации Цюрихской группы [21] описывается дальнейшее развитие гибридного подхода к извлечению мнений по модели, которая была предложена в рассматривавшейся работе [42]. М. Кленнер анализирует глагольные группы и их влияние на сочетаемостную оценку. Он указывает на принципиальную разницу между различными видами оценки (оценки-факты, эмоциональные и моральные оценки). Автор вводит понятие оценочного глагольного фрейма [21: 135] и группирует глаголы, входящие в идентичные фреймы, в классы. Всего на момент публикации статьи глагольных классов было 12 (всего описано 80 глаголов). К достоинствам статьи мы относим разграничение глаголов по сочетаемостным классам [ср. 46] и автоматическое заполнение некоторых классов при помощи учета отрицательных конструкций.

Из работ группы Оксфордского университета рассмотрим следующие — [37; 36; 38]. Выбор относительно старых работ объясняется тем, что в 2010 году из группы выделилась компания TheySay Ltd[131], теоретические принципы и алгоритмы которой являются коммерческой тайной.

Первая из перечисленных статей описывает подход к определению оценки предложений и клауз, для которых (по мнению авторов) не подходят статистические методы обработки текстов. В качестве входных данных в описываемом решении используются результаты синтаксического анализа текста.

[131] http://www.theysay.io/

Алгоритм навигации по дереву синтаксического разбора движется от листьев к вершине, таким образом, вначале анализируются именные группы, затем глагольные. Авторы отмечают, что прилагательные и наречия чаще всего выступают в качестве модификаторов оценки, но их следует отличать от другой функциональной лексики, например артиклей. Анализ ошибок [37: 381—382] показывает, что на долю семантического компонента приходится более 36% всех ошибок, ещё 24% вызваны прагматическими факторами, остальные 40% — ошибками более низкого уровня, в т.ч. орфографическими.

В следующей из анализируемых работ предлагается метод извлечения мнений относительно множества объектов в рамках одного и того же предложения. Важным наблюдением авторов является необходимость анализировать не только эксплицитно выраженные мнения, но также и имплицитные оценки и коннотации [36: 258]. При подсчете оценки в работе используется понятие грамматического расстояния, при помощи которого производится определение релевантности оценки к анализируемому объекту.

В последней из рассмотренных работ предлагается гибридный подход к извлечению мнений, в котором лингвистическая предобработка сведена к минимуму [38: 37]. Данный подход заключается в использовании словаря на всех этапах (и обучение модели, и декодирования), в то время как эвристические правила оценочной сочетаемости применяются только при обучении статистической модели. Получившаяся в результате модель анализа не уступает модели на правилах, но выигрывает у нее за счет более быстрого создания готовой модели анализа.

Необходимо отметить, что первая статья Оксфордской группы критикуется в работах Цюрихской группы [22; 42] за моделирование нормативного языка и игнорирование значимости предобработки. В целом, это свидетельствует о близости подходов в идеологической части и о конкуренции на коммерческом и академическом уровнях.

В заключение, можно констатировать, что в части основных тенденций в применении лингвистических методов к автоматическому извлечению мнений

сохраняется главенство словарного компонента, а также стремление к отказу от чисто эвристических или статистических методов в пользу гибридных. Тенденция к гибридным методам, по нашему мнению, объясняется невозможностью описать в рамках существующих лингвистических моделей оценки пограничные случаи сочетаемости (т.н. *polarity conflicts*), а также более высоким быстродействием статистических алгоритмов[132]. К недостаткам существующих методов можно отнести игнорирование семантических классов слов, а также отсутствие этапа фильтрации неоценочных конструкций и инкорпорации неравноправно-оценочных глаголов (см. п. 2.4.1.5.3). Отдельную проблему составляют имплицитные оценки, которые, к сожалению, находятся вне рассмотрения существующих лингвистических подходов.

2.4.1.5.2. Сила оценки[133]

Как уже отмечалось выше, одной из наиболее часто используемых классификаций оценочной лексики является классификация на оценочные и оценочно-усилительные слова. Оценочно-усилительные слова в свою очередь делятся на усилители (положительное усиление) и уменьшители (отрицательное усиление). Для описания взаимодействия оценочных и оценочно-усилительных слов используется оценочная шкала. Оценочная шкала, таким образом, является неотъемлемой частью систем автоматического извлечения мнений. Сама шкала практически никогда не входит в систему, представляя собой инструкцию по разметке исходного корпуса, на основе которого составляется словарь системы. В то же время, элементы шкалы всегда присутствуют в словаре.

В ранних работах шкала состояла из трех столбцов — отрицательной, положительной и нейтральной оценок. Дальнейшие исследования показали значительно большее варьирование степеней оценок. Увеличение степеней оценок (до 5) привело к незначительному улучшению качества работы системы [40: 25; 33–34].

---

[132] а также в определенной степени бо́льшим интересом научного сообщества в области компьютерной лингвистики (в значительной степени инженеров и математиков) к методам машинного обучения.

[133] данный раздел представляет собой переработанные варианты публикаций автора [Куликов, 2010, 2010а].

В более поздних работах было отмечено, что не только лексическая полярность влияет на оценку текста в целом. Было обнаружено, что синтагмы принимают оценочные значения не существительных, а прилагательных. Работы по лексической сочетаемости и ее влиянию на оценку синтагм привели к введению в практику разработчиков систем автоматического извлечения мнений понятия шифтеров (переключателей оценки), усилителей и уменьшителей оценок.

На современном этапе появились работы, призывающие к пересмотру параметров оценки [22]. В качестве параметров для шкалы оценки используются данные синтаксического и дискурсивного анализа [40: 35; 52].

Оценочные шкалы должны учитывать информацию на всех текстовых уровнях: графическом, морфологическом, синтаксическом и текстовом. Мы предлагаем использовать шкалу из 21 столбца (10 отрицательных, 10 положительных и 1 нейтральный) [121: 36]. При этом лексические свойства слов, влияющие на оценку, должны влиять не более чем на 5 столбцов, остальные столбцы должны формироваться на основе правил перехода, заносимых в матрицу переходов. Переход по оценочной шкале моделирует процесс восходящего синтаксического анализа. Графические факторы учитываются сразу после словарных (лексических). К морфологическим свойствам относятся маркеры превосходной степени и отрицательные приставки. Синтаксические и текстовые факторы влияют только при их наличии в тексте. При этом своего максимального значения (±10) оценка для объекта анализа может достичь только в случае присутствия в анализируемом документе текстовых факторов. В качестве формализма, описывающего правила перехода, можно использовать контекстологический словарь [140; 138], оперирующий лексическими функциями Magn, Bon, Plus, Minus, Anti. Более подробно подобный контекстологический словарь описан в нашей работе [116]. В данном исследовании мы отказались от контекстологического подхода из-за его излишней трудоемкости.

2.4.1.5.3. Оценочные классы лексики

Классификация оценочной лексики представляет собой одну из наиболее сложных проблем в области автоматического извлечения мнений. Традиционное деление (Табоада, Кленнер) оценочной лексики на 6 классов (положительный, отрицательный, нейтральный, усилительный, уменьшительный и класс переключателей) не в полной мере отражают лингвистическую действительность. Данный факт можно проиллюстрировать на упоминавшемся примере дублирования слова *pretty* в словаре SO-CAL.

Таким образом, необходимо использовать более дробную классификацию, которая бы учитывала сочетаемостные свойства слов.[134] Нам представляется необходимым ввести дополнительные классы негативных усилителей, позитивных усилителей[135], субъектно-нейтральных слов и словосочетаний, а также однореферентных слов. В результате полный набор классов оценочной лексики будет следующим:

1. положительные слова и словосочетания;
2. отрицательные слова и словосочетания;
3. нейтральные слова и словосочетания;[136]
4. усилители оценки;
5. уменьшители оценки;
6. оценочные переключатели (шифтеры);
7. положительные усилители;
8. отрицательные усилители;
9. субъектно-нейтральные слова и словосочетания;
10. субъект-позитивные глаголы;
11. субъект-негативные глаголы;
12. однореферентные слова.

2.4.1.5.4. Учет частеречной информации в словаре оценочной лексики

[134] такой подход полностью соответствует принципам интегрального описания языка, но в нашем случае, вместо семантики слова учитывается прагматика.
[135] нами не было обнаружено соответствующих классов для уменьшительных слов и словосочетаний.
[136] не отражаются в словаре, за исключением особых случаев, о которых речь пойдет в п. 2.4.1.5.5.

Прежде чем перейти к рассмотрению особенностей представления словосочетаний в оценочных словарях, рассмотрим принципы учета частей речи. Как упоминалось выше (п. 2.4.4.5.1), разграничение словарей оценочной лексики по частям речи используется в ряде систем, например PolArt [21] и системе «Аналитический Курьер» [46].

Необходимость учета части речи при классификации оценочной лексики объясняется различными сочетаемостными свойствами, которыми обладают разные части речи. В качестве примера можно сравнить следующие оценочные классы: положительные усилители, отрицательные усилители и субъектно-нейтральные слова и словосочетания. В первые два класса могут входить только прилагательные и наречия, а в последний — только глаголы и девербативы. Определенные части речи могут содержать дополнительные средства усиления (прилагательные), в то время как для имен существительных такое невозможно. Кроме того, глаголы в различных залоговых и неличных формах могут иметь оценочное управление, которое будет отличаться от оценочного управления личных форм.

Специальным оценочным классом для глаголов (который не включен в предыдущую классификацию) является класс донора оценки. Данный класс состоит из глаголов нескольких лексико-семантических классов, в первую очередь, движения. Сами по себе эти глаголы не несут оценочной информации, но в отдельных контекстах (особенно в форме отрицания) выступают в функции негативных глаголов. В качестве примера можно привести предложение «*Машина не поедет, т.к. бензин кончился*».

Существует два способа представления частеречной информации в оценочном словаре. Первый заключается в создании отдельных списков слов с указанием на часть речи (и соответствующий оценочный тип). Данное указание может быть выражено не эксплицитно, а при помощи настроек в программном коде. Второй подход заключается в кодировании указаний на часть речи в специальном поле базы данных. Первый способ обеспечивает возможность более

быстрого пополнения словника, в то время как второй способствует минимизации ошибок[137].

2.4.1.5.5. Учет фразеологических и терминологических сочетаний в оценочном словаре

Известно, что оценка терминологических и фразеологических сочетаний не является простой суммой оценки каждого компонента рассматриваемого словосочетания. Именно поэтому терминологические и квазитерминологические словосочетания следует заносить в словарь целиком. Тип оценки словосочетания при этом будет зависеть от типа синтаксически главного слова в словосочетании.

Фразеологические выражения бывают нескольких типов. С точки зрения выражаемой ими оценки можно выделить фразеологизмы с вариативной частью и нерасчленяемые фразеологизмы. Фразеологизмы первого типа заносятся в словарь с особой структурой, в котором вариативная позиция маркируется спецсимволом. В качестве примера подобного кодирования можно привести фразеологизм «*как баран на новые ворота*» для следующих предложений.

Как Барак на новые ворота[138]

Турция: как Туран на новые ворота[139]

В словаре фразеологизм будет представлен в виде записи: *как * на новые ворота*.

Нерасчленяемые фразеологизмы следует заносить в отдельный словарь, т.к. они могут влиять на оценку только в случаях сложного предложения или последовательного следования за текстом, содержащим объект анализа.

Описав все данные, которые учитываются в оценочных словарях, рассмотрим схему взаимодействия различных словарей оценочной лексики между собой (см. Рисунок 12). На первом этапе анализа мнений используется словарь (база данных) оценочных словосочетаний, содержащий различные терминологические и фразеологические сочетания. Далее опционально

[137] при условии, что множественные значения для одного и того же слова запрещены.
[138] http://beseder.ru/news/entryid/436
[139] http://football.ua/countrieselse/141135-gruppa-a-turcyja-kak-turan-na-novye-vorota.html

применяется словарь специализированной лексики (если к системе общего назначения подключены отдельные предметные области со своей оценочной спецификой). На следующем этапе используется общий оценочный словарь, с помощью которого размечаются оценочными тегами слова без разметки. Последним этапом является использование слов, включенных в правила оценочной сочетаемости.

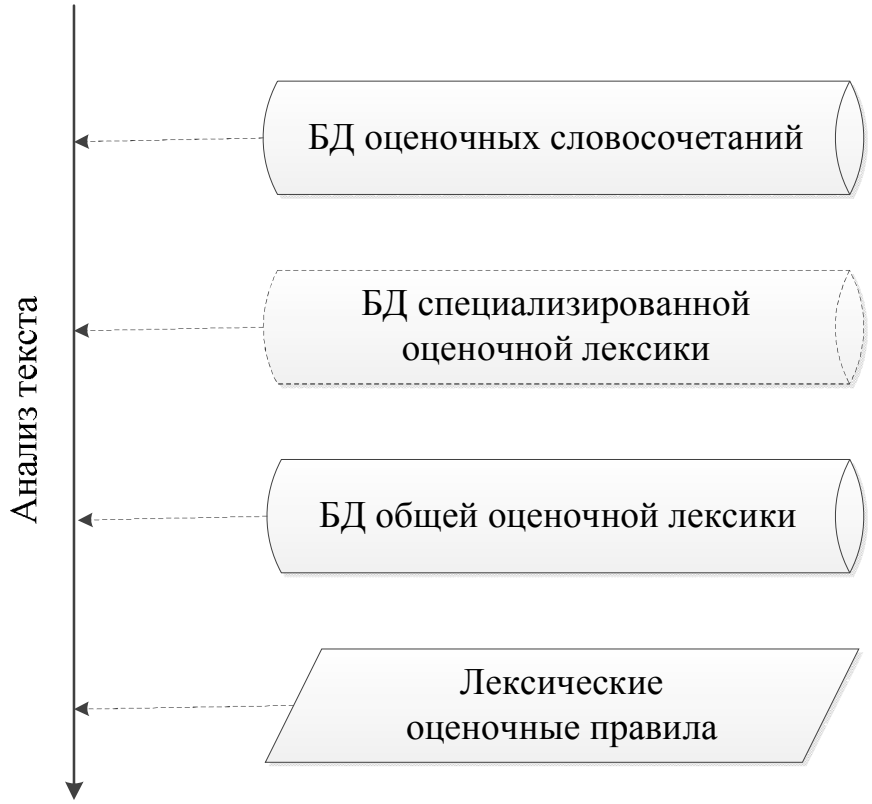

Рисунок 12. Последовательность применения словарей в системе автоматического извлечения мнений общего назначения

2.4.1.6. Правила сочетаемости оценочных слов

Правила сочетаемости оценочных слов оперируют тремя типами категорий — синтаксическими, семантическими и прагматическими. Необходимо сделать отступление касательно использования морфологической (словообразовательной) информации. На этапе применения правил оценочной сочетаемости словообразовательная информация уже учтена в оценочном словаре и попадает в категорию прагматической информации.

Использование синтаксической информации предопределяет уровневую модель применения правил. Таким образом, прагматический анализ в предлагаемом нами методе будет восходящим. В данном параграфе мы рассмотрим собственно правила сочетания именных, предложных и глагольных групп (п. 2.4.1.6.1). Далее в качестве частного случая каждой из названных групп будут проанализированы лексикализованные правила, в рамках которых мы опишем проблему переключателей (шифтеров) и уменьшителей оценки (п. 2.4.1.6.2). В заключительной части параграфа приводятся дискурсивные правила, а также указаны их связи с компонентом причинно-следственных отношений (п. 2.4.1.6.3).

2.4.1.6.1. Уровневая модель правил

Уровневая модель правил, предлагаемая нами, во многом сходна с моделями М. Кленнера и С. Пульмана, кроме того, сильное влияние на ее эволюцию оказали взгляды А.Н. Соловьева. От перечисленных подходов наш подход отличает набор оценочных классов (см. 2.4.1.5.3), а также правила, учитывающие семантические классы слов.

Первым этапом является оценочное связывание элементов внутри именных групп [ср. альтернативный подход в Куликов, 2010]. Классификация именных групп при этом следующая: группы прилагательного[140], группы существительного и генитивные группы.

В группе прилагательного происходит оценочное связывание прилагательных и наречных модификаторов. Наречные модификаторы при этом могут относиться как к однозначным классам (положительные, отрицательные, усилительные или нейтральные), так и к двойственным классам (положительные или отрицательные усилители). При объединении наречных модификаторов однозначных классов с прилагательными правила определения результирующей оценки просты:

1) AdvPos + AdjPos => AdjPos;

[140] сюда же примыкают причастия.

2) AdvNeg + AdjNeg => AdjNeg;
3) AdvIntens + AdjPos => AdjPos;
4) AdvIntens + AdjNeg => AdjNeg;
5) AdvIntens + AdjIntens => AdjIntens.

В случае потенциального конфликта оценок (AdvNeg + AdjPos; AdvPos + AdjNeg) алгоритм должен переводить первое слово в двойственный класс (динамическое изменение типа). Например, предложения «*На сей раз вот этот кошмарно красивый триллер, который посмотрела по наводке*[141]» (AdvNeg + AdjPos) и «*Восхитительно ужасный грибочек!*[142]» (AdvPos + AdjNeg) в действительности являются примерами усилителей.[143]

Иными словами, двойственный класс оценки является не словарным (статическим), а функциональным (динамическим). Для наречий это проявляется следующим образом: в естественных контекстах (к таковым мы относим глагольные) наречие выступает как полноценное оценочное слово (*поступил кошмарно*; *рисует восхитительно*), а в остальных случаях ведет себя как оценочный усилитель.

Ещё одним случаем оценочной сочетаемости в рамках группы прилагательного является употребление нескольких прилагательных. При этом прилагательные могут выступать как однородные члены предложения в именной части составного сказуемого или как определения при одном существительном. При совпадении типа оценки у прилагательных суммирующая оценка будет также равна этому типу. В случае конфликта оценок выбор итоговой оценки осуществляется при помощи конфигурации. При этом используются данные о типе оценки (этические, эстетические и т.п.), а из файла конфигурации берется информация о более важной из данного набора оценок. В случае невозможности определения оценки данным способом в качестве итоговой выбирается оценка наиболее левого прилагательного.

---

[141] http://merelana.livejournal.com/754391.html
[142] https://illustrators.ru/illustrations/766397
[143] здесь имеет место наложение различных типов оценок, но в рамках шкалы «положительный / отрицательный» указанные оценочные оттенки не имеют большого значения.

Группа существительного представляет собой сочетание определения и существительного. В случае совпадения полярностей оценок у всех элементов конструкции совокупная оценка группы существительного также будет равна данной оценке. При отсутствии у одного из элементов оценочного свойства общая оценка будет равна оценке того элемента, у которого она выражена[144]. В случае конфликта оценок выбор итоговой оценки аналогичен определению оценки для нескольких прилагательных. В случае невозможности определения оценки данным способом в качестве итоговой выбирается оценка прилагательного.

Оценка генитивных групп (т.е. состоящих из двух элементов, вторым из которых является существительное в генитиве) зависит от семантического класса каждого из слов. При наличии генитивных зависимых у девербативов и деадъективов в качестве модификаторов выступают именно данные классы слов [ср. 46]. В случаях оценочных конфликтов оценочный приоритет будет у существительных-модификаторов. Отдельный случай составляет совпадение семантического типа у обоих элементов именной группы, например: *Увы, сколько бы мы ни ставили под сомнение наличие общечеловеческих универсалий, единственная альтернатива в данном случае, – расовая теория и оправдание геноцида*[145]. В таком случае также применяются настройки конфигурации и/или более дробная семантическая классификация. В случае с детальной семантической информацией (учитывающей априорные оценки у событий и их типов) позитивный модификатор[146] для негативных событий ведет к отрицательной оценке всей группы. При наличии нескольких вложенных генитивных последовательностей производится восходящий анализ. Например, в предложении «*Отрицание геноцида армян не является преступлением, решил конституционный совет Франции*[147]» анализ подчеркнутой генитивной группы сначала будет проводиться для последовательности *геноцид армян*, а затем для всей последовательности *отрицание геноцида армян*.

---

[144] Такой элемент иногда называется «оценочным донором».
[145] http://n-europe.eu/content/?p=2814
[146] при этом не все позитивные девербативы приводят к такому результату. Подобные различия задаются в конфигурации модели оценки.
[147] http://www.gazeta.ru/politics/2012/02/29_a_4016413.shtml

Следующим этапом оценочного связывания является определение оценки для предложных конструкций. Для большинства предлогов данный этап не представляет сложности, т.к. предлоги, чаще всего, нейтральны. Ненейтральные предлоги указываются в лексикализованных правилах, которые рассмотрены в следующем параграфе (2.4.1.6.2).

Последним этапом оценочного связывания в рамках простых клауз является определения оценки для глагольных групп[148]. Этап подразделяется на следующие подэтапы, в которых моделируются следующие объекты — составное глагольное сказуемое, глагольная группа с наречием, составное именное сказуемое.

В рамках составного глагольного сказуемого наиболее значимым оценочным элементом является зависимое слово (чаще всего, инфинитив). При этом оценка модальных глаголов с зависимыми компонентами относится лексикализованным правилам. Примеры, приведенные ниже, демонстрируют различные сочетаемостные модели для глагольной вершины *решить* и зависимого элемента *опередить*.

В то время как по микроблогам звезд прокатилась волна взломов, Дима Билан решил опередить мошенников.[149]

Но медучерждение решило опередить событие.[150]

Японцы, как пишет портал Slash Gear, решили его опередить.[151]

Мы поняли, что на трассе мы точно не ***сможем*** *обогнать пилотов команды Lotus, поэтому мы решили попытаться опередить их за счет стратегии.*[152]

Результирующая оценка для глагольных групп с наречиями определяется по правилам, которые схожи с правилами определения оценки группы прилагательного с зависимым наречием. Наиболее значимой моделью является сочетание нейтрального глагола (например, семантического поля *действие*) с

[148] Параграф опирается на работу автора [112]
[149] http://www.vokrug.tv/article/show/Dima_Bilan_zashchitil_svoi_mikroblog_ot_moshennikov_49850/
[150] http://www.kp.ru/daily/26441/3312783/
[151] http://www.infox.ru/auto/car/2015/10/02/Na_dorogah_YAponii_p.phtml
[152] http://f1report.ru/news/f1-16251.html

оценочным наречием. При этом позиция наречия относительно глагола не имеет значения (см. примеры ниже).

Автомобиль с номерами "ААА" нагло едет по обочине.[153]

Затем общий план - автобус уже едет правильно, движение правостороннее.[154]

Как и для составного глагольного сказуемого, для составного именного сказуемого наиболее значимым элементом является зависимая (именная часть), например: *Но вот мне лично стало ясно: Хрущ вообще был дурак и хуже дурака, а в данном случае он сделал какую-то совсем уж гадость.*[155]

При определении оценки глаголов целесообразно учитывать залог глагола и его оценочные модели управления. Особенно важно различение залогов для глаголов следующих оценочно-сочетаемостных типов: субъектно-нейтральных, субъект-позитивных и субъект-негативных глаголов. Для глаголов данных типов изменение залога с действительного на страдательный ведет к изменению оценочного типа на противоположный, что вызвано их ориентированностью на семантический, а не синтаксический субъект. Рассмотрим парные примеры для таких глаголов:

*Матч не разочаровал двух красоток - **ПСЖ** победил, а звездный нападающий Златан Ибрагимович вошел в историю команды.*[156] (субъект-позитивный глагол)

*Потому, что **католическая церковь** в них была побеждена протестантизмом и "реформирована".*[157] (субъект-негативная глагольная группа)

***Экс-премьер Украины** раскритиковала социально-экономическую политику Киева <...>.*[158] (субъект-нейтральный глагол)

*Ранее ANZ был раскритикован **экологами** за финансирование угольной шахты в Новом Южном Уэльсе.*[159] (объект-нейтральный глагол)

---

[153] http://www.e1.ru/news/spool/news_id-407456.html
[154] http://www.kuban.kp.ru/daily/25692.3/895111/
[155] http://4pera.ru/news/feysbuchnye_truth/pochemu_krym_russkiy_poluostrov/
[156] http://gigafootball.net/news/89141
[157] http://www.nakanune.ru/articles/110929
[158] http://www.segodnia.ru/news/167271

Рассмотрим способы определения оценочных моделей глагольного управления. Для этого проанализируем понятия синтаксической сочетаемости и глагольного управления в прикладной лингвистике.

«Синтаксическая сочетаемость слова — совокупность и свойства потенциально возможных при нем синтаксических связей, набор и условия реализации синтаксических связей» [109: 81]. К понятию синтаксической сочетаемости примыкает понятие модели управления. Наибольшее значение модели управлении получили в теориях языка, в которых синтаксис представлен грамматикой зависимостей, в частности «Теория Смысл<=>Текст». Традиционно под моделью глагольного управления понимается свойство глагольных предикатов присоединять актанты в определенных падежах.

К настоящему времени в отечественной науке не выработано универсального лексикографического представления моделей управления. Так, И.А. Мельчук [145: 104] на материале английского языка приводит следующий пример неоднозначной модели управления, в которых позицию второго актанта одного и того же глагольного предиката занимают слова с разными семантическими ролями: supply food to the peasants vs. supply peasants with food. Мельчук отмечает, что «необходима Модель Управления, где привязка для заглавной лексемы L задается непосредственным перечислением» [Ibid.]. Таким образом, модель управления глагола to supply имеет следующий вид: supply (CommAct: Anim|Inan)(CommPass: Inanim|Anim), где CommAct — актуализированный участник коммуникативного акта, а CommPass — неактуализированный участник, после двоеточия даются свойства актантов.

Особое внимание моделям глагольного управления уделяется специалистами по синтаксическому анализу и синтезу текстов. Для задач анализа текста реализация той или иной модели управления позволяет разрешить достаточно большой процент омонимичных ситуаций. Подобные задачи решаются и при синтезе текстов, но следует сделать оговорку, что для систем

---

[159] http://bankir.ru/novosti/20151006/the-guardian-bankovskaya-gruppa-anz-vlozhit-10-mlrd-v-chistuyu-energetiku-10112620/

симуляции диалога (в т.ч. при создании чат-ботов) данный подход может привести к ряду негативных последствий.

