

Создание инструмента автоматического анализа текста в интересах социо-гуманитарных исследований.

Часть 1. Методические и методологические аспекты*

С. Н. Ениколопов^I, Ю.М. Кузнецова^{II}, И.В. Смирнов^{II}, М.А. Станкевич^{II}, Н.В. Чудова^{II}

^I Научный центр психического здоровья» РАН, Москва, Россия

^{II} Институт системного анализа Федерального исследовательского центра «Информатика и управление» РАН, Москва, Россия

Аннотация. На концептуальном уровне рассматриваются подходы к использованию средств искусственного интеллекта в социо-гуманитарных исследованиях. Мировой тенденцией последних десятилетий является вовлечение в исследовательскую практику данных, извлекаемых из социальных медиа, что позволяет решать задачи мониторинга, анализа, прогноза и управления по отношению к таким параметрам сетевой коммуникации, как мнения, установки, эмоции и поведенческие схемы коммуникантов. В целях развития средств анализа сетевого контента предлагается использование метода реляционно-ситуационного анализа, опирающегося на синтаксемный анализ Г.А. Золотовой и концепцию неоднородных семантических сетей Г.С. Осипова. Предлагаемый подход позволяет реализовать при анализе текста деятельностную парадигму, в рамках которой текст рассматривается как продукт речемыслительной деятельности автора, что открывает возможности для применения аппарата, выработанного отечественной психологией для изучения когнитивных процессов, личности, мотивации, эмоциональных состояний, индивидуальных различий, а также базового ментального образования – картины мира.

Ключевые слова: социо-гуманитарные исследования, искусственный интеллект, автоматический анализ текста, реляционно-ситуационный анализ, предикатно-синтаксемная структура.

DOI 10.14357/20718594190203

Введение

Одной из традиционных задач искусственного интеллекта (ИИ) является моделирование поведения человека и его психической активности – рассуждений, принятия решений, обучения, общения и проч. Центральным звеном, опосредствующим такую активность, является оперирование естественным языком. В связи с этим работы в области интеллектуального анализа текста в той мере, в какой они опираются на представления лингвистики о речевой деятельности, представляют собой не только практический, но и теоретический интерес для ИИ.

С другой стороны, и практический результат таких работ (в виде программного продукта) может быть интересен для проведения фундаментальных исследований – уже не в ИИ, а в различных социо-гуманитарных дисциплинах, в которых основным материалом для анализа выступают именно тексты испытуемых (у психологов), респондентов (у социологов и этнографов), исторических деятелей и простых обывателей (у историков), политиков (у политологов) и т.д.

В ИСА РАН уже второе десятилетие на основе реляционно-ситуационного анализа (РСА) разрабатываются средства для поиска и сравнения документов, а недавно была поставлена задача создания на основе РСА инструмента анализа

*Работа выполнена при финансовой поддержке РФФИ (№17-29-02247).

✉ Смирнов Иван Валентинович. E-mail: ivs@isa.ru

текстов для социо-гуманитарных исследований. Для ИИ эта задача является прикладной, однако результаты исследований гуманитариев, полученные с помощью созданного инструмента, могут быть впоследствии интегрированы в модели порождения и понимания текста.

Цель данной статьи – дать первое представление о новом методе, его технических характеристиках и особенностях получаемых с его помощью результатов.

1. Компьютерные технологии в социо-гуманитарных исследованиях

Одной из ключевых тенденций развития социо-гуманитарного сектора современной науки является активное применение компьютерных технологий на всех этапах исследовательской работы, начиная с выделения предмета и планирования исследования и заканчивая представлением полученных результатов: «По завершении долгих споров об отношениях между гуманитарными науками и цифровыми технологиями «цифра» отвоевала свое место в праве создания новой отрасли гуманитаристики - Digital Humanities, которая сформировалась в США и Британии еще в середине XX века. Это явилось ярким подтверждением более широкого процесса движения к интегративности научного знания. Внешние факторы подобной интегративности связаны с техническим использованием цифровых технологий в научном творчестве и культурном пространстве – это и создание поисковых систем, цифровых библиотек, музейных коллекций, профессиональных Интернет-сообществ, свободным доступом к информации, обмену данными, проведение медицинских исследований. Важным моментом здесь является управление процессами образования. Внутренние (глубинные) факторы связаны с непосредственным использованием цифровых технологий как комплекса прикладных методологических разработок применительно к гуманитарным наукам. Цифровая составляющая распространяется практически на все виды исследований культуры и связана с культурными практиками, в которых используется язык Soft Media: компьютерный дизайн, редактирование, создание баз данных (текстовых, библиотечных, онлайн энциклопедий, средств коммуникации в виде чатов, мессенджеров, социальных сетей и др.), моделирование информации, цифровых со-