Как уже отмечалось в Главе 1, впервые идея необходимости учета моделей глагольного управления для нужд автоматического извлечения мнений была высказана в работе Заенен 2004 г. [60]. В большинстве работ, опирающихся на идеи Заенен, рассмотрению подвергаются лишь отдельные значения слов, оценка которых зафиксирована в семантических ресурсах типа WodNet. Для русского языка подобных ресурсов достаточного объема не существует, поэтому оценочные значения глаголов целесообразно задавать через их модели управления.

Наибольшее значение учет моделей глагольного управления имеет для разграничения фактов и оценок (см. например, работу [Арутюнова, 1988]). Самыми значимыми при таком разделении являются глаголы говорения, которые вводят субъективные высказывания (в том числе оценочные суждения). С другой стороны, глаголы говорения могут и непосредственно входить в состав оценочных высказываний в качестве их ядра.

Для выявления правил разграничения моделей мы провели исследование[160] двух оценочных глаголов (*негодовать* и *считать*) с разными моделями управления. Следует заметить, что некоторые девербативы ведут себя как полнозначные глаголы (с точки зрения оценочной структуры предложения). Поэтому мы рассматриваем модели девербативов (как полупредикатных слов) вместе с соответствующими им глаголами.

Глагол «негодовать» (273 вхождения) может выступать в качестве глагола, вводящего оценочное суждение. Приведем примеры[161] трех самых распространенных конструкций с данным глаголом.

1) « Туристы фактически стали заложниками спора хозяйствующих субъектов, что совершенно недопустимо!» — негодуют в Ростуризме.

[160] Исследование проводилось на материале русскоязычного корпуса газетных текстов, объем которого превышает 25,5 млн. словоупотреблений.
[161] Примеры взяты из исследовательского корпуса без привязки к оригинальному источнику.

2) Но насилие — предсказуемый и неизбежный результат в тех случаях, когда молодые люди сыты и живут в обществе, где их слишком много и где они негодуют на это самое общество, поскольку понимают, что оно не в состоянии их востребовать.

3) Римляне негодуют: Берлускони тянет резину.

4) Другие интернет-пользователи с негодованием констатировали, что "президент отплясывает на костях россиян", и обвинили Медведева в предательстве родины в связи с тем, что он танцевал под "American boy".

5) В итоге Асдераки присудила победу в розыгрыше австралийке, что вызвало негодование Уильямс.

В приведенных примерах отражены следующие модели управления: 1) (SubjSent) Verb (Adj), 2) (Subj) Verb (OpObj), 3) (Subj) Verb (SubjSent), 4) (Subj) NounVerbMod Verb (SubjSent), 5) (SubjSent) <AnaphoricRel>Verb NounVerbMod (SemAct).

Приведем расшифровки метаязыковых обозначений: SubjSent — предложение с субъективной оценкой, Verb — основной глагольный предикат, Adj — обстоятельственный компонент, OpObj — объект, на который направлено мнение, Subj — субъект высказывания, NounVerbMod — отглагольное существительное (девербатив), выступающее в роли модификатора оценки, AnaphoricRel — анафорическое отношение, SemAct — семантический актант (субъект оценочного высказывания в глубинно-синтаксической структуре).

Следует отметить, что модели 1), 3), 4) и 5) вводят оценочное суждение, оставаясь при этом нейтральными, и лишь модель 2) является собственно оценочной.

Глагол «считать» (и его модификации) составляют в общей сложности 24527 вхождений в корпус. Этот глагол помимо оценочного значения (наиболее частотного в газетном подъязыке) имеет и фактитивное значение («*проводить арифметические действия над объектами*»).

Приведем несколько примеров употребления разных моделей управления с данным глаголом.

1) С точки зрения аппетита к риску, это можно считать небольшой победой в неприятной войне.

2) Началом его сольной деятельности можно считать памятное выступление 14 октября 1983 года.

3) Проблем с ЖЖ не было, если только не считать споров с Антоном Носиком.

4) А это, как принято считать, обязательное условие для существования жизни.

5) Это позволило считать закон самой успешной социальной реформой.

6) Достаточно, если они будут уметь считать до ста…

Модели управления, представленные в примерах, следующие: 1) <…> ModVerb Verb (OpObj) (Adj) 2) (SemAct) ModVerb Verb (OpObj) (Adj) 3) (SubjSent) (Verb (OpObj)) 4) (SubjSent (PassVerb Verb)) 5) (Subj) PassVerb Verb (Obj) (OpObj) 6) (Subj) ModVerb Inf Verb (Number). Из всех конструкций 6) не является оценочной, а 4) вводит оценочное суждение. Если выйти за рамки нашей схемы, то можно разграничить оценочные значения от неоценочных для данного глагола тем, что у оценочных глаголов одним из актантов является существительное, обозначающее действие (победа, выступление, спор, реформа).

Оценка глагольной группы является доминирующей при определении итоговой оценки предложения относительно объекта (подходы М. Кленнера и А.Н. Соловьева). Данное решение базируется на вербоцентрической модели предложения и кажется лингвистически обоснованным. Такие подходы можно назвать синтаксически-ориентированными. Ключевым недостатком чисто синтаксически-ориентированных подходов является невозможность учесть скрытую оценку (т.е. оценку отдельных слов или именных групп). Поэтому в конфигурации модели анализа необходимо указывать типы оценок, которые представляются системой конечному пользователю. При этом дублирующиеся оценки не отображаются в итоговом результате. Например, если отрицательно охарактеризованный объект совершает негативное действие, то конечному пользователю видна только оценка действия. Но если такое же действие

совершается положительно охарактеризованным объектом, то итоговый результат будет зависеть от пользовательских настроек. Таким образом достигается моделирование приоритетов пользователя (ср. подход Р. Шенка, описанный в Главе 1). В случае отсутствия пользовательской конфигурации схема анализа следующая:

$$\exists\, Verb_{ev} \rightarrow Obj_{ev} = Verb_{ev} \;\&\; \nexists\, Verb_{ev} \rightarrow Obj_{ev} = (NP_{ev}|W_{ev}), \qquad (3)$$

где $Verb_{ev}$ — оценочный глагол; Obj_{ev} — итоговая оценка объекта анализа; NP_{ev} — оценка именной группы, содержащей объект анализа; W_{ev} — лексикализованная оценка объекта анализа.

Ключевым отличием от модели А.Н. Соловьева является отсутствие в клаузе не глагольного предиката вообще, а только оценочного глагольного предиката.

2.4.1.6.2. Лексикализованные правила[162]

Как отмечалось в предыдущем параграфе, некоторые типы оценочных отношений между фрагментами предложений следует определять не при помощи оценочных типов, а при помощи лексикализованных правил. Рассмотрим некоторые относительные характеристики оценочных слов.[163]

В первую очередь необходимо определить целесообразность разделения оценочной лексики между словарем и правилами. Для этого рассмотрим соотношение между лексикализованными и нелексикализованными правилами.[164] Из диаграммы (Рисунок 13) видно, что слова, входящие в лексикализованные правила, составляют только 0,003 % всех оценочных слов. При столь незначительном объеме правил возможна их оперативная ручная коррекция.

[162] Раздел представляет собой расширенную версию доклада автора [24].
[163] Исследование проводилось на материале ЛО системы *Аналитический Курьер*, количественные данные приведены по состоянию на 31.12.2013 г.
[164] Общее количество правил — более 170.

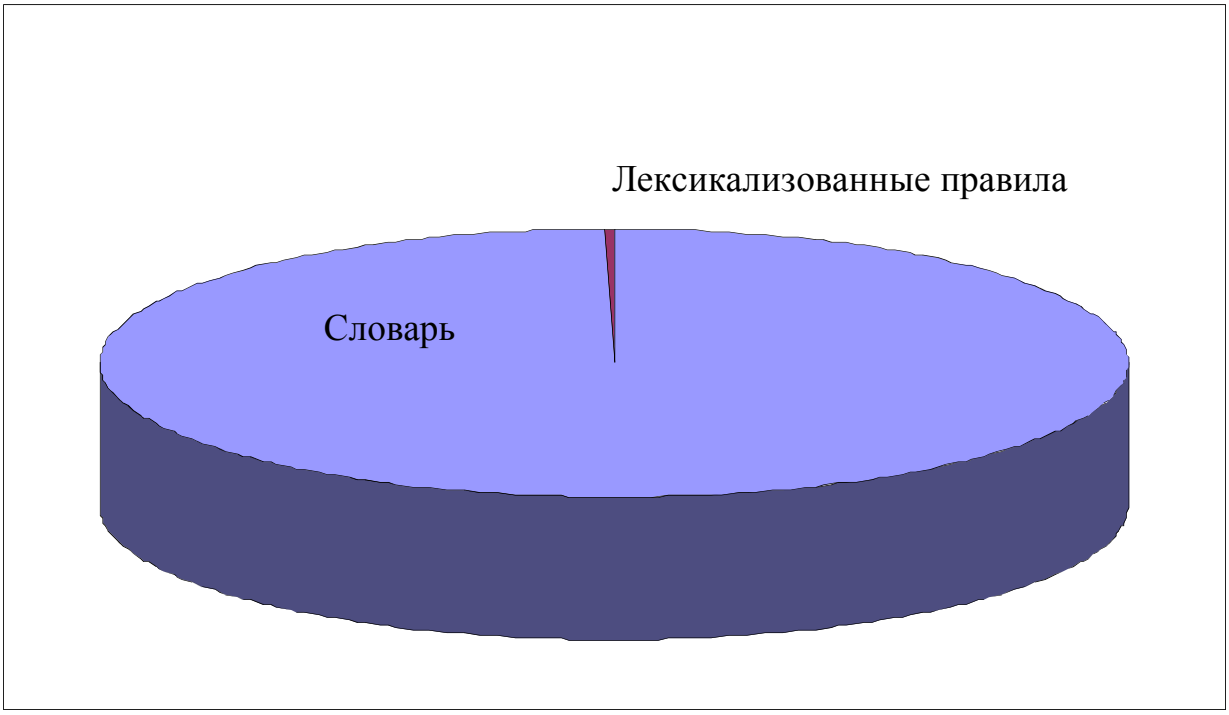

Рисунок 13. Соотношение словарных и лексикализованных типов

Для определения состава лексикализованных правил рассмотрим соотношение между двумя классами слов-модификаторов — усилителями и уменьшителями оценки. Из общего количества модификаторов только 4 % приходится на уменьшители оценки (Рисунок 14). Данный класс лексики является сильно периферийным и характеризуется тенденцией к закрытости класса. В силу этого уменьшители оценки можно компактно описать в правилах анализа.

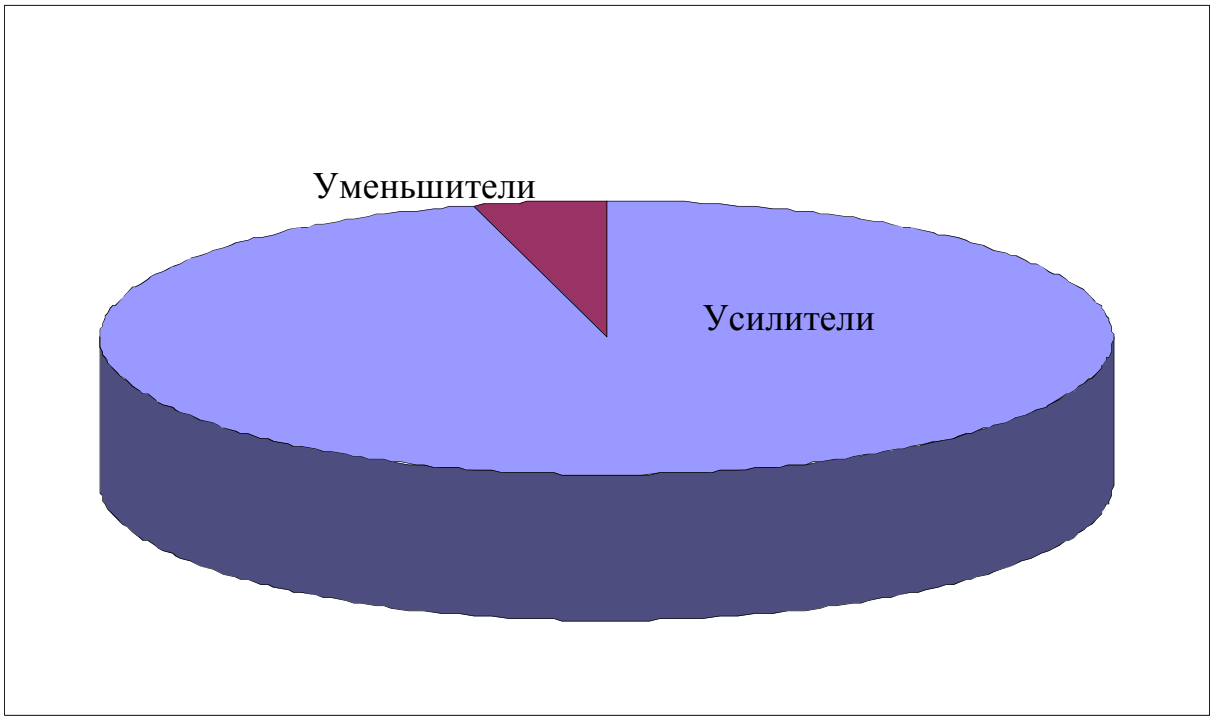

Рисунок 14. Соотношение усилителей и уменьшителей оценки

Другим закрытым классом оценочной лексики являются переключатели (шифтеры) оценки. В данный класс входят отрицательные частицы и некоторые коллокации. Оценочные переключатели существенно уступают уменьшителям оценки по количественному составу (Рисунок 15). Из общего количества элементов лексикализованных правил переключатели оценки составляют только 12 %, в то время как оставшиеся 88 % приходится на долю оценочных уменьшителей.

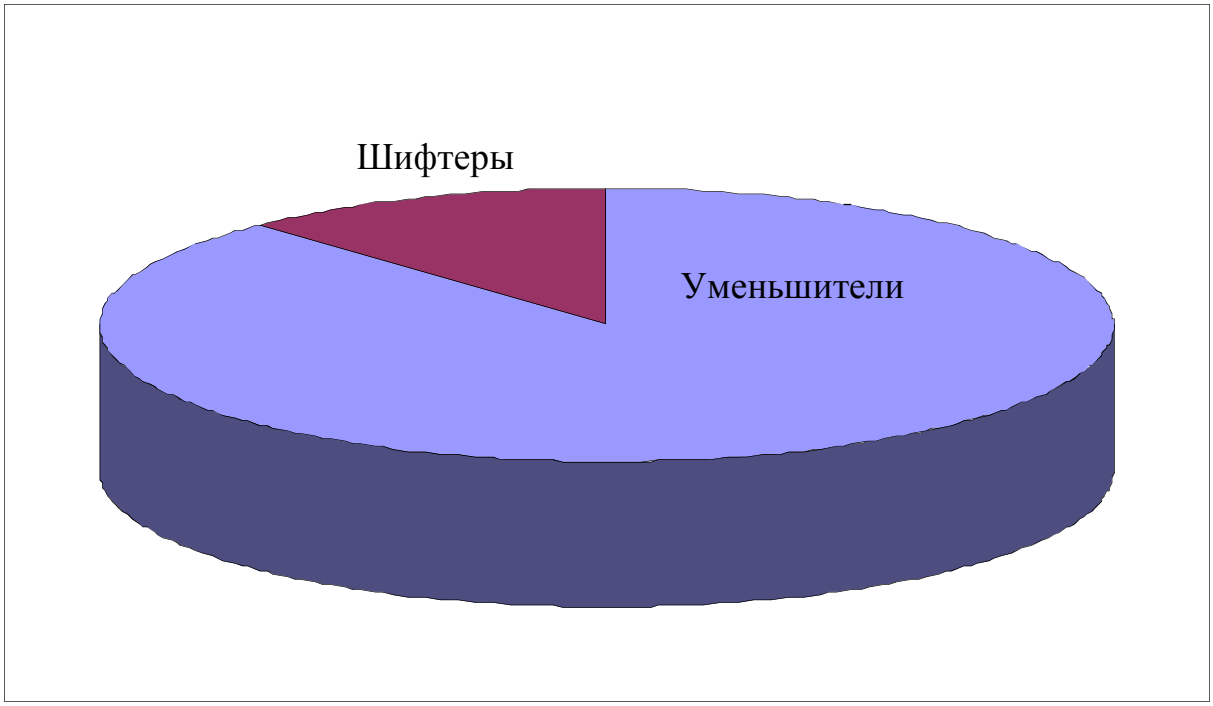

Рисунок 15. Соотношение шифтеров и уменьшителей

Несмотря на несущественное количество элементов, входящих в указанные классы оценочных слов, их роль в системе оценочных правил очень велика. 22 % всех правил содержат тот или иной лексикализованный компонент (Рисунок 16).

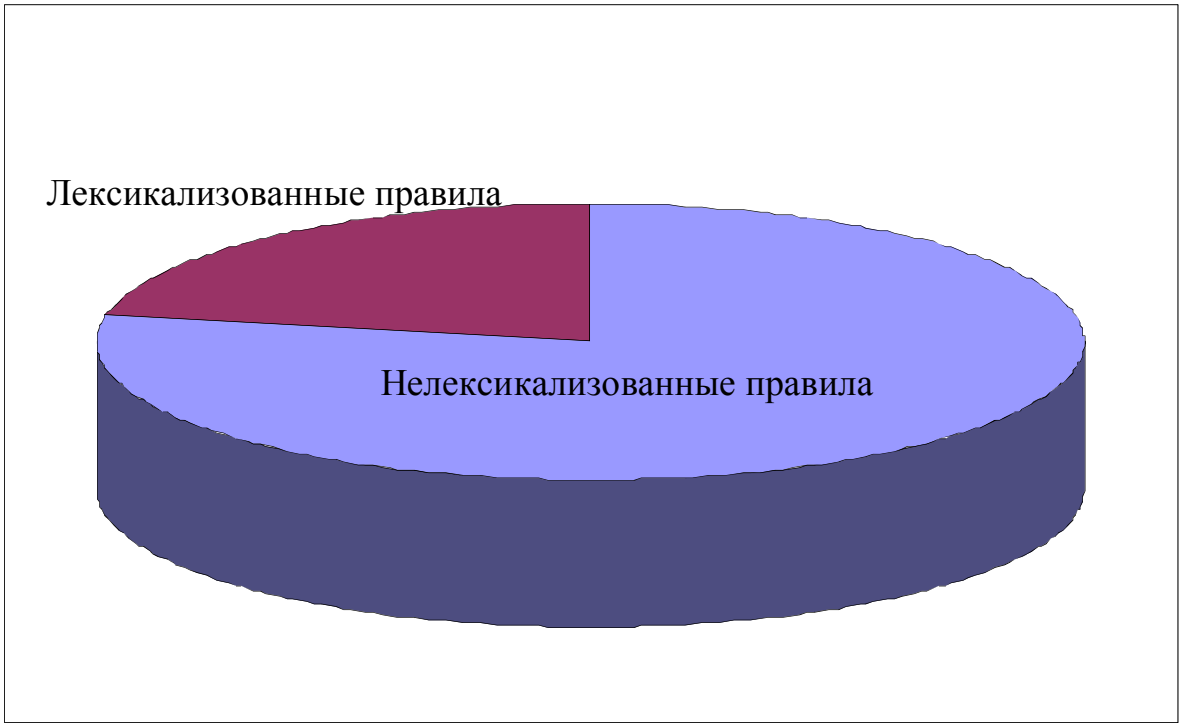

Рисунок 16. Соотношение правил с шифтерами и уменьшителями и без них

Как и следовало ожидать от класса, состоящего практически исключительно из частотной функциональной лексики, правил с переключателями оценки существенно больше. Данные правила составляют 95% от всех лексикализованных правил.

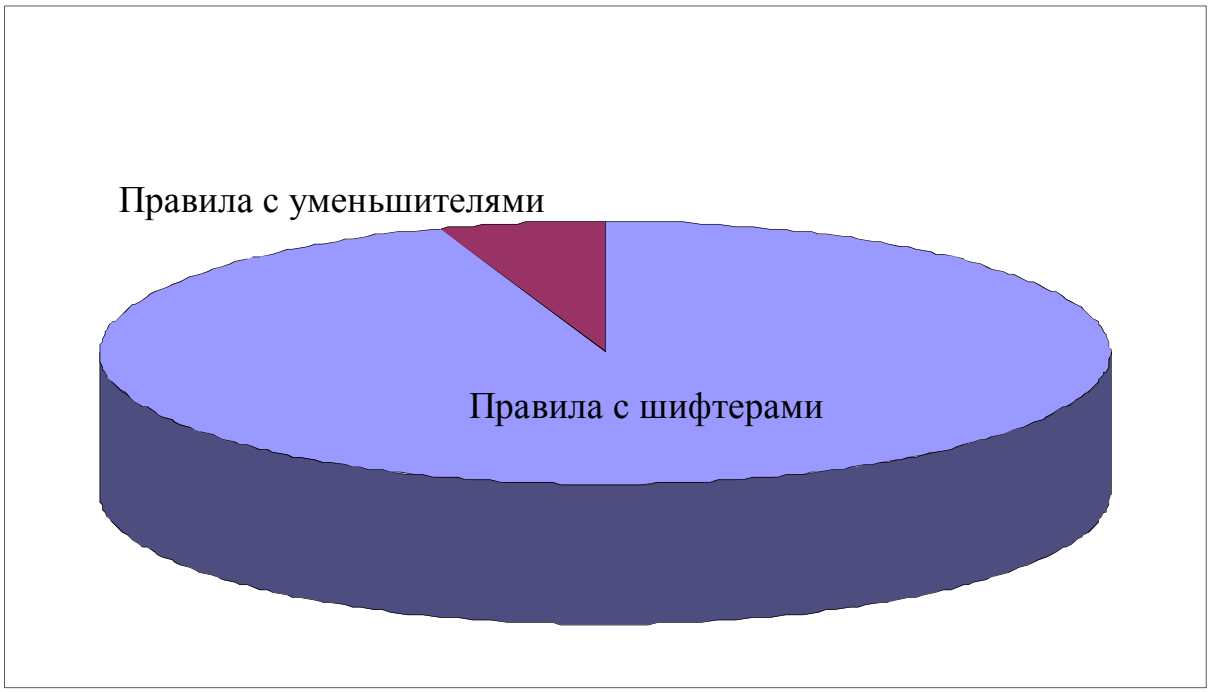

Рисунок 17. Соотношение правил с шифтерами и с уменьшителями оценки

Проведенное нами исследование демонстрирует, что представление уменьшителей и переключателей оценки в виде правил, а не словарей, направлено на более компактное описание оценочной лексики, за счет сокращения количества словарей, а также позволяет более оперативно исправлять возникающие ошибки, при помощи дополнительных ограничений на оценочную сочетаемость.

2.4.1.6.3. Дискурсивные правила[165]

Помимо лексикализованных правил, которые описывают случаи уменьшения или смены оценки, также существуют правила, описывающие лексикализованные межклаузные отношения. Вводные следственно-оценочные маркеры[166] такие как *хорошо, что* определяются на данном этапе. Следственно-оценочные маркеры относятся к трем сочетаемостным категориям — усилителям

[165] под дискурсивными правилами мы здесь понимаем оценочные отношения, которые влияют не непосредственно на оценку объекта анализа, а на ситуацию, в которую вовлечен объект.
[166] см. п. 2.4.1.4.

(если полярность оценки совпадает с полярностью зависимой клаузы), переключателям и донорам оценки.

Другой группой правил, которые можно отнести к дискурсивным, являются правила вывода результирующей оценки для объекта анализа на основе всех вхождений объекта в анализируемый текст. Для идентификации всех упоминаний объекта в тексте используются данные, полученные на этапе идентификации анафорических и кореферентных отношений. При этом возможны два варианта, которые зависят от типа модели объекта. При объектной модели сопоставляются все упоминания объекта. В случае расхождения полярностей оценки выбираются те, которые имеют наибольший вес в конфигурации. При объектно-аспектной модели в итоговое представление выводятся результирующие оценки по объекту и по каждому из его аспектов.

2.4.1.7. Недостатки машинного обучения применительно к задаче автоматического извлечения мнений

Подавляющее большинство работ в области автоматического извлечения мнений опираются на методы машинного обучения [библиография в 26, 108]. При использовании различных методов машинного обучения задача автоматического извлечения мнений рассматривается как частный случай классификации текста. В работах, которые оперируют объектами, при таком подходе считается достаточным простого вхождения объекта анализа в позитивный или негативный текст [174, 50]. Подобное упрощение оправдано для текстов, содержащих только один объект [ср. 26]. Именно по этой причине оценочные факты[167], чаще всего, не учитываются в подобных исследованиях [40].

Критики [например, 50: 328, 26: 31] отмечают, что адаптация к новой предметной области и/или сайту ведет к существенному снижению точности классификатора. Это связано как с оценочной спецификой различных предметных областей, так и со стилистическими особенностями текстов различных жанров. В результате применения классификатора к тексту, который существенно

[167] для решения задачи разграничения оценок от фактов используется этап определения субъективных текстов и предложений.

отличается от текстов в обучающей коллекции, нарушается принцип однородности анализируемых данных. В таких случаях статистические методы неприменимы к задачам анализа текста [168; 139: 74].

Другой недостаток машинного обучения заключается в их «непредсказуемости» [48: 188]. Именно в этом заключается причина отказа ряда коммерческих компаний от статистического определения оценочности. Для исправления ошибки при эвристическом подходе необходимо либо создать новое правило, либо изменить существующее. Для достижения такого же результата при помощи статистических методов необходимо либо переобучить модель при помощи другого алгоритма,[168] либо выбрать новые параметры, либо увеличить объем обучающего корпуса.

Большинство методов обучения с учителем, применяемых для решения задачи оценочной классификации текстов, оперируют комбинацией метода «мешка слов» (*bag-of-words*) и оценочных шифтеров. В качестве обучающего корпуса при этом используются массив текстов, в котором каждый текст имеет общую разметку, а также содержит оценки для каждого слова в корпусе. Иногда также учитываются части речи оценочных слов. На основе этих параметров строится векторное представление текста, которое подается на вход алгоритму классификации, например, опорных векторов. Применение метода опорных векторов объясняется его высокими показателями в плане точности классификации. Декодирование модели проводится (чаще всего) при помощи модели косинусного подобия.

При подготовке обучающего корпуса требуется придерживаться принципов равномерного распределения положительных и отрицательных текстов, в противном случае может произойти т.н. перекос модели. В таком случае система присваивает больший вес тому типу текстов, который лучше представлен в обучающем корпусе.

---

[168] при этом для каждого варианта неясно, будет ли исправлена ошибка, не произойдет ли «переобучения» или не упадет ли качество системы в целом.

Для преодоления проблем эвристического и статистического подхода рядом исследователей был предложен гибридный подход к извлечению мнений. Наибольшее распространение получили методы автоматического вывода правил [например, 43, 38]. По заявлениям авторов автоматический вывод правил из корпуса позволяет учесть те случаи, которые не были обнаружены экспертом. Кроме того, использование машинного обучения способствует формированию компактных правил.

Основной трудностью гибридного подхода является создание обучающего корпуса. В отличие от чисто статистических подходов здесь требования к сбалансированности корпуса несколько ниже.

Подводя итог анализу статистических методов относительно задачи извлечения мнений, можно констатировать, что основными проблемами данных методов являются: 1) чрезмерное упрощение задачи извлечения мнений, 2) невозможность использования обученной модели на принципиально новом материале без переобучения, 3) сложность в исправлении ошибок модели.

2.4.2. Специализированная система автоматического извлечения мнений

Существует довольно много различных типов систем моделирования эмоций, частным случаем которых являются системы автоматического извлечения мнений. Рассмотрим основные направления систем автоматического извлечения мнений [Рисунок 18].

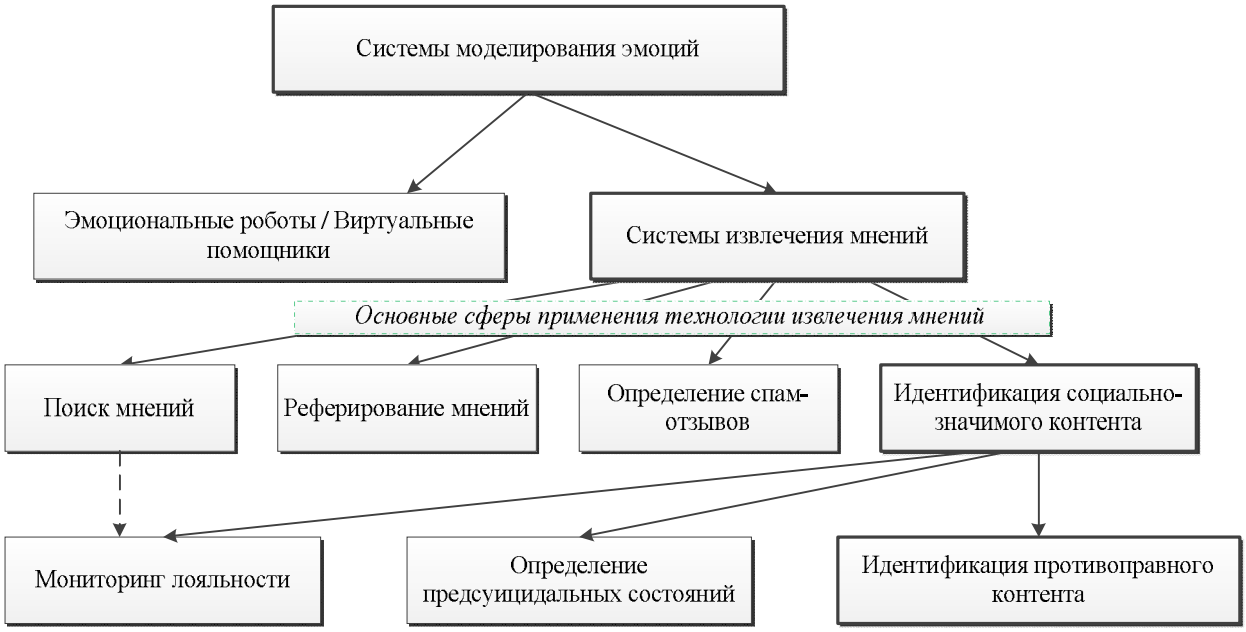

Рисунок 18. Упрощенная классификация систем моделирования эмоций

К числу наиболее значимых сфер применения технологии автоматического извлечения мнений мы относим следующие: поиск мнений, реферирование мнений, определение спам-отзывов и идентификацию социально-значимого контента. Поиск мнений является наиболее общей задачей. Его применение в задаче мониторинга лояльности рассмотрено ниже (см. п. 2.4.3.1). Задачи автоматического реферирования мнений и определения спам-отзывов в значительной степени относятся к пограничным направлениям компьютерной лингвистики — автоматического реферирования текста и классификации контента. Анализ указанных областей автоматического извлечения мнений выходит за рамки нашего исследования. Наиболее молодой областью извлечения мнений[169] мы считаем идентификацию социально-значимого контента. Эта область также сильно дифференцирована. Мы выделяем в ее рамках следующие ключевые направления: мониторинг лояльности (применяется в маркетинге, политических и социологических исследованиях), определение предсуицидальных состояний и идентификация противоправного контента (см. пп. 2.4.3.2 и 2.4.3.3). Определение предсуицидальных состояний при помощи социальных сетей является одним из наиболее перспективных направлений сразу

[169] за исключением анализа пользовательской лояльности, которая относится к первым коммерческим приложениям автоматического извлечения мнений..

в нескольких областях компьютерного анализа данных. К ним следует отнести компьютерную криминалистику, автоматическую классификацию текстов, исследование социальных медиа, социологическую прогностику, психиатрию и автоматическое извлечение мнений. Несмотря на актуальность и колоссальную практическую значимость изучение данного направления не входит в задачи нашего диссертационного исследования. В первую очередь это связано с методологией современных исследований в данной области знаний, которые опираются исключительно на данные машинного обучения [например, 29].

Вначале мы рассмотрим общие принципы построения специализированных систем автоматического извлечения мнений, а затем опишем особенности реализации отдельных классов систем.

2.4.2.1. Особенности специализированных систем автоматического извлечения мнений

В отличие от систем автоматического извлечения мнений общего назначения, специализированные системы направлены на обработку узкой предметной области или решение частной задачи. Самыми известными из отечественных систем подобного типа являются ИАС «Призма»[170] и «Новостной терминал ГЛАСС»[171] компании «Медиалогия». В названных информационных системах используются заранее сформированные списки объектов анализа (не более 20). Но подобное понимание специализированной системы автоматического извлечения мнений не является единственным. Схема взаимодействия словарных компонентов специализированной системы автоматического извлечения мнений представлена ниже [Рисунок 19]:

[170] предназначена для анализа текстов социальных медиа.
[171] система автоматического мониторинга текстов СМИ.

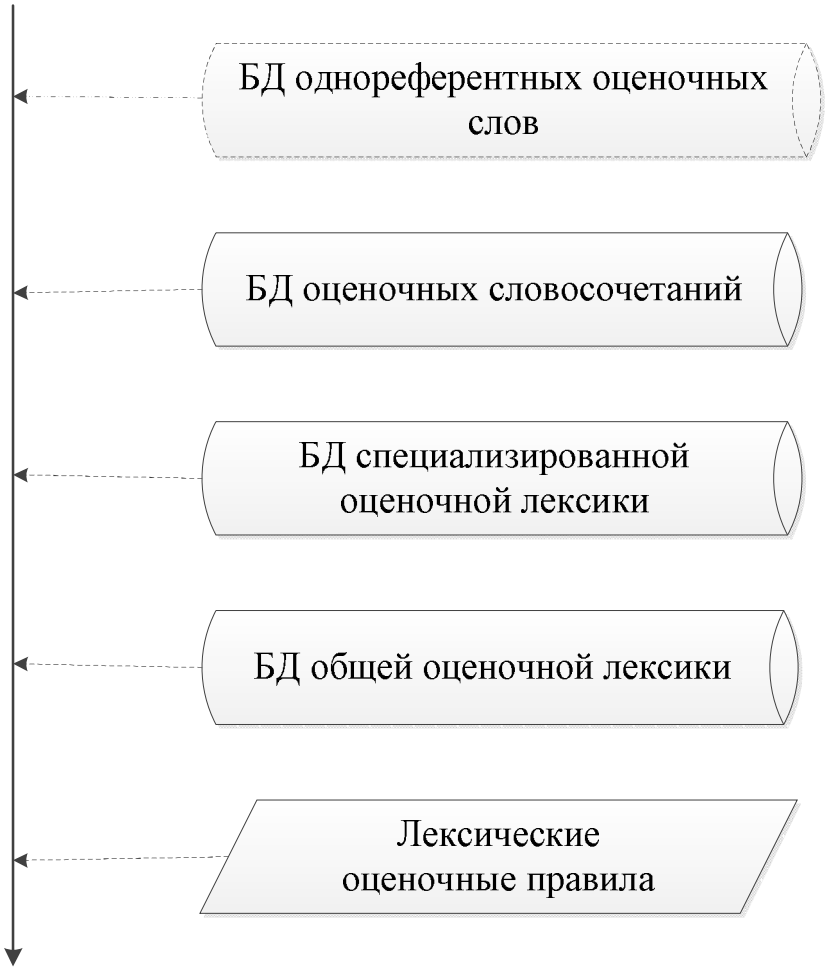

Рисунок 19. Упрощенная схема взаимодействия словарных компонентов специализированной системы автоматического извлечения мнений

Если в системе общего назначения оценка присваивается при помощи общего словаря, а специализированные словари являются, скорее, исключением, чем правилом, то в специализированной системе поиск проводится сначала в специализированном оценочном словаре, и только при отсутствии в нем анализируемого слова — в общем оценочном словаре.

На приведенной нами схеме в качестве первого этапа указано определение имплицитной объектно-ориентированной оценки. В системе общего назначения подобная лексика рассматривается в качестве либо положительной, либо отрицательной, но в отдельный класс не выводится.[172]

За счет введения дополнительных этапов анализа достигается максимально точное моделирование предметной области. Описанная схема является гибкой,

[172] хотя может применяться для идентификации объекта.

т.к. специализированные словари могут подключаться для любой предметной области. При этом также возможно использование нескольких словарей, последовательность обработки которых указывается в конфигурации системы.

2.4.2.2. Использование модели мира в системах автоматического извлечения мнений

Модель мира в специализированных системах применяется для интерпретации оценочных фактов. Кроме того, бывают узкоспециализированные решения. Например, в системах анализа эффективности банков в модель мира включаются математические сравнения (доходность по показателям, изменение клиентской базы и др.).

Наиболее часто используемой моделью мира является *модель продавца*. При этой модели продажи активов рассматриваются как негативные факты. Продажи пассивов, наоборот, — как позитивные. Рост цены на объекты продажи рассматриваются с положительной точки зрения, а на издержки — с отрицательной.

Существенным ограничением подобного подхода мы считаем невозможность комбинации различных моделей мира. Это связано с детальной проработкой одной конкретной модели мира, зачастую, под конкретного заказчика.

2.4.3. Система идентификации противозаконного контента

Одной из сфер применения технологии автоматического извлечения мнений является информационная безопасность[173]. Существует множество сфер использования технологий анализа текста для самых разнообразных задач этой комплексной области. Ключевыми, на наш взгляд, областями следует признать онлайн-мониторинг, автоматическую модерацию комментариев пользователей и программные средства ограничения доступа к веб-контенту.

[173] См., например, анонс конференции Text Analytics for Cybersecurity and Online Safety: http://www.ta-cos.org/program-2016

Перечисленные области информационной безопасности имеют принципиально разных заказчиков. Если онлайн-мониторингом пользуются практически все — маркетологи, социологи, политологи, HR-агентства, государственные структуры, то автоматической модерацией комментариев занимаются только онлайн-СМИ и социальные сети. Последний тип систем является исключительно прерогативой государства.

2.4.3.1. Системы онлайн-мониторинга

В настоящее время онлайн-мониторинг разрабатывается и поддерживается различными отечественными компаниями. Это и системные интеграторы (например, «Ай-Теко»[174] и «Медиалогия»-IBS), и специализированные компании и стартапы (например, «Айкубаз», «Ингейт Синергия»[175], «YouScan»), и государственные структуры (например, «Интерфакс» и «РИА Новости»).

Ключевыми отличиями перечисленных систем друг от друга являются набор источников, аналитических инструментов (в т.ч. визуализация), количество семантических типов сущностей и методология автоматического извлечения мнений. Базовыми задачами систем онлайн-мониторинга являются определение объекта анализа и выявление его взаимосвязей с другими объектами.

Наибольшее значение при онлайн-мониторинге уделяется динамике объекта (упоминания, изменение лояльности, локализация). Из этих аспектов к автоматическому извлечению мнений относится только лояльность (соотношения положительных и отрицательных упоминаний объекта). Измерение индекса лояльности часто сопровождается инструментами оперативного реагирования, например, при помощи кнопки «Связаться с автором сообщения». В некоторых системах мониторинга также стоятся рейтинги лояльности по источникам (что особенно важно при идентификации заказных кампаний, направленных на ухудшение репутации объекта).

Для построения индекса лояльности используется *поиск мнений*. Он заключается в идентификации объекта анализа в заданном оценочном контексте.

[174] системы «Аналитический Курьер» и BrandAnalytics
[175] Babkee.ru

При этом возможны две стратегии. Первая стратегия (т.н. *обработка на лету*) предусматривает заранее сформированный список объектов, которые и ищутся по проиндексированной базе источников. Затем для каждого документа (при использовании документной классификации) или каждого предложения с объектом анализа определяется оценка. Далее в зависимости от типа оценки проводится ее добавление в специализированную базу данных. На заключительном этапе проводится сравнение числа полей каждой из баз данных с оценками по объекту мониторинга. Вторая стратегия подразумевает автоматическое определение возможных объектов мониторинга и извлечение оценок относительно каждого из обнаруженных объектов. На последующих этапах анализа реализация данной стратегии не отличается от предыдущей. Принципиальное различие между стратегиями заключается в требованиях к аппаратному обеспечению системы мониторинга.

2.4.3.2. Системы автоматической модерации комментариев

В последние годы в мире резко возрос уровень вражды и нетерпимости. Особенно легко проследить эту тенденцию в текстах блогов, социальных медиа и новостных комментариях. Психологи объясняют данный феномен чувством безнаказанности, а также анонимностью (или псевдонимностью) в онлайн-общении. Между тем, государственные структуры, занимающиеся надзором в сфере Интернета (например, Роскомнадзор), предписывают владельцам сетевых ресурсов соблюдать законодательство в части проявлений нетерпимости.

Существует три способа модерации комментариев и публичных сообщений в Интернете. Первый способ — ручная модерация, осуществляемая экспертом. Данный подход обладает наибольшей точностью, но требует значительных временны*х* и финансовых затрат. Именно по этой причине ряд изданий, например, российская версия делового журнала Forbes, в определенный момент времени отключил возможность комментировать новости.

Второй способ — автоматизированная модерация комментариев. Автоматизированная модерация комментариев пользователей может

осуществляться в двух режимах. Первый режим можно условно назвать «жадным»[176]. При наличии в тексте комментария признаков потенциальных нарушений законодательства текст скрывается от пользователя и направляется модератору на проверку. Второй режим («ленивый») подразумевает изменение цвета текста комментария (например, Рисунок 20) или его скрытие от пользователя (например, Рисунок 21). При «ленивой» модерации комментарий также может быть удален модератором после проверки. Наиболее распространенным приемом является проверка комментария на наличие ключевых слов. К ключевым словам относятся, например, следующие — «революция», «протест», а также нецензурная лексика.

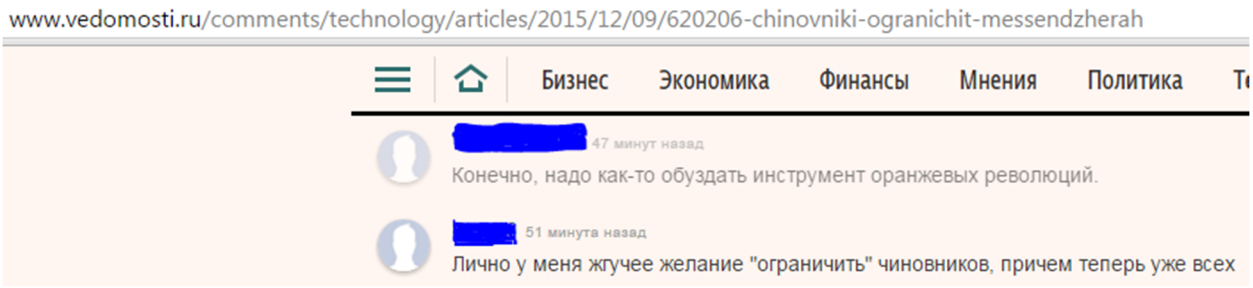

Рисунок 20. Пример «ленивой» автоматизированной модерации комментариев с изменением цвета шрифта

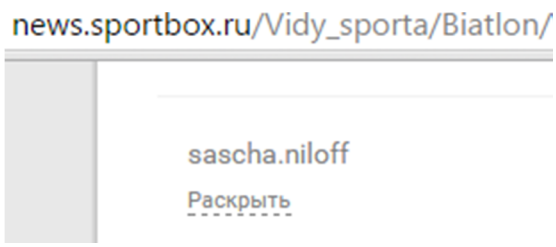

Рисунок 21. Пример «ленивой» автоматизированной модерации комментариев со скрытием текста

Ключевым недостатком простого словарного подхода является отсутствие средств «нейтрализации» оценки. В последнее время в социальных сетях стали блокировать аккаунты пользователей, которые цитируют классиков литературы («*хохол*», «*жид*» у Н.В. Гоголя, «*negro*» у классиков американской литературы XIX века.). К наиболее значимым нейтрализаторам можно отнести: цитаты из

[176] По аналогии с регулярными выражениями.

общепризнанных источников (Библия, Коран) или классиков литературы, а также употребление согласованных лексических маркеров, например *слово*, *термин*, *понятие*.

Третий способ — автоматическая модерация комментариев. При автоматической модерации комментариев текст комментария, подпадающий под запрещенный шаблон, автоматически удаляется из пула комментариев. Данный метод вызывает наибольшее количество нареканий у пользователей и применяется сравнительно редко.

У последних двух способов модерации комментариев есть один ключевой недостаток, сказывающийся на полноте идентификации негативно-оценочных случаев. Это языковая игра и оценочное словообразование. В настоящий момент эти проблемы не имеют однозначного решения ни в компьютерной, ни в теоретической лингвистике.

2.4.3.3. Системы ограничения доступа к веб-контенту

Ограничение доступа к веб-контенту является значимой опцией и для информационно-поисковых систем, и для веб-браузеров. В данных системах фильтрация контента проводится аналогично автоматической модерации комментариев.

В последние годы становится массовой тенденцией государственный контроль над Интернет-пространством. Это может проявляться в ограничении на доступ к определенным ресурсам (например, списку доменных имен, как в Китае, КНДР, Иране). Другой тенденцией становятся системы государственного мониторинга за противоправным контентом в Интернете. Подобным мониторингом в РФ занимается целый ряд ведомств, наиболее значимым из которых является Роскомнадзор.

В Роскомнадзоре совместно с ГлавНИВЦ АП РФ разработана система автоматического определения противоправного контента [Васечкин и др., 2014]. Система (насколько можно судить из открытых источников) представляет собой набор статистических классификаторов по параметрам, которые соответствуют

определениям УК РФ (национальная, религиозная рознь, экстремизм и др.). Из данных частного технического задания неясно, применяются ли в системе фильтры, подобные указанным в предыдущем параграфе.

В силу вышеизложенного рассмотрим возможную структуру системы автоматической идентификации противоправного контента, опираясь на лингвистические принципы. Известно, что в Гражданском Кодексе СССР, а следом за ним и в ГК РФ разделены понятия «факт» и «оценочное суждение» [131: 8]. Исходя из этого необходимо, чтобы в компоненте фильтрации сообщений/комментариев учитывался тип высказывания по этому параметру. Кроме того, в компоненте фильтрации должны учитываться комбинаторные фильтры, приведенные выше. Более того, система должна быть не автоматической, а автоматизированной (АРМ лингвокриминалиста). Мы считаем, что при текущем уровне развития технологий прагматического анализа не следует применять ее для вынесения (пусть и предварительного) решения, которое может повлечь уголовную ответственность. С другой стороны, однозначно идентифицируемые случаи должны вводиться в БД словосочетаний, по которой повторное решение эксперта не требуется (по аналогии с технологией Translation Memory[177]). Ниже (

[177] Впрочем, при этом порог совпадения должен быть равен только 1.

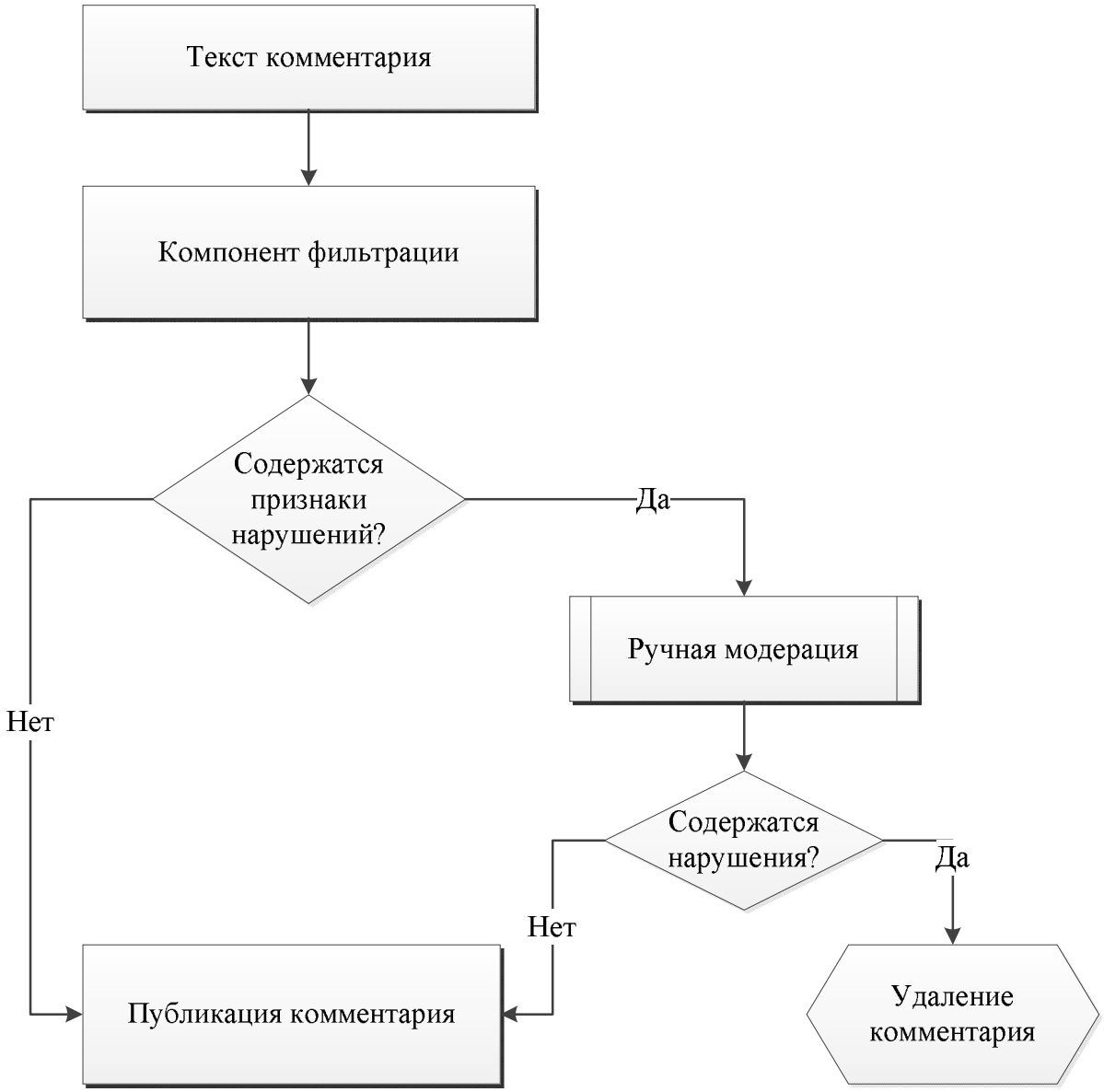

Рисунок **22**) приведена схема подобной системы. Она состоит из следующих компонентов:

1. Компонент фильтрации.
2. БД признаков нарушения.
3. АРМ ручной модерации комментариев, с банков данных законодательных актов в области уголовного права.
4. Компонент разметки комментариев (теггирование осуществляется по значениям «СодержитНарушение» и «НеСодержитНарушений»).
5. Механизм взаимодействия с сервером, на котором физически хранится текст комментария.
6. Программный комплекс администрирования БД.

Более подробная схема, включающая возможность автоматизации процесса, представлена на следующей схеме (Рисунок 23). Отличительной чертой данного варианта является компонент однозначной идентификации нарушений, который представлен ниже (Рисунок 24). Специализированным модулем в компоненте однозначной идентификации нарушений является конфигурационная схема (Рисунок 25). Данная схема позволяет настраивать параметры правонарушений, например, требуется ли анализировать случаи обвинений в тех или иных правонарушениях. Возможный способ идентификации подобных обвинений при помощи базы данных тезаурусного типа описан в Главе 3. На значимость использования специализированных баз данных по проявлениям вербальной агрессии указывает ряд лингвокриминалистов, например [157: 113].

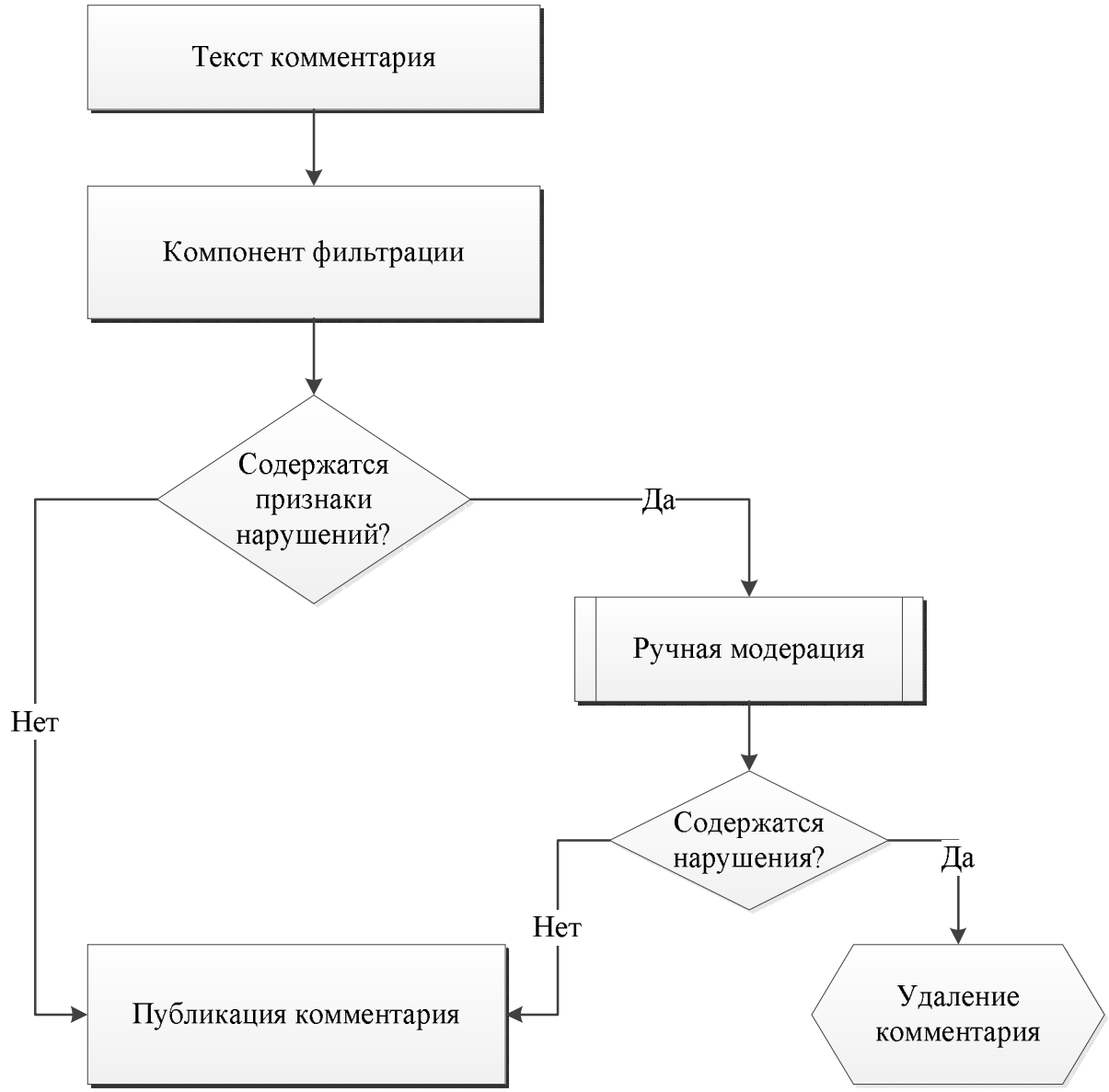

Рисунок 22. Упрощенная структура системы автоматической блокировки контента

Рисунок 23. Упрощенная схема автоматизации АРМ модерирования

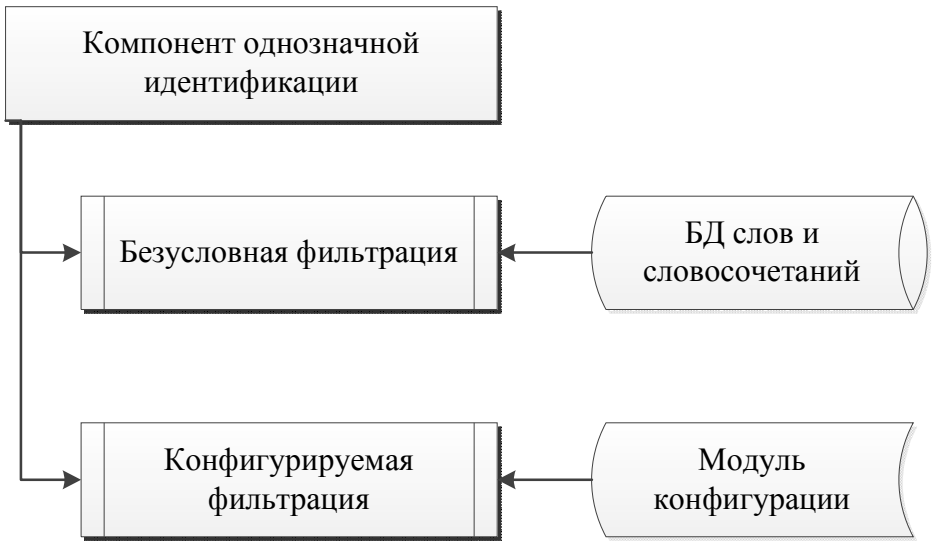

Рисунок 24. Схема компонента однозначной идентификации нарушений

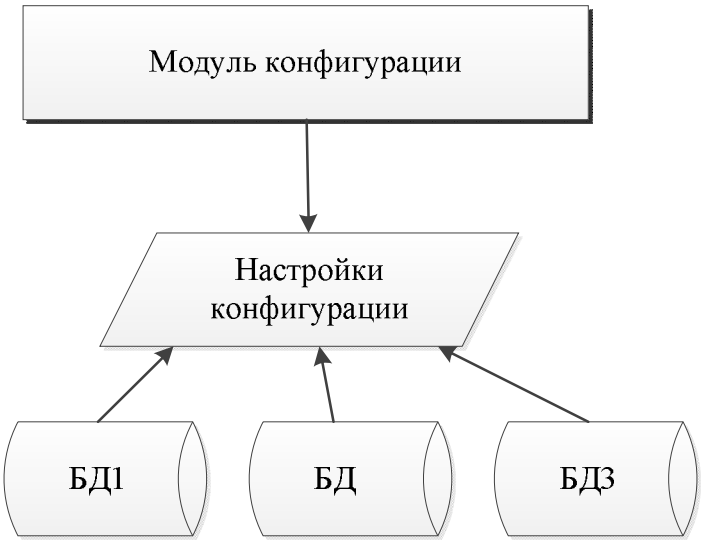

Рисунок 25. Схема модуля конфигурации

Выводы к Главе 2

1. Значительную роль для автоматического извлечения мнений играет предобработка текста. В компонент предобработки должны входить средства анализа экстралингвистической информации, представленной в html-коде веб-документов. Обязательным этапом предварительного анализа текста является классификация текстов с целью сокращения оценочной омонимии для различных предметных областей.

2. Наибольшее значение для автоматического извлечения мнений играет морфологический анализ текста, результаты которого используются при определении объекта и субъекта оценки, а также при составлении словаря оценочной лексики.

3. Этап фильтрации псевдооценочных конструкций в значительной степени опирается результаты семантического анализа текста. На данном этапе обработки проводится сокращение как количества потенциальных объектов анализа (что способствует высокой скорости обработки текста), так и априорно неоценочных конструкций (что повышает точность анализа).

4. Объектная модель анализа является наиболее лингвистически обоснованной. В случаях, когда объект анализа имеет сложную структуру, необходимо использовать объект-аспектную модель. Для получения наиболее полной информации об объекте и его свойствах целесообразно анализировать объекты не изолированно, а учитывая их кореферентные отношения.

5. Ключевым понятием в автоматическом извлечении мнений является модель мнения. Данная модель должна наиболее полно моделировать коммуникативную ситуацию, которая подвергается анализу. По этой причине необходимо разделять отношения между объектом мнения и субъектом (носителем) мнения. При этом субъекты должны различаться от канала передачи мнения. Кроме того, оказывается полезным разделять непосредственно оценочные контексты от причинно-следственных, в т.ч. лексикализованных.

6. Методы машинного обучения, господствующие в современной практике создания систем автоматического извлечения мнений, не позволяют оперативно исправлять ошибки классификации, а также трудно адаптируются к новым предметным областям. Объектные модели извлечения мнений, использующие машинное обучение, чересчур упрощают задачу идентификации объекта и не способны правильно определить оценку для разных объектов в рамках одного оценочного факта.

7. Ключевым компонентом в системе извлечения мнений является словарь оценочной лексики. Данный компонент представляет собой сложный программный продукт, состоящий из нескольких баз данных, а также набора лексикализованных правил. Для различных вариантов систем автоматического извлечения мнений необходима различная организация словарного компонента.

8. Системы автоматической идентификации противоправного контента обладают наиболее сложной организацией как словарного компонента, так и правил отнесения контента к противоправному. Являясь частным случаем экспертных систем, системы указанного типа выступают, как правило, в виде автоматизированного рабочего места эксперта-(лингво)криминалиста.

Глава 3. Разработка принципов автоматизированного составления ресурсов для автоматического извлечения мнений

В данной главе рассмотрены принципы создания специализированных ресурсов, которые могут быть использованы при разработке систем автоматического извлечения мнений. В разделе 3.1 представлена методика автоматизации создания первичного словаря оценочных прилагательных. Обобщенное описание формата для корпуса текстов, предназначенного для тестирования систем автоматического извлечения мнений на уровне объектов, дается в разделе 3.2. В разделе 3.3 описана структура и принципы использования лексической базы данных по этнофобонимам. В разделе 3.4 приведен способ автоматического заполнения полей лексической базы данных по этнофобонимам с использованием технологии морфемного синтеза.

3.1. Методика автоматизации создания первичного словаря оценочных прилагательных[178]

Как уже сообщалось выше (Глава 2) наиболее значимым компонентом любой системы автоматического извлечения мнений является словарь оценочной лексики. Целью данного раздела является описание и апробация двухступенчатой методики, использующей неспециализированный корпус текстов для автоматизированного составления списков оценочной лексики. Составленные списки могут быть использованы для идентификации предложений, несущих оценки относительно широкого класса объектов искусства (кино, театр, книги).

Существующая практика создания словарей оценочной лексики подразумевает активное использование различных автоматизированных средств. Методы ручной правки словарей применяются довольно редко (например, [78]), в основном, при редактировании полученных словарей. В настоящее время существует большое количество лингвистических ресурсов, предназначенных для нужд разработчиков систем автоматического извлечения мнений. Данные ресурсы можно разделить на два типа — размеченные оценочными тегами корпусы текстов и словари оценочной лексики. У каждого из названных источников информации есть свои особенности и недостатки.

Рассмотрим обобщенную методику использования готовых словарей. Существует два основных типа оценочных словарей (в виде списка или в виде базы данных). На практике в подавляющем большинстве случаев при генерации нового словаря оценочной лексики используют комбинированный подход, при котором список оценочных слов представляется в виде запросов к базе данных для извлечения связанных слов, учитывая таксономические связи между словами. Самым часто используемым ресурсом является тезаурус WordNet и различные его варианты для разных языков (например, EuroWordNet, MultiWordNet), а также специализированные версии, ориентированные на автоматическое извлечение мнений (SentiWordNet[179], SentiWords[180] и WordNet-Affect[181]). Самым

[178] В основе данного раздела лежит статья автора [110].
[179] http://sentiwordnet.isti.cnr.it/

значительным преимуществом использования тезаурусов типа WordNet считается возможность автоматически извлекать антонимы и синонимы для конкретных значений слов. Одной из основных проблем при использовании WordNet следует признать отсутствие единого стандарта при разграничении лексических значений для различных языков и различный объем самих словарей [31]. Другой не менее значимой проблемой является необходимость применения технологий автоматического определения лексических значений, которые в настоящее время для русского языка не дают приемлемого качества. Иногда методы, использующие в качестве источников оценочных словарей другие словарные ресурсы, подвергаются критике в плане отражения устаревшей лексики, не соответствующей лексике реальных текстов.

Применение корпусных технологий в рамках автоматического извлечения мнений способствует привлечению более актуальных языковых данных и позволяет значительно более оперативно увеличивать объем словаря. Самым известным подходом считается расширение уже готового словаря за счет логических связей между известными словами и словами, отсутствующими в исходном словаре. Ключевым фактором при данном подходе является правильный отбор материала при формировании корпуса текстов. Чаще всего используют данные с одного или нескольких сайтов, имеющих сходные названия, идентичные разделы и т.п. Таким образом, используются узкоспециализированные корпусы. Ключевым недостатком подобных подходов следует признать сложность адаптации к новой предметной области [26], где определенные оценочные слова могут нести противоположную оценку или быть нейтральными. Отдельную проблему составляет соотносимость оценок на различных сайтах (например, преобразование трехбалльной шкалы в пяти- или десятибалльную). Кроме того, сложно оценить репрезентативность подобных корпусов текстов.

---

[180] https://hlt.fbk.eu/technologies/sentiwords
[181] http://wndomains.fbk.eu/wnaffect.html

В настоящее время существует довольно много словарей оценочной лексики, предназначенных для различных нужд автоматического извлечения мнений. Рассмотрим некоторые из таких словарей, составленных для русского и нидерландского языков. Выбор языков обусловлен незначительным количеством лингвистических ресурсов свободного доступа (open-source) для обоих языков. Словарь И.И. Четверкина и Н.В. Лукашевич [175] представляет собой первичный словарь для определения фрагментов текста, несущих оценочную информацию. Словарь был составлен при помощи следующей методики: вначале был составлен корпус текстов отзывов о фильмах; затем были выявлены специальные оценочные слова на основе частотного распределения между оценочными отзывами о фильмах, описаниями фильмов и общеязыковым корпусом (тексты СМИ). В результате был получен частотный словарь (объемом 18362 слова), из которого посредством ручной разметки были извлечены 4079 оценочных слов. Затем различными статистическими процедурами была проверена значимость данных слов для предметной области. Применение данной методики для других предметных областей показало несущественное расхождение в лексическом составе. После объединения всех списков в один общий объем словаря составляет 5000 лемм. Анализ содержания словаря позволяет выявить некоторые особенности словника: 1) объекты оценки не отделены от оценочных слов (например, меню, фотография, телевизор), 2) в словаре присутствуют неоценочные или узкоспециализированные оценочные слова (например, советский, голливудский, режиссерский, русскоязычный), 3) некоторые слова получили неверную лемматизацию (например, блютусый (правильная лемма «блютус» либо единичное прилагательное), фентезь (правильная лемма «фентези»), фотикий (правильная лемма «фотик»)), 4) исходные текстовые массивы для разных предметных областей были разных объемов, что ведет к несбалансированности общего корпуса и нарушению частотного принципа формирования словника (доказательством может служить разная значимость для слов, обозначающих объекты). При составлении словаря использовались традиционные для информационного поиска методы классификации текстов.

Выбор в качестве источников отзывов только сайтов Имхонет и Яндекс.Маркет может повлиять на применимость словаря для отзывов с других сайтов. Словарь Дэ Шмедта и Далемана [12] содержит только оценочные прилагательные. Словарь был составлен на основе коллекции из 14000 отзывов о книгах. Для отбора прилагательных были отобраны 1100 прилагательных с частотой свыше 4. Затем список был отфильтрован вручную. Получившийся объем словаря составил 1044 лемм. В дальнейшем словник был расширен автоматическими методами с последующей ручной проверкой до 5407. В качестве ключевых методов использовались два: критерий сочетаемости с одинаковыми существительными и автоматическое расширение при помощи синонимического тезауруса. При составлении словаря акцент делался на фламандском варианте нидерландского языка, что может привести к неправильной интерпретации оценочных значений некоторых прилагательных в процессе обработки голландского варианта нидерландского языка.

Перейдем к рассмотрению двухступенчатой методики формирования первичного оценочного словаря. Под первичным оценочным словарем мы понимаем список оценочных слов, применяемых на этапе выделения оценочных предложений из текста. На первом этапе на основе сверхбольшого морфологически размеченного корпуса текстов (свыше 1 млрд. словоупотреблений) создается квазисинонимический тезаурус понятий предметной области. Его формирование происходит следующим образом: задается лемматизированная форма слова, интуитивно рассматриваемого в качестве базового для конкретной предметной области. Такими словами будут, например, «фильм» для отзывов о киноиндустрии, «отель» для гостиничного сектора и т.п. Затем производится поиск похожих слов (по контекстному окружению и ряду других параметров). В результате получается список слов, которые достаточно полно покрывают предметную область. В дальнейшем можно провести подобный поиск и для наиболее статистически близких квазисинонимов (коэффициент близости от 0,8). Этот шаг увеличивает полноту покрытия. Перед переходом ко второму этапу из списка слов удаляются низкочастотные синонимы

(коэффициент близости до 0,05). Данный шаг необходим для снижения эффекта поискового шума. На втором этапе для каждого слова из полученного тезауруса отбираются имена прилагательные (в данной работе мы специально не разграничиваем причастия и прилагательные), встречающиеся на расстоянии 1 слева от слова в запросе. Выбор имен прилагательных связан с тем, что именно они наиболее сильно модифицируют оценку слов в текстах Сети Интернет, где значительный процент предложений не имеет глаголов или предикативных слов (см. также [12]). Выбор расстояния обусловлен значительным объемом анализируемых данных, что понижает вероятность того, что связанное оценочное прилагательное не будет встречаться непосредственно в препозиции к оцениваемому слову. В результате второго этапа формируется частотный список всех прилагательных. Из данного списка удаляются низкочастотные прилагательные (с встречаемостью ниже 10 словоупотреблений в корпусе). Эта процедура способствует избавлению от ошибок парсинга hmtl, ошибок морфологического анализа низкочастотных слов и сокращает время ручной постобработки полученного списка. Последним этапом при создании первичного оценочного словаря является ручная фильтрация полученного списка оценочных слов. На необходимость подобной фильтрации указывает, например, Bing Liu [26]. С целью некоторой автоматизации ручной работы целесообразно использовать уже готовые списки заведомо неоценочных слов (например, список прилагательных, обозначающих цвета) и использовать словообразовательные шаблоны. Использование словообразовательных шаблонов обусловлено тем фактом, что целый ряд прилагательных образуется от имен и фамилий, а также географических названий (наиболее распространенной моделью для подобных слов, видимо, является окончание корня на согласную букву с последующим суффиксом –ский, например, *парижский*, *советский*, *сельский*, *бийский*). Предлагаемая методика является близкой к методике Дэ Шмедта и Далемана [12] с той разницей, что 1) не используется расширение словаря при помощи готовых словарей и 2) в качестве первичного списка берется список существительных, а не прилагательных. Одной из причин подобного выбора является направленность

нашей методики на выявление предложений, несущих оценку, а не на выявление самой оценки. Другой сходной работой является [14]. Основным отличием является использование при формировании слов, описывающих предметную область, не относительно статичных тезаурусов (например, Wikipedia), а корпуса текстов.

В целях апробации описанной методики мы использовали технологию Sketch Engine [20]. В качестве материала был использован корпус ruTenTen [16]. Объем данного корпуса веб-текстов на русском языке составляет 15,8 млрд. словоупотреблений. В корпусе представлены тексты 2011 года. Для морфологической разметки текстов применялся анализатор TreeTagger. В качестве основной причины выбора неспециализированного Интернет-корпуса является принцип независимости словаря от специфических особенностей текстовой выборки. В качестве моделируемой области была выбрана область отзывов о фильмах (как частный случай отзывов на объекты искусства). На первом этапе был составлен квазисинонимический тезаурус. Была использована опция автоматического построения тезауруса по поисковому слову, в нашем случае «фильм». Получившийся в результате список содержит следующие группы (в скобках указаны связанные слова): сериал (кинофильм, мультфильм); книга (картина, рассказ, роман, произведение, музыка, статья, сюжет, песня, текст, сказка, литература, творчество, пьеса); кино (комедия, драма); игра (история, проект, программа, реклама, сайт, событие, материал, образ, тема, идея, слово, вещь, жизнь, стиль, работа, модель, группа, команда); спектакль (выступление, концерт, сцена, театр); ролик (клип); фотография (новость, картинка, запись); серия (версия). Очевидно, что для дальнейшего исследования необходима фильтрация некоторых многозначных слов, использование которых будет приводить к большому шуму. Затем были построены частотные сочетания прилагательных и существительных из тезаурусной группы. В рамках данного эксперимента мы ограничились тремя словами (фильм, кинофильм, кино). После удаления низкочастотных слов получилось три списка прилагательных (2307, 596 и 1087). Число общих слов для всех списков составило 170 слов. После

фильтрации неоценочных слов мы получили общий список в 1514 оценочных слов, который был расширен при помощи регулярных выражений до 2966[182]. В ходе анализа неоценочных слов были выявлены следующие ошибки: ошибки парсинга html (отделяется окончание слова стоящее после дефиса или тире, например, -ское, остающиеся непечатные символы, например, nbspфантастичекое) и ошибки нормализации значений несловарных прилагательных (например, голливудским, голливудских, голливудского, голливудском). Таким образом, встает необходимость увеличения точности работы морфологического анализа слов, отсутствующих в словаре, для русскоязычных текстов.

В результате анализа ошибок методики было выявлено, что фильтрацию низкочастотных слов необходимо приводить после повторной лемматизации с применением морфоанализатора, отличающегося от используемого в системе TreeTagger[183]. После повторной лемматизации целесообразно произвести пересчет частот. Относительно небольшой объем словаря в дальнейшем предлагается увеличить за счет снижения частотного порога до четырех вхождений в коллокацию со словом из квазисинонимического тезауруса. Другим способом пополнения словаря может стать привлечение прилагательных, сочетающихся с другими словами из тезауруса[184].

3.2. Формат размеченного корпуса текстов на уровне объектов

В настоящее время происходит значительный рост числа систем автоматической обработки текста. Ручное сопоставление результатов их работы в силу больших объемов данных, необходимых для объективного тестирования, не представляется возможным. Поэтому появляется множество проектов (MUC, TREC, NIST, LREC, Romip), целью которых является выработка объективных критериев автоматической оценки таких систем.

---

[182] См. Приложение 2, где описана процедура расширения словаря.
[183] http://www.cis.uni-muenchen.de/~schmid/tools/TreeTagger/
[184] в рамках единого синсета

За почти 20 лет существования соревнований по автоматическому оцениванию систем обработки текста выработаны основные универсальные критерии такой оценки. К ним относятся: 1) наличие единого формата представления данных, 2) наличие эталонного корпуса, размеченного вручную экспертами, 3) наличие метрики, согласно которой проводится оценивание. В данном параграфе мы остановимся на формате представления данных при автоматической оценке систем автоматического извлечения мнений на уровне объектов. В качестве метрики оценивания нам представляется целесообразным использовать традиционные для информационного поиска меры полноты и точности, а также F-меру (меру Ван Ризбергена).[185]

Автоматизация оценки работы систем автоматического извлечения мнений представляет собой отдельную задачу в силу специфики формулировки проблемы. Хотя в общем виде автоматическое извлечение мнений (см. Главу 2) занимается определением отношения автора текста к проблемам, описываемым в тексте, на практике оно сводится к одной из трех задач — оценочной классификации текста статистическими методами, оценочной классификации предложений и оценочной классификации объектов в тексте. Существующие методики оценки (например, Romip-2012) учитывают только первый аспект, что приводит к искаженному представлению сведений о качестве систем, основанных на других принципах. При этом ни одна из существующих схем разметки не поддерживает полный структурный шаблон мнения.

Отметим, что наиболее распространенные форматы представления метаописаний (XML, XML-RDF) позволяют приписывать неограниченное количество тегов метаданных. Поэтому именно названные форматы целесообразно использовать для описания тестовых корпусов в области автоматического извлечения мнений. Добавление к элементу XML с общей оценкой текста атрибута с объектом оценки позволит (например, в рецензиях на фильмы) учитывать не только фильм в целом, но и игру актеров, сценарий и др. Добавление атрибута со свойствами объекта {фильм; актер, режиссер, сценарий,

[185] поэтому данный аспект мы подробно не рассматриваем.

эффекты, …} позволит оценивать полноту выдачи систем, учитывающих свойства/характеристики объектов. Добавление атрибута с информацией об авторе высказывания позволит выявить глубину текстового анализа — анафорические отношения, анализ прямой и косвенной речи, анализ метаданных и др. Общее количество узлов в элементе должно соответствовать модели мнения. Исходя из этого, в качестве факультативных атрибутов можно задать основание оценки, ее силу и т.п..

Использование всех описанных выше атрибутов в едином размеченном документе позволит применять общий формат для оценивания реализации любой из известных задач в рамках автоматического извлечения мнений. Таким образом, использование подобного формата представления документов является, по нашему мнению, универсальным для задачи многоаспектной оценки указанных систем. Пример разметки текста приведен в Приложении 3.

3.2.1. Сравнение предлагаемого подхода с существующими решениями

В последние годы появилось несколько проектов, направленных на многоуровневое представление оценки текста. Для анализа мы выбрали следующие (фрагменты размеченных данных приведены в Приложении 3): подход А.Ю. Суторминой [163], подход SentiRuEval-2015 и подход Б. Лю [28]. Выбор именно этих работ обусловлен нижеизложенными соображениями. А.Ю. Сутормина придерживается лингвистического подхода и учитывает оценку относительно объекта. Стандарт разметки SentiRuEval-2015 является единственным стандартом соревнований[186] по сравнению качества существующих систем автоматического определения оценки для русского языка. Формат разметки, предлагаемый Б. Лю, опирается на принципы объект-аспектной модели мнения.

Подход А.Ю. Суторминой был предложен в магистерской диссертации 2014 года.[187] Автор придерживается формата *stand-off annotation*. Как упоминалось в Главе 1, данный формат ориентирован, прежде всего, на машинную обработку и

[186] На момент написания параграфа.
[187] Корпус доступен по адресу: https://github.com/AlyssaSut/assessment_words

сложен для восприятия человеком. А.Ю. Сутормина указывает, что «В обучающем корпусе нами были выделены объекты оценки, выражающие оценку слова (по частям речи), а также слова, на наш взгляд переносящие оценку, т.е. служащие выражению имплицитной оценки.» [163: 21]. Такой формат корпуса нацелен на определение объекта и отдельных оценочных слов и не может применяться в общих задачах тестирования качества системы автоматического извлечения мнений по всем критериям оценки.

Подход SentiRuEval-2015 нацелен на анализ пользовательских отзывов. В качестве модели анализа рассматривается объект-аспектная. Объект анализа задается на основе метаинформации, а аспектные категории заданы конечным списком. При отсутствии в размеченном тексте какой-либо из категорий данной категории приписывается пустое значение. Интересным представляется решение создателей корпуса выбирать в качестве единиц описания свойств объекта словоформы без лемматизации, а также предложные группы. По нашему мнению, отбор глаголов (*гниют*, *цветет*) и глагольных групп (*дорогу держит*) в качестве оцениваемых свойств объекта не оправдан с лингвистической точки зрения. При выборе итогового значения авторы не учитывают структурные маркеры (*общее впечатление*). В разметке отсутствует автор текста, что обусловлено требованиями анонимизации пользователей. Так же в разметке нет указания на время (в данном случае, формально не выражено, но ассоциируется с моментом написания отзыва). В разметке задается тип оценки (эксплицитная, имплицитная, факт [45: 2-3]). При этом неявная (имплицитная) оценка понимается только в контексте выражения свойства при помощи оценочного слова. Подобное определение сказывается на полноте представления в разметке сравнительных характеристик (в рассматриваемом тексте — расход топлива, который можно отнести к категории *цены*).

Подход Лю (2015) нацелен на тестирование качества определения оценки при автоматическом выводе правил [28].[188] Из трех доступных корпусов текстов данной коллекции мы выбрали тексты о микрофонах (D8). В данном подкорпусе

[188] http://www.cs.uic.edu/~liub/FBS/CustomerReviews-3-domains.rar

не проведено различие между непосредственным объектом анализа и его свойствами (аспектом), что противоречит модели мнения, предложенной одним из авторов данной работы [26]. Информация об авторе мнения и критериях оценки не приводится.

Таким образом, можно констатировать, что существующие ресурсы для оценивания качества работы систем автоматического извлечения мнений моделируют лишь отдельные этапы задачи извлечения мнений, а именно идентификацию объекта (или списка его свойств) и его оценку. Основание оценки определяется лишь с точки зрения способов их формального выражения.

3.3. Структура и принципы использования лексической базы данных по этнофобонимам[189]

Как отмечалось в Главе 2, одним из аспектов автоматического извлечения мнений является идентификация участков текста, содержащих различного рода ненормативную лексику, в т.ч. контент противоправного характера. В связи с ростом ксенофобских настроений в Интернете многие Интернет-СМИ и сайты вводят предварительную модерацию комментариев. В настоящее время в Российской Федерации ведется разработка системы, позволяющей в автоматическом режиме осуществлять идентификацию и классификацию противоправного контента [79: 5-6]. В разрабатываемой системе учитываются различного рода призывы и иные прямые проявления ненависти. В данном разделе мы рассматриваем случай обвинения лиц в подобных проявлениях при помощи лексических средств и описываем возможный способ автоматического обнаружения подобных обвинений на примере этнофобии.

3.3.1. Однореферентные оценочные слова

Сочетаемостные свойства различных классов оценочных слов неоднократно становились объектом изучения в работах по автоматическому извлечению мнений. При этом предметом изучения являлись валентностные свойства глаголов и типы оценочных прилагательных (усилители, уменьшители и

[189] В основе раздела лежит статья автора [119].

нейтрализаторы оценки) (см. Главу 2). Референтные свойства оценочных прилагательных получили подробное освещение в работах многих исследователей, обзор которых приводит Б.Лю [26]. Основным наблюдением стало следующее: в различных подъязыках и для различных референтов одни и те же оценочные слова могут нести различную оценку. Например, предложение *A* ***funny*** *movie* – несет положительную оценку для объекта '*movie*' ('*фильм*'). Но тоже прилагательное может нести отрицательную оценку в контекстах поведения прибора или устройства, например, *a site dedicated to all those embarrassing* ***funny*** *auto-correct texts from iphones*. Разница в восприятии оценочных слов может также зависеть от участников ситуации, например, предложение *В IV квартале текущего года установилась* ***высокая*** *цена на бензин* будет негативно воспринято автовладельцами и позитивно владельцами автозаправок.

Следует заметить, что некоторые оценочные слова отражают лишь конкретные объекты или классы объектов. Сначала рассмотрим некоторые случаи словообразования оценочных слов, характеризующих отдельные классы объектов. Наиболее частым случаем является использование приставки *недо-*. Прототипическим словом для данного дериватологического класса является нацистский термин «*недочеловек*»[190]. В качестве современных примеров можно привести слова *недобанк* и *недомагазин*. В качестве примера употребления можно привести такое предложение: *Вывод: БРС – это недобанк, это кредитная лавка, можно пойти работать, если у вас нет опыта работы (не на длительный срок) и есть желание узнать всё о кредитах картах и т.д.*[191] В качестве корня, который подвергается модификации, может использоваться любое родовое понятие, например, нижнего уровня иерархии — *-банк*, *-магазин*, *-продавец*, *-министр*, etc., верхнего уровня — *-организация*, *-учреждение*, *-сотрудник*, etc. Другим способом словообразования подобных слов является словосложение. При этом один из компонентов сложного слова обязательно должен быть выражен

---

[190] Ср. значение приставки *недо-* в [159: 802]: «сложная приставка, придающая слову значение неполноты, недостаточности действия, отсутствия нужной меры, нормы, степени и т.п.». Таким образом, можно говорить о вторичном значении приставки, со смещением значения из снижения степени слова (соответствие лексической функции AntiMagn) в придание негативного значения (ЛФ AntiBon).

[191] https://www.banki.ru/services/official/bank/response/523933/

оценочным словом.[192] К таким словам можно отнести политические неологизмы «*пжив*» ('*партия жуликов и воров*' — референт *политическая партия «Единая Россия»*[193]) и «*воршебник*» (*'вор+волшебник*' — изначальным референтом был *глава ЦИК РФ Чуров*, затем в качестве референта стал подразумеваться любой сотрудник избирательной комиссии). В качестве примеров использования можно привести следующие предложения. *В своем интервью на Эхе воршебник Чуров прямо обвинил телевизионщиков в искажении фактов.*[194] *Потому что до ручки нас так или иначе доведут эти воршебники изумрудного города.*[195] *Но вопрос остается открытым: действительно ли ЕдРо больше не воспринимается как ПЖиВ?*[196]

Оценочные слова, относящиеся к отдельным объектам, образуются по сходным деривационным правилам. При этом референт может как содержаться в слове, так и задаваться имплицитно. К первому типу относятся различные способы «персонофобии», которые рассматриваются ниже. Второй тип включает слова, характеризующие действия или свойства конкретного лица, типа «*нанопремьер*» и «*либералиссимус*». Например, *Нанопремьер Медведев высказался в духе нашистки Светы из Иваново: ”зарплата растет, социальные условия улучшаются, детей в школах стали кормить лучше”.*[197] *В девяностых годах на вопрос о национальности либералиссимус ответил журналистам, что мама у него русская а папа — юрист.*[198]

3.3.2. Этнофобия, социофобия, персонофобия и религиофобия

В последнее время в прикладном языкознании наблюдается тенденция, направленная на активное привлечение технологий, основанных на знаниях. К числу подобных технологий относятся, прежде всего, размеченные корпусы текстов и специализированные лингвистические ресурсы (например, тезаурусы,

---

192 В данную группу входят слова, обозначающие различные «фобии» и «филии», которые рассмотрены ниже.

193 Включая членов данной партии

194 http://nearch78.livejournal.com/8178.html

195 http://www.komarovskiy.net/forum/viewtopic.php?t=22630&postdays=0&postorder=asc&start=675&sid=409b049e00561699b920fde7066d5139

196 http://svpressa.ru/politic/article/63081/

197 http://sturmnovosti.co/view.php?COLLCC=1448022774&id=38688

198 http://tomsk.fm/watch/47431?from=157652

лексические базы данных типа WordNet, VerbNet, FrameNet), онтологии). В системе Роскомнадзора [79: 12] предполагается использование редактируемых пользователем словарей[199], что позволяет подключать дополнительные внешние ресурсы.

В системах, ориентированных на фильтрацию противоправного контента, принята тройная классификация лексики (применительно к словам, обозначающим ксенофобию). Эта классификация опирается на типы правонарушений из Уголовного Кодекса РФ, а именно — возбуждение социальной, этнической и религиозной ненависти (розни). Таким образом, из различных типов ксенофобии выбираются только термины, относящиеся к социофобии, этнофобии и религиофобии. Необходимо отметить, что данные явления на текстовом уровне соотносятся с оскорблениями (чаще всего, путем сравнения с неодушевленными предметами). Другим значимым явлением в Интернет-коммуникации являются обвинения в разжигании того или иного вида розни. Данные обвинения строятся на основе приписывания тому или иному человеку свойств этно-, социо- или религиофобов. Наиболее частыми моделями являются следующие: 1) X есть prefix-фоб-(ка), где X — имя человека, prefix — указание на конкретное проявление ксенофобии, в скобках указан факультативный компонент, относящий слово к лицу женского пола; 2) prefix-фобск* (действие) X. К первой модели относятся, например, следующие предложения: *Однако, отметил он, "решать, что вредоносно, а что полезно, должны не исламофобы, не владеющие даже арабской графикой, а признанные во всем мире специалисты-исламоведы".*[200] *Рогозин - ксенофоб, юдофоб, американофоб? Ох, сомневаюсь я…*[201] Приведенные ниже примеры иллюстрируют использование второй модели. *После скандала с высказываниями российских политиков о казахстанской государственности и о нагнетании в этой стране русофобских настроений в Казахстане заявили о возможности*

199 без уточнения данного термина

200 http://www.newsru.com/religy/22sep2014/gainutdin_spisok.html

201 http://democracy.ru/article.php?id=1081

"негативных вызовов", связанных с безопасностью.[202] *Равно как и о том, как абсурдно было вменять девушкам надуманный христианофобский мотив вместо очевидного политического*.[203]

Анализ текстов при помощи регулярного выражения \w+(?>фоб)\w* демонстрирует наличие форм, которые можно отнести к «персонофобии». Например, *а ты бушафил или обамафоб? Я - путинофоб и медведефоб*[204]. Отличия персонофобии от других проявлений ксенофобии очень существенные. Во-первых, персонофобия не является преступлением. Во-вторых, она вписывается в традиционную картину мира, занимая примерно ту же нишу, что слова *враг* и *недруг*. При этом необходимо отметить, что явные проявления неприязни к конкретному лицу (особенно при помощи расистских высказываний) могут привести к восприятию явления персонофобии как чего-то недопустимого, например, *Что касается моих предпочтений, то могу сказать, что я не за Обаму, я против "обамофобии", как очень точно выразился уважаемый Буквоед*[205].

Компонентный анализ ксенофобонимов[206] позволяет довольно легко выявить деривационные модели. Помимо указанных выше моделей с компонентом *–фоб*, допустимы компоненты *–ненавистник/ца*. Различия между этими словообразовательными моделями заключаются в силе негативной оценки.

Разница в словообразовании между основными классами ксенофобонимов заключаются в референтном статусе первого деривационного компонента. Для социофобонимов — это наименования социальных групп, например, сексуальных меньшинств, представителей власти, футбольных фанатов и др. Для религиофобонимов — это наименования конфессий. Для этнофобонимов — это названия этнических групп. Персонофобонимы образуются от фамилий или прозвищ конкретных людей, чаще всего, широко известных.

---

202 http://www.newsru.com/world/18sep2014/kazakhstan.html
203 http://www.atheism.ru/library/efimov_1.phtml
204 http://otvet.mail.ru/question/18545290
205 http://berkovich-zametki.com/Guestbook/guestbook_okt2012_2.html
206 Т.е. слов, обозначающих различные проявления ксенофобии.

Отличительной чертой этнофобонимов следует признать возможность замены названия этнической группы на название страны, ассоциируемой с данной этнической группой. Например, слово *англофобия,* образованное от топонима *Англия*, имеет тот же референт[207], что и *англичанофобия*. Частотное соотношение между названием этнической группы и названием страны зависит от двух параметров: 1) длины слова (чем длиннее слово, тем менее частотным будет образованный от него этнофобоним), 2) удобство произнесения (чем сложнее фонетический комплекс на стыке частей этнофобонима, тем реже он будет использоваться).

3.3.3. Тезаурусное описание этнофобонимов

В практике информационного поиска давно успешно применяются разнообразные информационно-поисковые тезаурусы. Их основным назначением является увеличение полноты выдачи за счет увеличения длины запроса синонимами (и некоторыми другими семантическими отношениями) слов в запросе.

В отечественной науке уже предпринимались попытки тезаурусного описания фобий [172: 30]. В отличие от данного исследования в процитированной статье под фобиями понимаются психические заболевания. Ещё одним отличием от нашего исследования является многоязычность корпусного словаря (тезауруса) В.И. Хайруллина. Принципиальным представляется и выбор английского языка в качестве исходного. Подобный подход облегчает составление тезауруса, однако не способствует учету межчастеречного варьирования во флективном языке при функционировании лексических единиц в тексте.

Разработанная нами база данных по этнофобонимам в настоящее время насчитывает 85 лексических входов[208]. Под лексическим входом понимается единица языка, имеющая своим референтом страну (регион проживания) и ее жителей. Данная единица является объектом оценки и основным элементом (MAIN), подвергающимся модификациям. В качестве типов связей для

[207] с незначительными расхождениями в аспекте прямой/опосредованной референции.

[208] Фрагмент базы представлен в приложении. Для удобства восприятия исходный формат базы данных заменен на соответствующий ему формат xml.

оценочных отношений основного элемента используются следующие — синонимическая (SYN), агентивная (AGENT), общая (GENERIC), модификатор (MODIFIER). При комбинации признаков используется служебный тип MULT, который позволяет дополнять атрибут типа (или подтипа). Общая оценочная связь соответствует наименованию понятия (название фобии). Агентивная связь обозначает деятеля, совершающего действие этнофобного характера. Модификаторная связь описывает дополнительный оценочный параметр, который может выражаться действием или свойством субъекта. Параметр ID указывает на порядковый номер внутри блока. Параметр POS обозначает часть речи, к которой принадлежит слово. Параметр SENTIMENT описывает оценку объекта, приписываемую данным словом. Параметр INT соответствует интенсивности оценки (т.е. ее силе). В параметре VALUE задается лемматизированная форма слова.

Общую оценку, обозначаемую параметром GENERIC, можно проследить на приведенных ниже примерах. *Как все эти годы на выходе была русофобия, так и сейчас на выходе русофобия.*(объект оценки — Россия/русский). *Вот у нас почемуто нет укрофобии, нет латвофобии, литвофобии, германофобии, финофобии, американофобии.*[209] (объекты — Украина/украинец, Латвия/латвиец, Литва/литовец, Германия/германец, Финляндия/финн). *В регионе, а особенно у стран, которые непосредственно граничат с Китаем – это Казахстан и Кыргызстан – довольно высок уровень т.н. «китаефобии» или иначе «чайнафобии». О "чайнофобии", стратегии Пекина и китайском языке*[210]. (объект — Китай/китаец).

Носитель оценочного действия, описываемый при помощи параметра AGENT, может обозначаться не только при помощи собственно лексем с компонентом *–фоб*, но и другими способами, что можно продемонстрировать на следующих примерах. *Судя по стене в Одноклассниках, папаня тот еще*

---

209 https://www.facebook.com/764151920310984?comment_id=764812533578256

210 http://azh.kz/ru/news/view/19514

украинофоб и "антифашист", - сообщает журналист.[211] (объект — Украина/украинец). *Это человек, известный в мире как русофоб ещё с доперестроечных времён. Если сравнивать известного в кругах правозащитников политического заключённого не единожды, диссидента, антисоветчика Джамилева, и вполне успешного такого советского чиновника Чубарова, то понятно в чью сторону больше доверия скапливалось.*[212] (объекты — Россия/русский, СССР). *Вы, тролли, настолько ущербны, что не понимаете, что критикуя представителя нашего народа за малейшую провинность, идете на поводу не только и не столько у чеченофобов, сколько у кавказофобов, и, следовательно, у врагов России.*[213] (объекты — Чечня/чеченец, Кавказ/кавказец).

Модификаторы-прилагательные довольно часто игнорируются разработчиками отечественных систем автоматической обработки естественного языка при использовании зарубежных технологий, направленных на анализ английского языка, по причине большей вариативности средств передачи референциальных значений в текстах на русском языке. Это приводит к значительному падению полноты выдачи информационно-аналитических систем. Приведем примеры модификации оценочных агентов, которые обозначаются параметром MODIFIER. *Благо он неплохо разбирается в мусульманских течениях - несколько лет назад мне довелось читать его аналитическую записку о мусульманских сектах в Крыму, которые проводили свою русофобскую и антиправославную политику по линии меджлиса крымско-татарского народа и подконтрольного ему ДУМ Крыма", - отметил собеседник портала.*[214] (объект — русский, опосредованный объект — Россия). *Конашенков связал появившуюся информацию и с тем, что владельцем изданий, первыми сообщивших о "захвате", является олигарх Игорь Коломойский, с которым связывают антироссийскую пропаганду и другие действия против Москвы.*[215] (объект — Россия). *По части истерии, носящей укрофобский характер, российские СМИ дают фору всем*

---

211 http://www.newsru.com/russia/22aug2014/bmd.html
212 http://actualcomment.ru/daycomment/1494/
213 http://www.grozny-inform.ru/main.mhtml?Part=11&PubID=54515
214 http://www.newsru.com/religy/17sep2014/prokurorskaya_proverka.html
215 http://www.newsru.com/russia/22aug2014/bmd.html

мировым СМИ.[216] (объект — Украина/украинец). *Европобская "Европа за свободу и демократию" избрала своим кандидатом родившуюся в Сомали активистку за права женщин Айаан Хирси Али, ставшую известной на Западе и недавно получившую гражданство США.*[217] (объект — Европа). *Денис , Посмотрите,как на фоне амерофобской истерии,из Путина сделали эталон противозападной политики.*[218] (объект — Америка/США).

Существенным представляется следующее положение: база данных принципиально открыта и может быть использована для описания других типов ксенофобонимов. Например, параметром GENERIC для объекта ислам может быть описано существительное *исламофобия* в следующем контексте: *Комментируя гневный пост бывшего французского госсекретаря по делам семьи, глава французского бюро по наблюдению за исламофобией Абдалла Зекри заявил изданию France 24 , что заявленный Морано постулат о необходимости уважать ценности Французской Республики означает необходимость уважать ценности всех ее граждан - в том числе и мусульман.*[219]

Оценочные однореферентные слова в словаре-тезаурусе сгруппированы по словообразовательному принципу. Так для названия *беларус* в словаре существует 3 базовые оценочные словообразовательные модели: *беларусофоб* (модель «-фоб»), *беларусоненавистник* (модель «-ненавистник») и *небеларусофил* (модель «не-фил»). Порядок расположения этих моделей зависит от их оценочной силы. В настоящее время принята следующая градация «High->Extreme->Low». Внутри каждой словообразовательной модели лексемы расположены в порядке частей речи «Существительное-Прилагательное-Глагол». Внутри класса существительных вначале идет наиболее общее понятие, например, *беларусофобия*, затем его синоним *беларусофобство*, затем наименования деятеля *беларусофоб*.

---

216 https://www.facebook.com/770200476372795?comment_id=770228089703367
217 http://ria.ru/world/20140923/1025339630.html
218 http://vk.com/id2640759
219 http://www.newsru.com/religy/21aug2014/morano.html

3.4. Способ автоматического заполнения полей лексической базы данных по ксенофобонимам[220]

Качество систем автоматического извлечения мнений во многом определяется оценочным словарем [26]. В данном разделе мы рассмотрим один из способов расширения уже существующего словаря ксенофобонимов, сходного по структуре с представленным в предыдущем разделе, при помощи процедуры морфологического синтеза, которая была частично использована при расширении исходного списка первичных оценочных прилагательных в п. 3.1. Выбор технологии морфемного синтеза обусловлен его ориентированностью на языки флективного типа.

Традиционные подходы к составлению словарей оценочной лексики игнорируют базовое свойство языка как средства передачи информации — тенденцию к компрессии. Компрессия в языках флективного строя зачастую проявляется в виде замены потенциальных конструкций, содержащих имена существительные в генитиве, на конструкции с однокоренными прилагательными, а также однокоренными глаголами. В качестве примеров можно привести следующие предложения. *Срджа Трифкович, сторонник Слободана Милошевича, исламофоб (в интернете — писатель сербского происхождения, критик ислама — прим. Ред.).*[221] *Политический редактор известного издания Huffington Post в Великобритании призвал к введению санкций против СМИ, порочащих ислам и распространяющих исламофобские суждения, а также к их бойкоту.*[222] *Стоит напомнить, что Санторум известен своими исламофобскими настроениями и не раз высказывался оскорбительно об исламе и мусульманах.*[223] *«Центр борьбы с исламофобией не делает различий между критикой ислама и агрессивными действиями, тогда как критика христианской религии не рассматривается как христианофобский поступок.*[224] Рассматриваемая методика, таким образом, направлена на увеличение процента

---

[220] в основе раздела лежат работы автора [115] и [23].
[221] http://stockinfocus.ru/2014/11/14/rasshifrovka-vystuplenij-na-sovbez-oon/
[222] http://www.islamnews.ru/news-440705.html
[223] http://islam-today.ru/novosti/2014/11/12/eks-kandidat-v-prezidenty-ssa-sravnil-musulman-s-bolnymi-rakom/
[224] http://regions.ru/news/2488577/

покрытия автоматическим словарем произвольного текста, содержащего оценочную лексику.

С целью апробации методики морфемного синтеза было отобрано 20 словарных статей (по 10 на этнофобонимы и религиофобонимы[225]), что объясняется следующими факторами. Во-первых, существует значительный дисбаланс между количеством этносов и конфессий, представленных в словаре-тезаурусе. Во-вторых, данная выборка отражает все трудности, с которыми пришлось столкнуться при экспериментах по автоматической генерации каждого из разделов словаря.

3.4.1. Технология морфемного синтеза

Прежде чем перейти к описанию эксперимента по автоматическому синтезу словарных статей рассмотрим общие принципы использования морфемной информации в системах компьютерной обработки текстов. Морфемная информация может использоваться на этапах анализа и синтеза текстов. Целью морфемного анализа словоформ является получение информации о морфемной структуре словоформы, а также идентификация значений выделенных морфем. Автоматический анализ текста на морфемном уровне к исследованиям по компьютерной лингвистике привлекается редко. В основном, вычленение морфем проводилось для сложных слов немецкого языка [146]. В подобных случаях морфемный анализ может рассматриваться как этап морфологического (или даже первичного синтаксического) анализа текста, т.к. он направлен на получение информации о грамматических отношениях между элементами сложного слова. Для русского языка морфемный анализ использовался в рамках проекта словообразовательной разметки Национального корпуса русского языка [76]. Целью данного проекта было выявление наиболее продуктивных словообразовательных моделей в текстах на русском языке в диахронической перспективе. В обоих случаях использовались заранее известные списки основ (для немецкого языка в работе Меньшикова) или приставок, префиксоидов и

[225] социофобонимы в данном разделе не рассматриваются, т.к. их словообразовательные модели идентичны рассматриваемым по формальным признакам.

суффиксов (для русского языка в работе Березуцкой и Тагабилевой). Сходный подход к выделению оценочных суффиксов применялся в рамках работ по семантической разметке Национального корпуса русского языка [130]. В отличие от нашего (узкого) понимания оценочных отношений как противопоставления «плохой-хороший», в цитируемой работе категория оценки понимается более широко — в интересующем нас случае — как противопоставление по шкале «маленький-большой» (на примере диминутивных и аугментативных слов). Перейдем к рассмотрению задачи морфемного синтеза текста. Целью морфемного синтеза является получение нового слова (деривата) с заданными характеристиками из уже существующего слова. Необходимо отметить, что проблема морфемного синтеза слов для русского языка решалась, в основном, в рамках работ по машинному переводу [19]. Существенным недостатком разработки коллектива под руководством З.М. Шаляпиной является словарный [ручной] подход к вычленению суффиксального словообразования. Разработка словарей оценочной лексики, в которых учитывалась бы значительная лингвистическая информация, представляется бесперспективным подходом по причине излишней трудоемкости. Поэтому несмотря на некоторое сходство нашего подхода с подходом разработчиков системы RUMORS мы изначально отказались от расширения существующего словника словообразовательной информацией.

3.4.2. Выявление словообразовательных моделей этно- и религиофобонимов

Как уже было указано выше, в словаре-тезаурусе представлены 3 части речи — имена существительные и прилагательные, а также глаголы. Рассмотрим словообразование в каждой из этих частей речи по отдельности.

Имена существительные представлены в словаре четырьмя различными словообразовательными классами лексем. Первый класс (объект оценки) задается извне и не подвергается словообразовательному анализу. Второй класс можно обозначить как класс деятеля. Третий класс — понятие. Четвертый

словообразовательный класс составляют отрицания положительного отношения к объекту.

Компонентный анализ класса деятеля выделяет две квазисуффиксальные словообразовательные модели — *«-фоб»* и *«-ненавистник»*. При добавлении квазисуффикса к основе объекта с ней происходят следующие изменения: 1) отсекается суффикс (при наличии, например, *русский — русофоб*, *чеченец — чеченофоб*, но *казах — казахофоб*); 2) конечная йотированная гласная или «*й*» заменяется на «*е*» (например, *Россия — россиефоб*); 3) новая форма образуется от основы множественного числа (например, *христианин — христианоненавистник*). Следует отметить, что возможна и иная трактовка пункта 3 — как частного случая пункта 1. Для образования форм женского рода используется квазисуффикс *-ка* для модели «*-фоб*» и *-ца* для модели «*-ненавистник*» (при этом конечная *-к* отбрасывается).

Класс, обозначающий понятия, в своем словообразовании модифицирует класс деятеля. Для модели «*-фоб*» допустимы квазисуффиксы *-ия* (*туркофоб — туркофобия*) и *-ство* (*туркофоб — туркофобство*), которые присоединяются к концу слова. Для словообразовательной модели «-ненавистник» возможно использование только квазисуффикса *-ство*, при этом основа слова подвергается следующим изменениям: 1) конечная *-к* заменяется на *-ч*; 2) к модифицированной основе добавляется гласная *-е*. В качестве примера можно привести следующий: *татароненавистник — татароненавистничество*.

Класс отрицания положительного отношения к объекту полностью соответствует по суффиксальному словообразованию с моделью «*-фоб*» (кроме квазисуффикса *-ство*, перед которым добавляется *-ь*), при этом «*-фоб*» заменяется на «*-фил*». В части префиксальной деривации данным классом используется отрицательная приставка *не-*, например *русский — нерусофильство*.

Для того, чтобы выявить словообразовательные схемы имен прилагательных, мы проанализировали словообразовательные суффиксы, характерные для данной части речи, в морфологическом словаре информационно-поисковой системы Stocona (около 40 тысяч лемм) [152]. В результате

компонентного анализа были получены следующие суффиксы и сочетания суффиксов, характерные для прилагательных, обозначающих религиофобонимы и этнофобонимы: *-йский*, *-овский*, *-евский*, *-ский*, *-цкий*. Кроме того, анализ показал, что данные суффиксальные словообразовательные модели можно свести к двум: *-цкий* и *{C}+ский*, где *C* — согласная буква. Использование согласных букв способствует отсеканию прилагательных, образованных от однокоренных слов, которые обозначают, например, род деятельности (терроризм — террористич*ес*кий, но террорист — террорис*тс*кий или (редкая форма) террорист*овс*кий). Исключение составляет группа слов с компонентом *-ненавистник*. Корпусный анализ ксенофобонимов продемонстрировал, что одушевленные имена существительные женского рода из рассматриваемой выборки не имеют собственных моделей образования прилагательных. Так, пары *исламоненавистник-исламоненавистница* и *англофоб-англофобка* имеют общие дериваты-прилагательные *исламоненавистнический* и *англофобский*. Например, *С тех пор меня по сей день клеймят "московским чеченцем, защитившем в своем журнале своего приятеля-убийцу Расула Мирзаева" и даже самый кавказофобский и исламоненавистнический канал России - НТВ предлагал принять участие в съемке программы "Максимум", посвященной Расулу.*[226] *"Мистер Питкин" тоже страшно англофобский фильм.*[227] Данное наблюдение может быть объяснено двумя обстоятельствами. С одной стороны, сочетания из более четырех согласных звуков подряд представляют собой труднопроизносимый комплекс звуков. С другой стороны, прилагательное, образованное от существительного мужского рода, может использоваться для референции к однокоренному существительному женского рода.

Глагольное словообразование этнофобонимов и религиофобонимов не отличается разнообразием моделей. Единственной обнаруженной нами моделью является суффиксальный комплекс *-вать*, который добавляется к основе «понятийного» класса на *-фобство*. Наиболее частым употреблением данной

---

226 http://gareeb.livejournal.com/45804.html

227 http://dpmmax.livejournal.com/186656.html?thread=39568416#t39568416

модели являются причастные формы. Например, форма «*русофобствующий*» в поисковой выдаче информационно-поисковой системы Google© в 5,4 раза частотнее глагола в форме «*русофобствовал*» (12800 вхождений против 2370 вхождений). Например, предложение *Горноалтаец Манзыров под ником «нацик» русофобствовал в Интернете.*[228] ощущается более искусственным по сравнению с предложением *Патриарх Руси, Папа Римский, Патриарх русофобствующий и жуткая ошибка кардинала*[229].

Таким образом, можно сделать вывод о доминировании суффиксальной модели словообразования в этно- и религиофобонимах. Деривационные компоненты позволяют провести разграничение не только по частям речи, но и по силе оценки.

3.4.3. Отличия в базе данных при ручном и автоматическом заполнении полей

Ручное заполнение полей базы данных по ксенофобонимам способствует более компактному представлению информации. Рассмотрим причину этого на примере словарной статьи по этнофобомимам. Отличительной чертой этнофобонимов следует признать возможность замены названия этнической группы на название страны, ассоциируемой с данной этнической группой. Например, слово *англофобия,* образованное от топонима *Англия*, имеет тот же референт, что и *англичанофобия*. В результате мы получаем единую статью для двух референтов, автоматический морфемный вывод которых один из другого не представляется возможным в рамках описанного подхода. Подобное различие ведет к увеличению числа полей в автоматически сгенерированной базе данных. С другой стороны, при автоматической генерации сводится к минимуму человеческий фактор при заполнении полей базы.

3.4.4. Алгоритм использования морфемного синтеза при автоматическом заполнении полей базы-данных (тезауруса) по ксенофобонимам

---

228 http://pn14.info/?p=158687

229 http://www.zarusskiy.org/initiatives/2008/12/09/oshibka/

На вход программе подается входная строка базы данных, содержащая наименование объекта (с указанием соответствующего атрибута). Программа оперирует следующими понятиями: набор атрибутов, набор квазиаффиксов, набор связей квазиаффиксов с атрибутами.

Под набором атрибутов нами понимается список настраиваемых полей базы данных, к которым относятся: часть речи, тип связи с объектом, оценка, сила оценки и лемматизированная форма слова. Набор квазиаффиксов состоит преимущественно из квазисуффиксов (*-фоб*, *-ство*, и др.) и префикса *не-*. Ключевым понятием в программе является набор связей квазиаффиксов с атрибутами. Этот набор не только позволяет приписать квазиаффиксу конкретное значение атрибута, но и описывает последовательность выполнения действий с квазиаффиксами.

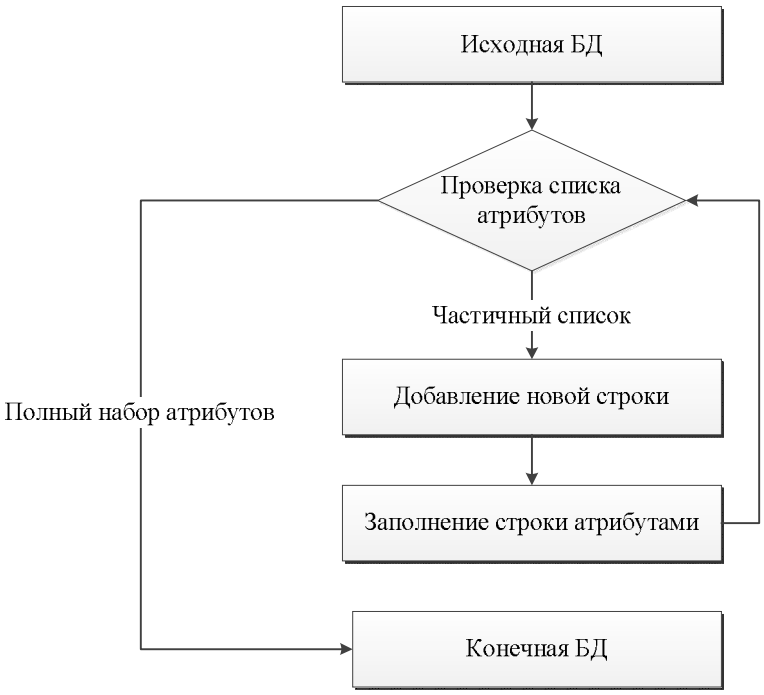

Рисунок 26. Блок-схема пополнения словаря

На рисунке (Рисунок 26) указана общая блок-схема пополнения словаря-тезауруса. Словарь-тезаурус представляет собой реляционную БД, единицами которой являются БД-словарные статьи. Алгоритм заканчивает работать при достижении конца списка объектов. Алгоритм может быть реализован при помощи SQL-скриптов.

3.4.5. Недостатки предлагаемого подхода

У предлагаемого нами подхода существует ряд недостатков, сказывающихся, в первую очередь, на полноте выдачи. Алгоритм не учитывает языковую игру, например, «наглый+Англия => Наглия». Например, *ЛОНДОН, **НАГЛИЯ***.[230] Существует ряд религий, которые имеют собственные (не выводимые из названия) религиофобонимы. Например, для *иудаизма* в качестве религиофобонима используется лексема *юдофоб*: *Скольких убийц в их кровавом деле поддержали такие теоретики юдофобии, как Федор Достоевский или Рихард Вагнер.*[231] Проблемой онтологического характера является наличие в подавляющем большинстве религий специальных терминов, обозначающих людей, которые по определенным признакам представляют угрозу для данной религии. Для христианской религии — это термины *ересь*, *схизма* и *идолопоклонство*. Для ислама — это названия противников внутри ислама (*такфир*) и вне ислама (*кяфир*). Все названные термины имеют свои дериваты. Приведем несколько примеров, доказывающих это утверждение: *Со ссылкой на "такфири", боевиков - суннитов, которые провозглашают последователей других направлений ислама неверными и, следовательно, законными целями в священной войне, на сайте аятоллы говорится: "Мусульмане должны быть осведомлены о такфири и их общих целях с высокомерными – посеять войну в мусульманском мире. Шииты и сунниты должны быть бдительными".*[232] *Он также заявил, что США, Израиль и Саудовская Аравия создали радикальный такфирский проект ("такфир" — в исламе обвинение в неверии, прим. ред.), оказавшись перед лицом исламского пробуждения, подразумевая их причастность к созданию салафитской экстремистской группировки "Исламское государство Ирака и Леванта" (ИГИЛ).*[233] *Ты увидишь, что город Махачкала называется Шамиль-Калой, ты увидишь какие-то фантастические новости: моджахеды приняли бой*

---

230 http://observer-lj.livejournal.com/1139969.html
231 http://mnenia.zahav.ru/Articles/702/krov_pasternaka
232 http://www.mignews.com/news/society/world/190614_121926_49030.html
233 http://9tv.co.il/news/2014/06/18/178287.html

против кяфиров.[234] Существуют также и общие нереферентные для отдельной конфессии негативные оценочные слова, например, *богохульство*. Хотя в задачи данной статьи не входит рассмотрение подобных слов, референтных для определенного класса объектов, необходимо отметить, что учет этой группы слов обязателен для комплексной системы автоматической идентификации противоправного контента (или указаний на его проявления). В качестве примера можно привести следующее предложение: *Да, эти вандальские выходки носили совершенно явный богохульный характер, однако во Франции нет закона с наказанием за богохульство*.[235] Существенной проблемой нашего подхода является неучет возможности вхождения этно- и религиофобонимов в состав сложных слов, например, *Однако невежественные социалисты-христианофобы времен Франсуа Олланда окончательно об этом забыли.*[236]

Для решения указанных проблем можно сократить объем генерируемой информации, например, подавая на вход форму «*юд*» для генерации лексемы «*юдофоб*». БД в таком случае будет подвергаться обязательному постредактированию. Анализ сложных слов можно учитывать при помощи поиска подстроки, что, несмотря на замедление скорости обработки, приведет к существенному росту покрытия БД.

При создании аналогичной базы данных на английском языке [23] было выявлено, что предлагаемый подход необходимо расширить как в части используемых частей речи (добавить наречия), так и в части прилагательных с другой референтной моделью (что справедливо и для русского языка, где модель *-ный* может употребляться в тех же контекстах, что и модель на *-ский*[237]).

---

[234] http://thekievtimes.ua/society/374806-konchita-pobeda-nad-evropoj.html

[235] http://armtoday.info/default.asp?Lang=_Ru&NewsID=117432

[236] http://armtoday.info/default.asp?Lang=_Ru&NewsID=117432

[237] что не отражено в исходном словаре, по которому определялся список квазисуффиксов. Пары прилагательных (-ский/-ный) присутствуют в текстах, например, *террористический*/*террористичный*, *русофобный* (http://poleev.livejournal.com/583.html)

Выводы к Главе 3

1. Создание первичного словаря оценочной лексики для широкой предметной области целесообразно проводить на основе неспециализованных массивов текстов сверхбольшого объема (свыше 10 млрд. словоупотреблений). Применение статистических процедур при формировании списка слов, описывающих предметную область, позволяет отсечь низкочастотные и нерелевантные синонимы.

2. Корпусы текстов, предназначенные для тестирования систем автоматического извлечения мнений, должны описывать общую модель мнения. Подобный формат описания, реализуемый при помощи языка XML, позволяет определить не только качество системы в целом, но и учесть качество выполнения каждой отдельной подзадачи.

3. Однореферентные оценочные слова, наиболее значимыми из которых являются ксенофобонимы, представляют собой один из пластов имплицитной оценочной лексики. Наиболее частым случаем использования ксенофобонимов в тексте следует признать приписывание человеку свойств, которые считаются неприемлемыми в российском правовом поле.

4. Наиболее адекватным способом формального описания однореферентных оценочных слов является тезаурус. Тезаурус позволяет описать разнообразные лингвистические свойства, такие как принадлежность к части речи, референциальные и прагматические свойства однореферентных оценочных слов.

5. Технология морфемного синтеза позволяет порождать формы слов с заданными параметрами. Таким образом, она может применяться для автоматического пополнения как простых списков слов, например, списков первичных оценочных слов, так и тезаурусных ресурсов, подобных тезаурусам по этно- и религиофобонимам.

Заключение

Многообразие задач, стоящих перед компьютерной лингвистикой в области анализа текстов в Интернет-пространстве, диктует новые требования к лингвистическому обеспечению систем автоматической обработки текстов. В последнее время все большую роль в инструментах автоматической обработки текстов играют компоненты автоматического извлечения мнений.

В зависимости от задач, стоящих перед модулем автоматического извлечения мнений, меняется состав его компонентов и их структура. Поэтому ключевую роль в успешном функционировании системы в целом начинает играть архитектурная составляющая. На первое место выдвигается формальная модель мнения, которая уникальна для каждого класса задач, решаемых при помощи технологии автоматического извлечения мнений.

Особенное внимание разработчики систем обработки текста уделяют сравнению собственных разработок с разработками конкурентов, что ведет к постоянному совершенствованию как методик сравнения, так и ресурсов, привлекаемых для этой цели.

Проведенное нами исследование позволило прийти к следующим выводам:

1. В современных системах автоматического извлечения мнений из текстов на русском языке применяются, преимущественно, статистические методы, а задача извлечения мнений рассматривается как частный случай классификации текстов.
2. Зародившись как междисциплинарное прикладное направление, находящееся на стыке компьютерной лингвистики и искусственного интеллекта, автоматическое извлечение мнений в настоящий момент не обладает устоявшейся терминологической системой.
3. В системах автоматической обработки текстов, ориентированных на язык Интернет-коммуникации, большое значение приобретает компонент предобработки текстов. В состав данного компонента необходимо включать инструменты анализа метатекстовой

информации, которые позволяют учесть априорно заданную прагматическую информацию.

4. Обязательным компонентом лингвистического обеспечения систем автоматического извлечения мнений является фильтрация нерелевантных объектов анализа. Фильтрация данных объектов осуществляется на основе структурных и семантических критериев.
5. Формальная модель мнения, используемая в системе автоматического извлечения мнений, должна учитывать особенности функционирования коммуникации в Интернет-пространстве. Поэтому она должна включать информацию об источнике мнения и канале передачи сообщения.
6. Наиболее адекватным способом описания объектов мы считаем объект-аспектную модель, которая позволяет учитывать свойства объекта мнения. При этом данная модель не является универсальной и не подходит для любого типа текстов.
7. Структура лингвистического обеспечения систем автоматического извлечения мнений в значительной степени зависит от конкретной задачи анализа текста. Наиболее сложной задачей мы считаем идентификацию противоправного контента, для решения которой требуется не только наличие особого набора специализированных оценочных словарей, но и значительно более жесткие требования к фильтрации входных данных.
8. При оценке качества систем автоматического анализа мнений мы считаем целесообразным использовать размеченные корпусы текстов, разметка которых полностью отражает модель мнения. Такая разметка позволяет определять качество работы не только системы в целом, но и отдельных компонентов.
9. Для автоматического определения некоторых случаев имплицитной оценки необходимо учитывать референтные свойства оценочных слов. Среди однореферентных оценочных слов особое социальное

значение имеют ксенофобонимы, обозначающие отрицательную оценку объекта заданного класса (например, этнической, религиозной или социальной группы и отдельного человека). Наиболее адекватным инструментом для формального представления однореферентных оценочных слов мы считаем тезаурус.

Литература

1. Balahur, A., Steinberger, R. Rethinking Sentiment Analysis in the News: from Theory to Practice and back // Proceedings of the '1st Workshop on Opinion Mining and Sentiment Analysis'. Sevilla, 2009. p. 1—12. Retrieved from: http://langtech.jrc.it/Documents/09_WOMSA-WS-Sevilla_Sentiment-Def_printed.pdf (26.02.2016).
2. Balahur, A., Steinberger, R., Kabadjov, M., Zavarella, V., van der Goot, E., Halkia, M., Pouliquen, B., Belyaeva, J. Sentimet Analysis in the News // 7th International Conference on Language Resources and Evaluation (LREC 2010). — pp. 2216—2220.
3. Bermingham, A., Smeaton, A. On Using Twitter to Monitor Political Sentiment and Predict Election Results // Proc. of SAAIP 2011, pages 2—10.
4. Biber, D., Conrad, S., Rippen, R. Corpus linguistics. Investigating language structure and use. — Cambridge: CUP, 2007.
5. Bučar, J., Povh, J. Sentiment Analysis in Web text: An overview. 2013. — In Recent Advances in Information Science. Proceedings of the 7th European Computing Conference (ECC '13). — Dubrovnik, 2013. — pp. 154—159.
6. Cambria, E., Hussain, A. Sentic Computing: Techniques, Tools, and Applications. 2nd edition. Springer, 2012. — xxiv, 155 pp.
7. Carbonell, J. G. Towards a Process Model of Human Personality Traits // Artificial Intelligence, Vol. 15, No. 1,2, November 1980, pp. 49—74.
8. Cardie, Cl. Review: Bing Liu. Sentiment Analysis and Opinion Mining // Computational Linguistics. 2014. Volume 40, Number 2. — pp. 511—513.
9. Carstensen K.-U. et al. (Hrsg.) Computerlinguistik und Sprachtechnologie: Eine Einfuerung. 3. ueberarbeitete und erweiterte Auflage. — Elsevier: Spektrum akademischer Verlag. — Berlin, etc., 2010. — xvi 736 S.
10. Chetviorkin, I. I., Loukachevitch, N. V. Extraction of Russian Sentiment Lexicon for Product Meta-Domain // In Proceedings of COLING 2012: Technical Papers , pages 593—610.
11. Corpus Linguistics. An International Handbook. Ed. by Lüdeling, A. and Kytö, M. [Text] : / A. Lüdeling, M. Kytö (eds.) — Berlin: DE GRUYTER MOUTON, 2008—2009. — XII, 1553 p.
12. De Smedt, T. and Daelemans, W. `“Vreselijk mooi!” (terribly beautiful): A Subjectivity Lexicon for Dutch Adjectives.' // In Proceedings of the Eight International Conference on Language Resources and Evaluation (LREC'12), Istanbul, Turkey, European Language Resources Association (ELRA), pp. 3568—3572, 2012.
13. Eidelman, V. Unsupervised Feature-Rich Clustering // Proceedings of COLING 2012: Posters, pages 275–286, COLING 2012, Mumbai, December 2012.
14. Fahrni, A., Klenner, M. Old Wine or Warm Beer: Target-Specific Sentiment Analysis of Adjectives // In: Proc.of the Symposium on Affective Language in Human and Machine, AISB 2008 Convention, 1st-2nd April 2008. University of Aberdeen, Aberdeen, Scotland, 60 - 63, 2008. [Электронный ресурс]:

http://www.zora.uzh.ch/8810/2/OldWineOrWarmBeerV.pdf . Дата доступа: 15.03.2016.
15. Hatzivassiloglou, V., McKeown K. R. Predicting the semantic orientation of adjectives // Proceedings of the 35th Annual Meeting of the Association for Computational Linguistics and of the 8th Conference of the European Chapter of the Association for Computational Linguistics, Madrid, Spain, 7—12 July 1997. — pp. 174—181.
16. Jakubíček, M., Kilgarriff, A., Kovář, V., Rychlý, P., Suchomel, V. The TenTen Corpus Family. // In Abstract Book of 7th International Corpus Linguistics Conference, Lancaster, July 2013. — pp. 125—127.
17. Joglekar, S.T., Kadam, S.A. Opinion Mining: A Tool for Market Intelligence. — International Journal of Advanced Research in Computer Science and Software Engineering, Volume 5, Issue 7, July 2015. — pp. 370—375.
18. Kadam, S.A., Joglekar, S.T. Sentiment Analysis: An Overview. — International Journal of Research in Engineering & Advanced Technology, Volume 1, Issue 4, Aug-Sept, 2013. [Электронный ресурс. Дата доступа: 05.03.2016]. http://www.ijreat.org/Papers%202013/Issue4/IJREATV1I4016.pdf
19. Kanovich M., Shalyapina Z. The Rumors System Of Russian Synthesis // In Proceedings of COLING-1994.
20. Kilgarriff, A., Rychly, P., Smrz, P., Tugwell, D. The Sketch Engine. // In Proc. of EURALEX 2004, Lorient, France; — pp. 105—116.
21. Klenner, M. (2015). Verb-centered Sentiment Inference with Description Logics. In: 6th Workshop on Computational Approaches to Subjectivity, Sentiment & Social Media Analysis, Lissabon, 17 September 2015 — 21 September 2015, 134—140.
22. Klenner, M., Petrakis, S; Fahrni, A. Robust Compositional Polarity Classification // Proceedings of RANLP`09, Borovets, Bulgaria, 2009. —pp. 180—184.
23. Kulikov ,S. Detecting Implicit Opinions with a Target-specific Thesaurus // In Abstracts of Clin25, Antwerpen, Belgium. 2015. — p. 24.
24. Kulikov, S. Opinion Lexicon Organisation in a rule-based Sentiment-analysis System // In Abstracts of Clin24, Leiden, the Netherlands. 2014. — p. 78.
25. Kulikov, S. What is web-based machine translation up to? // Proceedings of TRALOGY-2011 «Métiers et technologies de la traduction : quelles convergences pour l'avenir?» Paris, France, 2011 Electronic publication: http://lodel.irevues.inist.fr/tralogy/index.php?id=118
26. Liu, B. Sentiment Analysis and Opinion Mining. Synthesis Lectures on Human Language Technologies #16. 2012. — xiv, 165 p.
27. Liu, H. et al. A Model of Textual Affect Sensing using Real-World Knowledge // Proceedings of the IUI 2003, Miami, FL, USA, 2003. — pp. 125—132.
28. Liu, Q., Gao, Zh., Liu, B., Zhang, Y. Automated Rule Selection for Aspect Extraction in Opinion Mining // Proceedings of the Twenty-Fourth International Joint Conference on Artificial Intelligence (IJCAI 2015). — Buenos Aires, Argentina, 2015. — pp. 1291—1297.

29. Luyckx, K., Vaassen, F., Peersman, C., Daelemans, W. Fine-Grained Emotion Detection in Suicide Notes: A Thresholding Approach to Multi-Label Classification // Biomedical Informatics Insights 2012:5 (Suppl. 1). — pp.61—69.
30. Ma, T, Wan, X. Opinion Target Extraction in Chinese News Comments // Coling 2010: Poster Volume, pages 782–790, Beijing, August 2010.
31. Maks, I., Izquierdo, R., Vossen, P. Automatic generation of sentiment lexicons in five languages // In Abstracts of Clin24, Leiden, the Netherlands. 2014. — p.29.
32. McShane, M. A Multi-Faceted Approach to Reference Resolution in English and Russian // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог» (Бекасово, 4 — 8 июня 2014 г.). Вып. 13 (20). — М.: Изд-во РГГУ, 2014. —С. 391—409.
33. Measuring the Information Society Report 2015. — Geneva: International Telecommunication Union, 2015. — xiv, 235 p.
34. Medhat, W., Hassan, A., Korashy, H. Sentiment analysis algorithms and applications: A survey // Ain Shams Engineering Journal. Volume 5, Issue 4, December 2014, Pages 1093—1113.
35. Mitkov, R. (ed.) The Oxford Handbook of Computational Linguistics. [Text] : / R. Mitkov (ed.). — Oxford, etc. OUP, 2003. — xx, 786 p.
36. Moilanen, K., Pulman, S. Multi−entity Sentiment Scoring // In Proceedings of the International Conference RANLP−2009. — Pages 258–263. — Borovets, Bulgaria. September, 2009. Association for Computational Linguistics.
37. Moilanen, K., Pulman, S. Sentiment composition //Proceedings of RANLP`07, Borovets, Bulgaria, 2007. — pp. 378—382.
38. Moilanen, K., Pulman, S., Zhang, Y. Packed Feelings and Ordered Sentiments: Sentiment Parsing with Quasi−compositional Polarity Sequencing and Compression // In Proceedings of the 1st Workshop on Computational Approaches to Subjectivity and Sentiment Analysis (WASSA 2010) at the 19th European Conference on Artificial Intelligence (ECAI 2010). — Pages 36—43. August, 2010.
39. Paassen, B., Stöckel, A., Dickfelder, R., Göpfert, J.P., Kirchhoffer, T., Brazda, N., Müller, H.W., Klinger, R., Hartung, M., Cimiano, P. Ontology-based Extraction of Structured Information from Publications on Preclinical Experiments for Spinal Cord Injury Treatments // Proceedings of Third Workshop on Semantic Web and Information Extraction, pages 25—32, Dublin, Ireland, 24 August, 2014.
40. Pang, B., Lee, L. Opinion mining and sentiment analysis. //Foundations and Trends in Information Retrieval. Vol. 2. 2008. — 137 p.
41. Pang, B., Lee, L., Vaithyanathan, S. Thumbs up? Sentiment classification using machine learning techniques. // Proceedings of the 2002 Conference on Empirical Methods in Natural Language Processing, Philadelphia, Penn., 6—7 July 2002, — pp. 79—86.
42. Petrakis, S., Klenner, M. (2011). Learning theories for noun-phrase sentiment composition. // In: 8th International NLPCS Workshop, Kopenhagen, 20 August

2011 — 21 August 2011, pp. 179-188. Retrieve from: http://www.zora.uzh.ch/53322/1/nlpcs.pdf (26.02.2016).
43. Rentoumi, V., Petrakis, S., Klenner, M., Vouros, G.A., Karkaletsis, V. United we stand: improving sentiment analysis by joining machine learning and rule based methods. // 7th International Conference on Language Resources and Evaluation (LREC 2010) pp. 1089—1094.
44. Schank, R. C., Riesbeck, C. K. (eds.) Inside Computer Understanding: Five Programs Plus Miniatures. [Text] : / R.C. Schank, C.K. Riesbeck. — New Jersey: Erlbaum, 1981. — 400 p.
45. SentiRuEval: разметка тональности характеристик (аспектов) объектов. Руководство асессора, 2015. [Электронный ресурс] https://drive.google.com/folderview?id=0B7y8Oyhu03y_fjNIeEo3UFZObTVDQXBrSkNxOVlPaVAxNTJPR1Rpd2U1WEktUVNkcjd3Wms&usp=drive_web (Дата доступа: 10.04.2016)
46. Solovyev, A. N., Antonova, A. Ju., Pazelskaia, A. G. Using Sentiment-analysis for text information extraction// Компьютерная лингвистика и интеллекуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 11 (18). Том 1. — М.: РГГУ, 2012. — С. 616—626
47. Somasundaran, S. Discourse-level relations for Opinion Analysis, PhD Thesis, University of Pittsburgh. [unpublished]
48. Sonntag, J., Stede, M. Sentiment Analysis: What's Your Opinion? // In Text Mining. From Ontology Learning to Automated Text Processing Applications. Festschift in Honor of Gerhard Heyer. Biemann, C., Mehler, A. (eds.). Theory and Applications of Natural Language Processing. Book Series. — Heidelberg, etc. Springer, 2014. — pp. 177—199.
49. Taboada, M., Brooke, J., Tofiloski, M., Voll, K., Stede, M. Lexicon-Based Methods for Sentiment Analysis // Computational Linguistics, Vol. 37, №2, 2011. — pp. 267—307.
50. Taboada, M.T. Sentiment Analysis: An Overview of Linguistics. — Annual Review of Linguistics, 2016. №2. — pp. 325—347.
51. Turney, P. Thumbs up or thumbs down? Semantic orientation applied to unsupervised classification of reviews. //Proceedings of the 40th Annual Meeting of the Association for Computational Linguistics, Philadelphia, Penn., 7—12 July 2002. — pp. 417—424.
52. Vettigli, G. et al. Extracting Cause-Effect Relations in Natural Language Text // Proc. of CLIN 2014, p. 57.
53. Voutilainen, A. Part-of-Speech Tagging // Mitkov R. (ed.) The Oxford Handbook of Computational Linguistics. — Oxford, etc., 2003.
54. Wanner, F. et al. (2009) Visual Sentiment Analysis of RSS News Feeds Featuring the US Presidential Election in 2008 //Workshop on Visual Interfaces to the Social and the Semantic Web. 2009. [Electronic resource] [Last retrieved 26.05.2011] [url: http://sunsite.informatik.rwth-aachen.de/Publications/CEUR-WS/Vol-443/paper7.pdf]

55. Wiebe, J., Wilson, T., Bruce, R, Bell, M., Martin, M. Learning subjective language. // Computational Linguistics 30 (3), 2004: pp. 277—308.
56. Wiebe, J.M. Tracking point of view in narrative. // Computational Linguistics, 20 (2), 1994: pp. 233—287.
57. Wiebe, J.M., Wilson, T., Cardie, **C.** Annotating expressions of opinions and emotions in language //Language Resources and Evaluation, volume 39, (2005). — issue 2—3, pp. 165—210.
58. Wilks, Y., Bien, J. Beliefs, Points of View, and Multiple Environments //Cognitive Sciences 7, (1983). — pp. 95—119.
59. Xue, N., Poesio, M. (eds.) LAW IV. Proceedings of the Fourth Linguistic Annotation Workshop. [Text] : / N. Xue, M. Poesio. — Stroudsburg, PA. ACL, 2010. — xvi, 292 p.
60. Zaenen, A., Polanyi, L. Contextual valence shifters. // J. Shanahan, Y. Qu, and J. Wiebe (eds.), Computing Attitude and Affect in Text: Theory and Applications.//The Information Retrieval Series, Vol. 20, 2004. — Dordrecht, Springer. — pp. 1—9.
61. Zagibalov, T.E. Unsupervised and Rnowledge-poor Approaches to Sentiment Analysis. PhD Thesis. — Brighton: Sussex University, 2010. — xii, 181 p.
62. Zhekova, D. Towards Multilingual Coreference Resolution. — PhD Thesis, University of Bremen. 2013. — xxxiv, 299 pp.
63. Zhou, X., Wan, X., Xiao, J. Collective Opinion Target Extraction in Chinese Microblogs // Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing, pages 1840–1850, Seattle, Washington, USA, 18-21 October 2013.
64. Амирова, Т.А., Ольховиков, Б.А., Рождественский, Ю.В. История языкознания. [Текст] / Т. А. Амирова, Б. А. Ольховиков, Ю. В. Рождественский; под ред. С. Ф. Гончаренко. — 5-е изд., стер. — М.: «Академия», 2008. — 670 с.
65. Андреев, Н.Д. Машинный перевод и проблема языка-посредника// Вопросы языкознания, 1957. №5. — с. 117—121.
66. Андреев, Н.Д. Статистико-комбинаторные методы в теоретическом и прикладном языкознании. — Л.: Наука, 1967. — 404 с.
67. Апресян Ю.Д. (отв. ред.) Новый объяснительный словарь синонимов русского языка. М., 2004.
68. Апресян, Ю.Д. Избранные труды. [Текст] : в 2 т. Том I. Лексическая семантика. — 2-е изд., испр. и доп. — М., «Языки русской культуры», 1995. – 464 с.
69. Арутюнова, Н.Д. Типы языковых значений: Оценка. Событие. Факт. — М.: Наука, 1988. — 341 с.
70. Ахренова, Н.А. Разные страны, один Интернет? // Вестник Московского государственного областного университета. Серия: Лингвистика. 2009. № 2. — с. 186—191.
71. Белоногов Г.Г., Новоселов А.П. Автоматизация процессов накопления, поиска и обобщения информации. — М., 1979.

72. Белоногов, Г.Г. и др. Компьютерная лингвистика и перспективные информационные технологии : теория и практика построения систем автомат. обраб. текстовой информ. [Текст] : / Г. Г. Белоногов, Ю. П. Калинин, А. А. Хорошилов. — М.: Информационно-издат. агентство Русский мир, 2004. — 246 с.
73. Белоногов, Г.Г., Родионова, А.К. Морфологический словарь для автоматического анализа военных текстов // Сборник научных трудов №59. Отв. ред. Н.А. Криницкий. — М.: Министерство обороны СССР, 1965. — с. 119—211.
74. Беляева, Л.Н. Применение ЭВМ в лингвистических исследованиях и лингводидактике. — Л.: ЛГПИ им. А.И. Герцена. — 84 с.
75. Бенвенист, Э. Общая лингвистика. [Текст] : / Пер. с фр. / Под ред., с вступ. статьей и коммент. Ю. С. Степанова. — М.: «Прогресс» , 1974. — 446 с.
76. Березуцкая, Ю. Н., Тагабилева, М. Г. Словообразовательная разметка Национального корпуса русского языка: задачи и методы // В кн.: Компьютерная лингвистика и интеллектуальные технологии. По материалам ежегодной Международной конференции "Диалог" (2010) / Под общ. ред.: А. Е. Кибрик. Т. 9. Вып. 16. — М. : РГГУ, 2010. с. 499—506.
77. Бессмертный, И.А., Джалиашвили, З.О., Максимов, В.В., Маркин, Д.А. Лингвооценочное управление текстом // Материалы X Международной конференции "Применение новых технологий в образовании". — Троицк: Фонд новых технологий в образовании "Байтик", 1999. [Электронный ресурс] Дата доступа: 05.03.2015. http://bytic.trtk.ru/cue99M/cx1v6d3gym.html
78. Брунова, Е.Г. Методика составления оценочного лексикона для контент-анализа мнений // Language & Science. - Тюмень: ТюмГУ.– 2012. - Вып. 1. – URL: http://utmn.ru/docs/9317.pdf
79. Васечкин, Д.В., Зайцев, И.Н., Осипов, С.А. Система мониторинга средств массовых коммуникаций в Сети «Интернет». Подсистема обработки текстовой информации ИСМК. Частное техническое задание. — М., 2014. — 45 с. http://www.rfs-rf.ru/idc/groups/public/documents/grhc_native_files/017740.pdf (дата обращения: 01.10.14).
80. Васильев, Л.М. Методы современной лингвистики / Л. М. Васильев. — Уфа: БГУ, 1997. — 180 с.
81. Виноградов, В.А. Одушевленности / неодушевленности категория // Языкознание. Большой энциклопедический словарь. [Текст] : / В.Н. Ярцева и др. — 2-е изд. — М.: Большая Российская энциклопедия, 1998. — С. 342—343.
82. Виноградов, В.А. Собирательности категория // Языкознание. Большой энциклопедический словарь. [Текст] : / В.Н. Ярцева и др. — 2-е изд. — М.: Большая Российская энциклопедия, 1998. — с. 473.
83. Вольф, Е.М. Функциональная семантика оценки. [Текст] : / Е.М. Вольф; вступ. ст. Н.Д. Арутюновой и И.И. Челышевой. — Изд. 2-е изд., доп. — М.: URSS, 2005. — XXII, 259 с.

84. Герд А.С. Структура документа и словник терминологического словаря // Лексикология. Терминоведение. Стилистика. — М.-Рязань, 2003. — с. 71—73
85. Герд, А.С. Прикладная лингвистика. — СПб.; СПбГУ, 2005. — 268 с.
86. Говорунова, Л.Ю. Речевой жанр «Интернет-отзыв туриста» в русской и итальянской лингвокультурах. АКД. — Волгоград, 2014. — 26 с.
87. Головин, Б.Н., Кобрин Р.Ю. Лингвистические основы учения о терминах. — М.: Высшая школа, 1987. —104 с.
88. Горелов, И.Н. Невербальные компоненты коммуникации. [Текст] : / И.Н. Горелов; отв. ред. В. Н. Ярцева. — М.: Наука, 1980. — 104 с.
89. Городецкий, Б.Ю., Раскин, В.В. Методы семантического исследования ограниченного подъязыка. Публикации Отделения структурной и прикладной лингвистики. Вып. 5 «Автоматическая обработка речевой информации». Под общ.ред. В.А. Звегинцева. — М.: Изд-во Московского университета, 1971. — 414 с.
90. ГОСТ 34.003-90 «Информационная технология. Комплекс стандартов на автоматизированные системы. Автоматизированные системы. Термины и определения». — М.: Стандартинформ, 2009. — 15 с.
91. Дегтева, А.В. Автоматическое извлечение мнений (обзор). Рукопись. — СПб., 2006.
92. Дмитровская, М. А. Знание и мнение: образ мира, образ человека [Текст] : / М.А. Дмитровская // Логический анализ языка. Вып. 1. Знание и мнение: [Сб. ст.] / АН СССР, Ин-т языкознания; Отв. ред. Н. Д. Арутюнова. — М.: Наука, 1988. — С. 7—17.
93. Ермаков А.Е., Киселев С.Л. Лингвистическая модель для компьютерного анализа тональности публикаций СМИ // Компьютерная лингвистика и интеллектуальные технологии: труды Международной конференции Диалог’2005. — Москва, Наука, 2005. — [Электронный ресурс] http://www.rco.ru/?p=4439 (Дата: 10.04.2016)
94. Ефанова, Л.Г. Семантическая категория нормы в аспекте структурных составляющих нормативной оценки: монография. — Томск: Изд-во Томского государственного педагогического университета, 2012. — 220 с.
95. Ефанова, Л.Г. Функционально-семантическое поле нормы в русском языке: монография. — Томск: Изд-во Томского государственного педагогического университета, 2012. — 220 с.
96. Ефремова, Н.В. Диалогичность как один из механизмов текстообразования в процессе дискурсивной деятельности ученого // Известия ВГПУ. Филологические науки. №2. — С. 138—145.
97. Жирова, И.Г. Лингвистическая категория эмфатичность в антропоцентрическом аспекте: автореферат дис. на соиск. уч. степ. доктора филологических наук: 10.02.19. [Текст] : / И.Г. Жирова; [Место защиты: Моск. гос. обл. ун-т]. — М.: МГОУ, 2007. — 40 с.

98. Зализняк А. А. "Русское именное словоизменение" с приложением избранных работ по современному русскому языку и общему языкознанию. — М., 2002.
99. Зелко, В.М. Проблемы разработки лингвистического обеспечения системы китайско-русского информационного машинного перевода (Научно-экспериментальное исследование). АКД. Специальность: 10.02.19 «Теория языкознания». — М.: Институт языкознания, 1991. — 31 с.
100. Зубов, А.В., Зубова, И.И. Информационные технологии в лингвистике. — М.; Академия, 2004. — 208 с.
101. Ионова, С.А. Аппроксимация содержания вторичных текстов: автореф. дис. … д-ра. фил. наук. — Волгоград, 2006. — 37 с.
102. Исенбаева, Г.И. Методология порождения вторичного текста: когнитивный аспект: автореф. дис. … д-ра. фил. наук. — Уфа, 2010. — 37 с.
103. Кибрик, А.Е. Модель автоматического анализа письменного текста (на материале ограниченного военного подъязыка). — М.: МГУ, 1970. — 365 с.
104. Классы глаголов в функциональном аспекте: Сб. науч. тр. — Свердловск: УрГУ, 1986. — 160 с.
105. Клобуков, Е.В. Категория одушевленности/неодушевленности // Современный русский литературный язык: учебник для филологических специальностей пед. институтов / П.А. Лекант, Н.Г. Гольцова, В.П. Жуков и др.; Под ред. П. Леканта. — М.: Высш. шк., 1988. — С. 183—184.
106. Клышинский Э.С. Начальные этапы анализа текста // Автоматическая обработка текстов на естественном языке и компьютерная лингвистика / Большакова Е.И., Клышинский Э.С., Ландэ Д.В., Носков А.А., Пескова О.В., Ягунова Е.В. — М.: МИЭМ, 2011. [Электронный ресурс]: http://clschool.miem.edu.ru/uploads/swfupload/files/011a69a6f0c3a9c6291d6d375f12aa27e349cb67.pdf (Дата доступа: 10.07.2015)
107. Коваль, С.А. К построению классов лексем в компьютерной морфологии: имя прилагательное в русском языке // Структурная и прикладная лингвистика. Вып. 6: Межвуз. сб. / Под ред. А.С. Герда. — Спб.: изд-во С.-Петербургского ун-та, 2004. — с. 101—123.
108. Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 11 (18). Том 2. — М.: Изд-во РГГУ, 2012. — 137 с.
109. Котелова, Н.З. Значение слова и его сочетаемость. — Л.: Наука, 1975. — 164 с.
110. Куликов, С.Ю. Автоматизация составления оценочного словаря широкой предметной области [Текст] : (опыт использования неспециализированного корпуса текстов) / С. Ю. Куликов // Вестник Иркутского Государственного Технического Университета. — 2014. — № 8 (91). — С. 240—243.
111. Куликов, С.Ю. Автоматическое извлечение мнений как новая информационная технология // Теоретические и практические вопросы межкультурной коммуникации. Материалы научно-методической

конференции преподавателей, аспирантов и студентов МГОУ. Студенческие работы. М. 2010. С. 69—71.

112. Куликов, С.Ю. Использование моделей глагольного управления для задач автоматического извлечения мнений //Актуальные задачи лингвистики, лингводидактики и межкультурной коммуникации. Материалы 5-й Международной научно-практической конференции (20–21 сентября 2012): сборник научных трудов / отв. ред. Н.С. Шарафутдинова.— Ульяновск: УлГТУ, 2012.
113. Куликов, С.Ю. К вопросу о месте автоматического извлечения мнений в рамках компьютерной лингвистики // VI Международная научная конференция «Язык, культура, общество». Тезисы докладов. Том 1. — М., 2011. — С. 94.
114. Куликов, С.Ю. Контекстологическое определение оценки именной группы // XIV Международная научная конференция студентов-филологов. Тезисы. Часть 4. — СПб., 2011. — С. 26—27.
115. Куликов, С.Ю. Морфемный синтез как способ автоматического пополнения словаря оценочной лексики тезаурусного типа (на материале ксенофобонимов) / С. Ю. Куликов // Вопросы филологии. — 2015. — № 2 (50). — С. 91—98.
116. Куликов, С.Ю. Некоторые вопросы описания положительных оценок именных групп для задач автоматического извлечения мнений // Актуальные проблемы теоретической и прикладной лингвистики: сборник научных трудов. — Ульяновск, 2010. — С. 118—122.
117. Куликов, С.Ю. Об одном аспекте определения мнений на уровне объектов // Материалы Международного молодежного научного форума «ЛОМОНОСОВ-2013» / Отв. ред. А.И. Андреев, А.В. Андриянов, Е.А. Антипов, К.К. Андреев, М.В. Чистякова. [Электронный ресурс] — М.: МАКС Пресс, 2013. http://lomonosov-msu.ru/archive/Lomonosov_2013/2309/32293_8be6.doc
118. Куликов, С.Ю. Общение в Интернете и новые вызовы компьютерной лингвистике // Материалы Международного молодежного научного форума «ЛОМОНОСОВ-2012» / Отв. ред. А.И. Андреев, А.В. Андриянов, Е.А. Антипов, К.К. Андреев, М.В. Чистякова. [Электронный ресурс] — М.: МАКС Пресс, 2012. http://lomonosov-msu.ru/archive/Lomonosov_2012/1911/32293_74bc.doc
119. Куликов, С.Ю. Однореферентные оценочные слова и их формализация [Текст] (на примере этнофобонимов) / С. Ю. Куликов // Вопросы психолингвистики. — 2014. — № 4 (22). — С. 166—173.
120. Куликов, С.Ю. Определение автора высказывания при двойном цитировании // Проблемы языка: Сборник научных статей по материалам Второй конференции-школы «Проблемы языка: взгляд молодых ученых». — М.: Институт языкознания РАН, 2013. — С. 209—215.

121. Куликов, С.Ю. Оценочная шкала в системах автоматического извлечения мнений // XIII межвузовская научная конференция студентов-филологов. Тезисы. Часть 3. — СПб., 2010. — С. 35—36.
122. Куликов, С.Ю. Оценочные усилители в тексте // Перевод и когнитология в XXI веке. Сборник студенческих научных статей (по материалам конференции 13 апреля 2010 г.). — М., 2010. — С. 68—73.
123. Куликов, С.Ю. Проблема однородной выборки при отборе ресурсов для задач автоматического извлечения мнений // Прикладная лингвистика сегодня и завтра: актуальные проблемы: Материалы Межвузовского студенческого форума по прикладной лингвистике, 18 февраля 2011 г.: Вып.1. — Жуковский, 2011. — С. 47—49.
124. Куликов, С.Ю. Распознавание именованных сущностей как первый этап прагматического анализа текста // Русский язык: исторические судьбы и современность: V Международный конгресс исследователей русского языка (Москва, МГУ имени М. В. Ломоносова, филологический факультет, 18-21 марта 2014 г.): Труды и материалы / Составители: М. Л. Ремнёва, А. А. Поликарпов, О. В. Кукушкина. — М.: Изд-во Моск. ун-та, 2014. — С. 574.
125. Куликов, С.Ю. Способы выражения причин оценки в языке Интернета // Материалы Международного молодежного научного форума «ЛОМОНОСОВ-2014» / Отв. ред. А.И. Андреев, А.В. Андриянов, Е.А. Антипов. [Электронный ресурс] — М.: МАКС Пресс, 2014. http://lomonosov-msu.ru/archive/Lomonosov_2014/2713/2200_32293_2abfde.doc
126. Куликов, С.Ю. Способы выражения причин оценки в языке Интернета и их автоматическая идентификация // Актуальные проблемы филологической науки: взгляд нового поколения: Материалы XX–XXI Международных конференций студентов, аспирантов и молодых ученых «Ломоносов»: Секция «Филология» / Ред.-сост. А. Е. Беликов. — М.: Издательство Московского университета, 2015. — Выпуск 6. — 627—630.
127. Куликов, С.Ю. Теория подъязыков и автоматическое извлечение мнений // Сборник научных трудов, посвященный юбилею проф. Марчука Ю.Н. — М.: МГОУ, 2012. — С. 69—73.
128. Куликов, С.Ю. Типология оценочных шифтеров при автоматическом извлечении мнений //Проблемы современной лингвистики и методики преподавания иностранных языков: сб. тезисов науч.-практич. конф. для студентов / Под ред. И.Ю. Мигдаль. Коломна, 2010. — С. 78—81.
129. Куликов, С.Ю. Формальные свойства текста как фактор субъективности // Проблемы современной лингвистики и методики преподавания иностранных языков: сб. тезисов науч.-практич. конф. для студентов. — Коломна, 2011. — С. 99—102.
130. Кустова, Г.И., Ляшевская, О.Н., Падучева, Е.В., Рахилина, Е.В. Семантическая разметка лексики в Национальном корпусе русского языка: принципы, проблемы, перспективы // Национальный корпус русского языка: 2003-2005. Результаты и перспективы. — М., 2005. — с. 155—174.

131. Леонтьев, А.А., Сорокин, Ю.А., Базылев, В.Н., Бельчиков, Ю.А. Понятие чести и достоинства: психолингвистический анализ. — М. — Калуга: ИП Кошелев А.Б., 2009. — 72 с.
132. Леонтьева Н.Н. Автоматическое понимание текстов. — М.: «Академия», 2006. — 304 с.
133. Логический анализ языка. [Текст] : Вып. 1. Знание и мнение : [Сб. ст.] / АН СССР, Ин-т языкознания; Отв. ред. Н. Д. Арутюнова. — М.: Наука, 1988. — 125 с.
134. Логический анализ языка. Ментальные действия. Отв. ред. Н. Д. Арутюнова, Н.К. Рябцева. — М.: Наука, 1993. — 176 с.
135. Логический анализ языка. Язык речевых действий. Отв. ред. Н. Д. Арутюнова, Н.К. Рябцева. — М.: Наука, 1994. — 188 с.
136. Лукашевич, Н.В. Тезаурусы в задачах информационного поиска. — М., 2011.
137. Марчук, Ю.Н. Вычислительная лексикография. [Текст] : / Ю.Н. Марчук. — М.: ВЦП, 1976. — 184 с.
138. Марчук, Ю.Н. Компьютерная лингвистика. [Текст] : / Ю.Н. Марчук. — М.: «Восток-Запад», 2007. — 317 с.
139. Марчук, Ю.Н. Модели перевода. [Текст] : / Ю.Н. Марчук. — М.: «Академия», 2010. — 176 с.
140. Марчук, Ю.Н. Об алгоритмическом разрешении лексической многозначности (исследование автоматического контекстологического словаря). [Текст] : / Ю.Н. Марчук. [Место защиты МПИИЯ]. — АКД. — М.: МГПИИЯ, 1968. — 17 с.
141. Маслова, В.А. Параметры экспрессивности текста [Текст] : / В.А. Маслова. // Языковые механизмы экспрессивности (Отв. ред. В.Н. Телия). — М.: Наука, 1991. — С. 179—204.
142. Матвеева, Т.В. Тональность // Стилистический энциклопедический словарь русского языка. Под редакцией М.Н. Кожиной. — М:. «Флинта», «Наука», 2003. — [Электронный ресурс]: http://enc-dic.com/stylistic/Tonalnost-114.html (Дата доступа: 10.03.2016)
143. Мельчук, И.А. Автоматический синтаксический анализ. Том 1. Общие принципы. Внутрисегментный синтаксический анализ. Новосибирск, 1964.
144. Мельчук, И.А. Опыт теории лингвистических моделей "Смысл ⇔ Текст". — М., 1999.
145. Мельчук, И.А., Иорданская, Л.Н. Смысл и сочетаемость в словаре. — М.: Языки славянских культур, 2007. — 672 с.
146. Меньшиков, И.И. Определение семантических связей между компонентами сложного слова при автоматическом анализе немецких текстов // Проблемы изучения семантики языка. Часть I. — Днепропетровск, 1968. — С. 72—74.
147. Минина, М.А. Психолингвистический анализ семантики оценки (на материале глаголов движения). — АКД. М., 1995. — 22 с.

148. Минтусова, О.В. Моделирование морфологического компонента промышленных систем машинного перевода. АКД. Специальность: 10.02.19 «Теория языкознания». — М.: Институт языкознания, 1990. — 19 с.
149. Мустафина, А.Р. Синтаксическое сжатие: структура, семантика, иллокуция (на материале английского языка). АКД. Уфа, 2013. — 24 с.
150. Нелюбин, Л.Л. Компьютерная лингвистика и машинный перевод. — М.: ВЦП, 1991. — 151 с.
151. Новиков А. И. Текст и его смысловые доминанты / Под ред. Н. В. Васильевой, Н. М. Нестеровой, Н. П. Пешковой.— М.: Институт языкознания РАН, 2007. — 224 с.
152. Огарок, А. Стокона на РОМИП-2006 // Труды РОМИП'2006, Санкт-Петербург: НУ ЦСИ, 2006, с. 86—91.
153. Падучева, Е.В. Соответствие «смысл ⇔ текст» в исторической перспективе//Восток — Запад: Вторая международная конференция по модели «Смысл ⇔ Текст» (Отв. ред. Ю. Д. Апресян, Л. Л. Иомдин). — М, 2005. — С. 330—349
154. Пащенко, Н.А., Кнорина, Л.В., Молчанова, Т.В., Чепиго, Т.С., Шумилина, А.Л., Яровенко, О.И. Проблемы автоматизации индексирования и реферирования // Итоги науки и техники. Серия «Информатика». Т.7 Автоматизация индексирования и реферирования документов. — М.: ВИНИТИ, 1983. — с. 7—164.
155. Плунгян, В.А. Корпус как инструмент и как идеология: о некоторых уроках современной корпусной лингвистики// Русский язык в научном освещении, №2 (16), 2008. — С. 7—20.
156. Поляков, П.Ю., Калинина, М.В., Плешко, В.В. Исследование применимости методов тематической классификации в задаче классификации отзывов о книгах // Компьютерная лингвистика и интеллектуальные технологии: По материалам ежегодной Международной конференции «Диалог». Вып. 11 (18). Том 2. — М.: РГГУ, 2012. — С. 51—59.
157. Потапова, Р.К., Комалова, Л.Р. База данных русскоязычных текстов, содержащая единицы семантического поля «агрессия» // Вестник Московского государственного лингвистического университета, 2014. Серия Языкознание, том 19, № 705 «Семиотическая гетерогенность мужкультурной коммуникации». Ч. 1, с. 112-121.
158. Рябцева, Н.К. Информационные процессы и машинный перевод. [Текст] : / Н.К. Рябцева, отв. ред. Р.Г. Котов. — М.: Наука, 1986. — 168 с.
159. Словарь современного русского литературного языка. Том 7 «Н». Ред. Ф.П. Сороколетов, Ф.П. Филин. — М.-Л.: Изд-во АН СССР, 1958. — 1470 с.
160. Соболь, Е.Ю. Невербальные компоненты текстовой информации (на материале английской художественной литературы). [Текст] : / Е.Ю. Соболь, Монография. — М.: Издательство МГОУ, 2009. — 233 с.
161. Сунцова, Н.Л. Лингвистическая модель порождения вторичного текста: автореф. дис. … канд. фил. наук. — М., 1995. — 22 с.

162. Суперанская, А.В., Подольская, Н.В., Васильева, Н.В. Общая терминология: Вопросы теории. — М.: Либроком, 2012. — 248 с.
163. Сутормина, А.Ю. Выявление скрытой оценки. Магистерская диссертация. — М.: ВШЭ, 2014. — 52 с.[238]
164. Телия, В.Н. Предисловие [Текст] : / В.Н. Телия //Языковые механизмы экспрессивности. (Отв. ред. В.Н. Телия). — М.: Наука, 1991. — С. 3—4.
165. Телия, В.Н. Русская фразеология. Семантический, прагматический и лингвокультурологический аспекты. [Текст]: / В.Н. Телия. — М.: Школа «Языки русской культуры», 1996. – 288 с.
166. ТКС [Толково-комбинаторный словарь русского языка]. [Текст]:/ И.А. Мельчук (ред.) и др. — Вена: Wiener Slawische Almanach, 1984. — 992 с.
167. Толкунов, А.А. Модель оперативной аналитической обработки текстовых комментариев к законопроектам. АКД. — Орел: Академия ФСО, 2014. — 24 с.
168. Тулдава, Ю.А. Проблемы и методы квантитативно-системного исследования лексики. [Текст] : / Ю.А. Тулдава. — Тарту: ТГУ, 1987. — 203 с.
169. Убин, И.И. Лексические средства усиления и ослабления слов в тексте [Текст] : / И.И. Убин //Вычислительная лингвистика (Отв. ред. Е.К. Гусева). — М.: Наука, 1976. — С. 159—167.
170. Успенский, Б.А. Поэтика композиции. — М.: «Искусство», 1970. — 223 с.
171. Фенько, А.Б. Справочник социопата. ИД «Коммерсант», 2001. [Электронный ресурс: http://www.compromat.ru/page_11503.htm]
172. Хайруллин, В.И. Корпусный словарь фобиотерминологии // Научно-техническая информация. Сер. 2. — 2013. — № 4. — С. 30—32.
173. Чернышов, М.Ю. Вербальные средства выражения мыслей-скрепов как медиаторов смысловой интеграции текста. [Текст]:/ М.Ю. Чернышов. — М.: Издательство «Наука и право», 2010. — 168 с.
174. Четверкин, И.И. Автоматизированное формирование базы знаний для задачи анализа мнений: автореф. дис. … канд. физ.-мат. наук. — М., 2013. — 20 с.
175. Четверкин, И.И., Лукашевич, Н.В. Автоматическое извлечение оценочных слов для конкретной предметной области [Текст]:/ И.И. Четвёркин, Н.В. Лукашевич //Труды международной конференции «Диалог-2010». — М.: РГГУ, 2010. — С. 565—571.
176. Чиркин, Е.С. Некоторые проблемы автоматизированного извлечения данных из веб-страниц // Материалы научной конференции «Интернет и современное общество». — СПб., 2013. — С. 291—294.
177. Шпет, Г.Г. Внутренняя форма слова: Этюды и вариации на темы Гумбольта. [Текст] : / Г.Г. Шпет. — Изд. 3-е, стереотипное. — М.: КомКнига, 2006. — 216 с.

[238] Электронные приложения (словарь оценочной лексики и размеченный корпус): https://github.com/AlyssaSut/assessment_words

178. Ярцева, В.Н. (ред.) Языкознание. Большой энциклопедический словарь. [Текст] : / В.Н. Ярцева и др. — 2-е изд. — М.: Большая Российская энциклопедия, 1998. — 685 с.

Список иллюстративного материала

1. http://irek-murtazin.livejournal.com/1297976.html?thread=54227768
2. http://taki-net.livejournal.com/1283868.html
3. http://valery-kichin.livejournal.com/258940.html?thread=2769788&
4. http://www.otzyv.ru/hotel/turkey/Club_Hotel_Sera/
5. http://khabar.kz/ru/news/ekonomika/item/24696-v-kazakhstane-podorozhal-benzin
6. http://mikle1.livejournal.com/3385051.html
7. http://vk.com/wall-40684908_255150
8. http://www.kommersant.ru/doc/2799888
9. http://www.newsru.com/world/30sep2015/kerrysays.html
10. http://www.bbc.com/russian/russia/2015/08/150818_rus_press
11. http://www.utro.ru/articles/2015/08/05/1251575.shtml
12. http://www.novayagazeta.ru/politics/65753.html
13. http://echo.msk.ru/programs/korzun/1419892-echo/
14. http://www.omskinform.ru/news/88205
15. http://ria.ru/culture/20141018/1028968755.html#ixzz3GbppmAEj
16. http://books.imhonet.ru/element/1159123/
17. http://democrator.ru/complain/5726/5726/?page=2
18. http://www.odnako.org/blogs/chitat-3/
19. http://husainov.livejournal.com/292460.html?thread=2444652#t2444652
20. http://litrossia.ru/archive/112/writer/2632.php
21. http://www.otzyv.ru/read.php?id=188224
22. http://www.banki.ru/services/responses/bank/response/8321342/
23. http://www.mk.ru/economics/2016/01/17/importozameshhenie-luch-nadezhdy-ili-doroga-v-propast.html
24. http://expert.ru/2015/07/15/horoshij-primer-dlya-tsb/
25. http://www.metronews.ru/kolumnisty/horoshij-plohoj-effektivnyj-jozef-blatter/Tpoogu---2M4DdY4OKMEDI/
26. http://ru.tsn.ua/politika/klyuchevye-tezisy-brifinga-kolomoyskogo-horoshiy-muzhik-v-ukrnafte-i-sms-perepiska-s-ahmetovym-456620.html
27. http://www.poisknews.ru/theme/edu/15256/
28. http://lady.tut.by/news/food/457582.html
29. http://www.championat.com/football/article-226401-otstavka-fabio-kapello--v-obzore-zarubezhnykh-smi.html
30. http://vesti-ukr.com/strana/108161-ukraincev-vystrojat-v-polutoramesjachnuju-ochered-za-shengenom
31. http://www.inopressa.ru/article/20jul2015/times/miller.html
32. http://www.championat.com/football/news-2194473-timoshhuk-u-kajrata-ochen-khoroshij-podbor-igrokov.html
33. http://vesti-ukr.com/odessa/108129-chto-tormozit-razvitie-odessy-kak-turisticheskoj-stolicy
34. http://www.nakanune.ru/news/2015/7/16/22407839

35. http://www.pravda.com.ua/rus/news/2015/07/28/7075946/
36. http://kazanfirst.ru/feed/50644
37. http://news.finance.ua/ru/news/-/354953/yatsenyuk-skazal-kogda-v-dnepropetrovske-poyavitsya-politsiya-zarplatu-obeshhaet-takuyu-zhe-kak-i-v-kieve
38. http://www.tvc.ru/news/show/id/73340
39. http://www.gazeta.ru/business/news/2015/07/24/n_7404617.shtml
40. http://www.unian.net/politics/1102054-boris-filatov-snachala-lyudi-vyibirayut-verhovnuyu-radu-a-potom-govoryat-chto-j-za-idiotov-myi-tam-sobrali.html
41. http://www.inopressa.ru/article/27Jul2015/welt/rus_gold.html
42. http://nv.ua/world/countries/gensek-nato-podytozhil-rezultaty-ekstrennogo-zasedanija-po-turtsii-61344.html
43. http://riafan.ru/354951-obama-ya-horoshiy-prezident-no-ya-ustal
44. http://forumkavkaz.com/index.php?topic=53.1770
45. http://ria.ru/arab_riot/20130906/961375679.html
46. http://news2world.net/politicheskie-novosti/novosibirsk-sekonomit-na-viborah-v-odin-tur-bolshe-chem-ozhidalos.html
47. http://mr7.ru/articles/81301/
48. http://www.km.ru/v-rossii/2013/05/29/ekonomika-i-finansy/711933-press-sekretar-putina-otkazalsya-svyazyvat-ukhod-guri
49. http://webground.su/rubric/2010/07/22/nauka_fizika/retro/
50. http://obozrevatel.com/abroad/98567-patologicheskaya-tupost-belarusskij-politik-rasskazal-za-chto-putin-nenavidit-lukashenko.htm
51. http://www.mv.org.ua/news/109186-izvestnyi_rossiiskii_akter_priznal_chto_rossija_-_strana_tretego_mira.html
52. http://irubtsov.ru/2012/08/19/al-fa-bank-nenavidit-nadyozhny-h-klientov/
53. http://ncrim.ru/articles/view/25-09-2015-na-ulicah-sevastopolya-opasno-perehodit-dorogu
54. http://www.news-asia.ru/view/8806
55. http://www.gezitter.org/vybory/44101_kak_proshli_parlamentskie_vyiboryi/
56. http://mir-politika.ru/tags/%D0%BC%D0%BE%D0%B4%D0%B6%D0%B0%D1%85%D0%B5%D0%B4%D0%B4%D0%B8%D0%BD%D1%8B/
57. http://www.saper.etel.ru/history/musulman-vs.html
58. http://mreadz.com/new/index.php?id=346373&pages=98
59. http://pereformat.ru/2014/01/dna-genealogy-jews/
60. http://alexander-apel.narod.ru/library/dohodyaga/part19.htm
61. http://pirochem.net/index.php?id1=3&category=voennaya-istoriya&author=kiselev-vi&book=diviziiss2002&page=12
62. https://otvet.mail.ru/question/169019485
63. http://newsland.com/news/detail/id/576565/
64. http://samlib.ru/i/iwanowmiljuhin_j_z/zzz-1.shtml
65. http://old2.booknik.ru/colonnade/letters/lichnyyi-jeleznyyi-zanaves/

66. http://www.gday.ru/forum/%CE%E1%F9%E8%E9-%F4%EE%F0%F3%EC/217638-%E8%F1%F2%EE%F0%E8%FF-%EB%FE%E1%E2%E8-3.html
67. http://www.studfiles.ru/preview/1742725/page:25/
68. http://www.nn.ru/community/velo/main/uskorit_velosiped_staroy_tetki.html
69. http://riatribuna.ru/news/2010/12/23/3688/
70. http://m.echo.msk.ru/blogs/detail.php?ID=931642
71. http://www.championat.com/football/news-1935884-radimov-zenit---javnyj-favorit-matcha-so-spartakom.html
72. http://www.dni.ru/polit/2014/9/19/281226.html
73. http://cheger.livejournal.com/478311.html
74. http://awas1952.livejournal.com/3828452.html
75. http://gorod1277.org/?q=content/ne-ponravilos
76. http://forum.allods.ru/showthread.php?t=119664
77. http://beseder.ru/news/entryid/436
78. http://football.ua/countrieselse/141135-gruppa-a-turcyja-kak-turan-na-novye-vorota.html
79. http://merelana.livejournal.com/754391.html
80. https://illustrators.ru/illustrations/766397
81. http://n-europe.eu/content/?p=2814
82. http://www.gazeta.ru/politics/2012/02/29_a_4016413.shtml
83. http://www.vokrug.tv/article/show/Dima_Bilan_zashchitil_svoi_mikroblog_ot_moshennikov_49850/
84. http://www.kp.ru/daily/26441/3312783/
85. http://www.infox.ru/auto/car/2015/10/02/Na_dorogah_YAponii_p.phtml
86. http://f1report.ru/news/f1-16251.html
87. http://www.e1.ru/news/spool/news_id-407456.html
88. http://www.kuban.kp.ru/daily/25692.3/895111/
89. http://4pera.ru/news/feysbuchnye_truth/pochemu_krym_russkiy_poluostrov/
90. http://gigafootball.net/news/89141
91. http://www.nakanune.ru/articles/110929
92. http://www.segodnia.ru/news/167271
93. http://bankir.ru/novosti/20151006/the-guardian-bankovskaya-gruppa-anz-vlozhit-10-mlrd-v-chistuyu-energetiku-10112620/
94. https://www.banki.ru/services/official/bank/response/523933/
95. http://nearch78.livejournal.com/8178.html
96. http://www.komarovskiy.net/forum/viewtopic.php?t=22630&postdays=0&postorder=asc&start= 675&sid=409b049e00561699b920fde7066d5139
97. http://svpressa.ru/politic/article/63081/
98. http://sturmnovosti.co/view.php?COLLCC=1448022774&id=38688
99. http://tomsk.fm/watch/47431?from=157652
100. http://www.newsru.com/religy/22sep2014/gainutdin_spisok.html
101. http://democracy.ru/article.php?id=1081
102. http://www.newsru.com/world/18sep2014/kazakhstan.html

103. http://www.atheism.ru/library/efimov_1.phtml
104. http://otvet.mail.ru/question/18545290
105. http://berkovich-zametki.com/Guestbook/guestbook_okt2012_2.html
106. https://www.facebook.com/764151920310984?comment_id=764812533578256
107. http://azh.kz/ru/news/view/19514
108. http://www.newsru.com/russia/22aug2014/bmd.html
109. http://actualcomment.ru/daycomment/1494/
110. http://www.grozny-inform.ru/main.mhtml?Part=11&PubID=54515
111. http://www.newsru.com/religy/17sep2014/prokurorskaya_proverka.html
112. https://www.facebook.com/770200476372795?comment_id=770228089703367
113. http://ria.ru/world/20140923/1025339630.html
114. http://vk.com/id2640759
115. http://www.newsru.com/religy/21aug2014/morano.html
116. http://stockinfocus.ru/2014/11/14/rasshifrovka-vystuplenij-na-sovbez-oon/
117. http://www.islamnews.ru/news-440705.html
118. http://islam-today.ru/novosti/2014/11/12/eks-kandidat-v-prezidenty-ssa-sravnil-musulman-s-bolnymi-rakom/
119. http://regions.ru/news/2488577/
120. http://gareeb.livejournal.com/45804.html
121. http://dpmmax.livejournal.com/186656.html?thread=39568416#t39568416
122. http://pn14.info/?p=158687
123. http://www.zarusskiy.org/initiatives/2008/12/09/oshibka/
124. http://observer-lj.livejournal.com/1139969.html
125. http://mnenia.zahav.ru/Articles/702/krov_pasternaka
126. http://www.mignews.com/news/society/world/190614_121926_49030.html
127. http://9tv.co.il/news/2014/06/18/178287.html
128. http://thekievtimes.ua/society/374806-konchita-pobeda-nad-evropoj.html
129. http://armtoday.info/default.asp?Lang=_Ru&NewsID=117432
130. http://poleev.livejournal.com/583.html

Приложение 1. Список мотивационных целей

Полный список целей (разработан Р. Шенком):

1) P-health (Self, +)
2) P-health (Family, +)
3) A-possessions (Self, +)
4) P-possessions (Self, +)
5) A-scont (Self, others, +)
6) A-know (Self, x, +)
7) E-unpleasant activity (Self, -)
8) E-pleasant activity (Self, +)
9) P-health (Others, +)
10) P-anything (enemies, -)

Пример описания стереотипов поведения (стрелка вверх — увеличение номера относительно базового списка, стрелка вниз — уменьшение номера относительно базового списка):

Ambitious (1) A-possessions (Self, +) ↑
A-scont (Self, others, +) ↑
P-anything (Others, +) ↓
A-know (Self, +) ↓ (слабое)

Powerhungry (6) A-scont (Self, others, +) ↑

Приложение 2. Фрагмент оценочного словаря прилагательных

При расширении исходного списка прилагательных были использованы следующие процедуры:

- расширение списка при помощи префиксов;
- расширение списка при помощи суффиксов.

Процедуры выполняются последовательно.

Первичный словарь оценочной лексики доступен по веб-адресу: https://sourceforge.net/projects/wnlp-project/files/Simple%20adjective%20list/.

Приложение 3. Структура оценочной разметки корпусов текстов относительно объектов оценки

Формат разметки А.Ю. Сутормина[239]

```
<document>
  <header>
    <textfile>assessment_corpora_2/lj__steppenwolf_1353448.txt</textfile>
    <lang>english</lang>
    <complete>100%</complete>
  </header>
    <segments>
      <segment id='1' start='437' end='439' features='types;type;object'
state='active'/>
      <segment id='2' start='440' end='450' features='types;type;est_verbs;verb_neg'
state='active'/>
      <segment id='3' start='484' end='493' features='types;type;est_verbs;verb_neg'
state='active'/>
    </segments>
</document>
```

[239] Текст взят из архива: https://github.com/AlyssaSut/assessment_words/blob/master/%D0%BE%D1%86%D0%B5%D0%BD%D0%BA%D0%B0_%D0%BA%D0%BE%D1%80%D0%BF%D1%83%D1%81.rar

Формат разметки SentiRuEval-2015[240]

<review id="816831">

<meta>

<object>Mazda 6 седан</object>

</meta>

<text>В принципе машинка не плохая, объемом в 2.0 куба, легкий кузов, дорогу держит не плохо, приятна по салону, сделана в спорт стиле, по городу 9 литров не больше, по трассе 7, но есть и минусы что по ходовой слабенькая машина и кузова почти у всех гниют со временем, многие водители этого не замечают, не говорю что все 100% гниют. Общее впечатление : Слабая по ходовке,цветет кузов.</text>

<aspects>

<aspect category="Whole" from="11" mark="Rel" sentiment="positive" term="машинка" to="18" type="explicit"/>

<aspect category="Driveability" from="57" mark="Rel" sentiment="positive" term="кузов" to="62" type="explicit"/>

<aspect category="Driveability" from="64" mark="Rel" sentiment="positive" term="дорогу держит" to="77" type="explicit"/>

<aspect category="Appearance" from="99" mark="Rel" sentiment="positive" term="салону" to="105" type="explicit"/>

<aspect category="Appearance" from="115" mark="Rel" sentiment="neutral" term="в спорт стиле" to="128" type="explicit"/>

<aspect category="Driveability" from="130" mark="Rel" sentiment="neutral" term="по городу" to="139" type="explicit"/>

<aspect category="Driveability" from="160" mark="Rel" sentiment="neutral" term="по трассе" to="169" type="explicit"/>

<aspect category="Driveability" from="197" mark="Rel" sentiment="negative" term="ходовой" to="204" type="explicit"/>

[240] Текст взят с сайта соревнования: https://drive.google.com/folderview?id=0B7y8Oyhu03y_fjNIeEo3UFZObTVDQXBrSkNxOVlPaVAxNTJPR1Rpd2U1WEktUVNkcjd3Wms&usp=drive_web

<aspect category="Driveability" from="216" mark="Rel" sentiment="negative" term="машина" to="222" type="explicit"/>

<aspect category="Reliability" from="225" mark="Rel" sentiment="negative" term="кузова" to="231" type="explicit"/>

<aspect category="Reliability" from="245" mark="Rel" sentiment="negative" term="гниют" to="250" type="fct"/>

<aspect category="Reliability" from="322" mark="Irr" sentiment="negative" term="гниют" to="327" type="fct"/>

<aspect category="Driveability" from="359" mark="Rel" sentiment="negative" term="ходовке" to="366" type="explicit"/>

<aspect category="Reliability" from="367" mark="Rel" sentiment="negative" term="цветет" to="373" type="fct"/>

<aspect category="Reliability" from="374" mark="Rel" sentiment="negative" term="кузов" to="379" type="explicit"/>

</aspects>

<categories>

<category name="Comfort" sentiment="absence"/>

<category name="Appearance" sentiment="positive"/>

<category name="Reliability" sentiment="negative"/>

<category name="Safety" sentiment="absence"/>

<category name="Driveability" sentiment="both"/>

<category name="Whole" sentiment="both"/>

<category name="Costs" sentiment="absence"/>

</categories>

</review>

Формат разметки Б.Лю

<sentence id="6">

<text>The speakers have a very rich sound and good bass also , obviously not thumping bass for which you would need a huge subwoofer !!</text>

<aspectTerms>

<aspectTerm from="4" to="11" polarity="positive" term="speakers" pos="nn"/>

<aspectTerm from="30" to="34" polarity="positive" term="sound" pos="nn"/>

<aspectTerm from="45" to="49" polarity="positive" term="bass" pos="nn"/>

</aspectTerms>

</sentence>

Предлагаемый формат разметки[241]

<annotation>

<text>«Наша дизайн-концепция заключается в создании светлого и открытого пространства, объединяющего различные программы для городского сообщества. Это то, как с помощью архитектуры можно организовать жизнь сообщества», - цитирует главного архитектора Serie Architects Кристофера Ли Designboom.</text>

<opinion id=”1”>

<object from="1" to="21" term="Наша дизайн-концепция" sentiment="positive" config_type=”common”/>

<subject from=”225” to=”278” term=”главный архитектор Serie Architects Кристофер Ли”/>

<source from=”279” to=”288” term=”Designboom”/>

</opinion>

<opinion id=”2”>

<object from=”46” to”79” term=”светлое и открытое пространство” sentiment=”positive” config_type=”common”/>

<subject from=”225” to=”278” term=”главный архитектор Serie Architects Кристофер Ли”/>

<source from=”279” to=”288” term=”Designboom”/>

[241] Текст взят из статьи http://www.radidomapro.ru/ryedktzij/arkhitektura/design/oazis-restoranov-i-bolgnitz-slajd-schou-16974.php

</opinion>

<opinion id=”3”>

<object from=”201” to”210” term=”сообщество” sentiment=”positive” config_type=”common”/>

<subject from=”225” to=”278” term=”главный архитектор Serie Architects Кристофер Ли”/>

<source from=”279” to=”288” term=”Designboom”/>

</opinion>

</annotation>

Приложение 4. Фрагмент базы данных[242]

<ONTO-CLASS VALUE=”UNLAWFUL BEHAVIOUR MARKERS”>

<ONTO-SUBCLASS VALUE=”ENTHNIC PHOBIA”>

<ENTITY-CLASS ID=”1” VALUE=”RUSSIA”>

<ENTITY ID=”1” TYPE=”MAIN” POS=”NOUN” VALUE=”Россия” />

<ENTITY ID=”2” TYPE=”GENERIC” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”русофобия” />

<ENTITY ID=”3” TYPE=”SYN” SUBTYPE=”GENERIC” SYNID=”2” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”русофобство” />

<ENTITY ID=”4” TYPE=”AGENT” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”русофоб” />

<ENTITY ID=”5” TYPE=”MODIFIER” POS=”ADJECTIVE” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”русофобский” />

<ENTITY ID=”6” TYPE=”MULT” SUBTYPE=”MODIFIER,ACTION” POS=”VERB” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”русофобствовать” />

<ENTITY ID=”7” TYPE=”GENERIC” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”EXTREME” VALUE=”русоненавистничество” />

[242] Одним из источников рассматриваемой онтологии является «Справочник психопата» Анны Фенько [171].

<ENTITY ID=”8” TYPE=”AGENT” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”EXTREME” VALUE=”русоненавистник” />
<ENTITY ID=”9” TYPE=”MODIFIER” POS=”ADJECTIVE” SENTIMENT=”NEGATIVE” INT=”EXTREME” VALUE=”русоненавистнический” />
<ENTITY ID=”10” TYPE=”GENERIC” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”LOW” VALUE=”нерусофильство” />
<ENTITY ID=”11” TYPE=”AGENT” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”LOW” VALUE=”нерусофил” />
<ENTITY ID=”12” TYPE=”MODIFIER” POS=”ADJECTIVE” SENTIMENT=”NEGATIVE” INT=”LOW” VALUE=”нерусофильский” />
<ENTITY ID=”13” TYPE=”SYN” SUBTYPE=”GENERIC” SYNID=”2” POS=”NOUN” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”россиефобия” />
<ENTITY ID=”14” TYPE=”MODIFIER” POS=”ADJECTIVE” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”россиефобский” />
<ENTITY ID=”15” TYPE=”MODIFIER” POS=”ADJECTIVE” SENTIMENT=”NEGATIVE” INT=”HIGH” VALUE=”антироссийский” />
<ENTITY ID=”16” TYPE=”SYN” SUBTYPE=”MAIN” SYNID=”1” POS=”NOUN” VALUE=”русский” />
</ENTITY-CLASS>
</ONTO-SUBCLASS>
</ONTO-CLASS>