циологических исследований, цифровых исследований этносов и народов» [1].

Важным направлением применения компьютерных технологий в интересах гуманитарных исследований является анализ информации, содержащейся в *сетевых средствах массовой коммуникации*. С помощью методов статистики и машинного обучения на больших массивах данных сетевого контента выявляются демографические характеристики пользователей, исследуются процессы группирования и лидерства в сети, значимые темы обсуждения, локализация и динамика таких обсуждений, специфические формы онлайн-общения, признаки вбросов недостоверной (фейковой) информации и т.п. Особое внимание уделяется психологическим состояниям участников сетевого общения, к которым относят их эмоции, мнения, убеждения, связанные с различными сферами общественной и частной жизни (политика, потребление, медицина, межличностные отношения и т.п.).

Согласно обзору [2], аналитическая работа с социальными сетями может иметь как пассивный (мониторинг с целью достижения понимания происходящих в сетях процессов), так и активный (прогнозирование и управление с целью перевода социальной сети в требуемое состояние) характер. При этом *мониторинг* подразумевает сбор текстов, выявление связей между пользователями и ссылок на внешние ресурсы; *анализ* – вычисление количественных показателей (например, количество отправленных сообщений пользователя), которые затем обрабатываются с целью выявления статистических и структурных закономерностей; *прогноз* – идентификацию математической модели информационного процесса с использованием статистических моделей и моделей динамических процессов на графах; *управление* – предоставление рекомендаций, основанных на формальных процессах.

Сами информационные системы описываются по следующим параметрам:

- уровень анализа социальных сетей – мониторинг, анализ, прогнозирование, управление;
- модели социальных сетей – модели структуры (случайных графов, безмасштабных сетей), модели распространения информации (марковские модели, конечные автоматы, модели диффузии инноваций, модели заражения и др.);
- методы анализа данных – статистические, анализа графов, семантического анализа, сантмент-анализа;

- объекты анализа – сеть в целом (при помощи агрегированных показателей), подсети и сообщества, отдельно взятые пользователи (при помощи актор-ориентированных локальных показателей), сообщения (при помощи показателей упоминания соответствующих информационных объектов), мнений (при помощи показателей тональности сообщения относительно информационных объектов), внешних узлов или информационных ресурсов сети;

- режимы анализа данных – возможность ретроспективного анализа данных и / или анализа данных в режиме реального времени;

- режимы сбора данных – возможность ретроспективного сбора данных и / или сбора данных в режиме реального времени; всего объема данных или по определенной тематике;

- охват источников данных – классические онлайн-социальных сетей (Facebook, Vkontakte), блогов (LiveJournal), микроблогов (Twitter), сервисов обмена фотографиями и видео (YouTube, Flickr), форумов и т.п.;

- объемы обрабатываемых данных – модельные или промышленные (программные продукты BigData).

Реально существующие системы различаются тем, какие характеристики они сочетают в себе. Некоторые из перечисленных параметров, по оценке авторов, являются не реализованными, например, возможность управления сетями [2]. Однако в более поздних работах имеются сообщения о реализации данной функции, например, в [3] описана система, позволяющая контролировать уровень агрессивности сетевого общения.

Специфическими особенностями методов искусственного интеллекта, выделяющими их на фоне компьютерных технологий в целом, по определению Г.С. Осипова, является: *экспериментальность* или подбор при отсутствии точного метода решения проблемы модели, поведение которой оказывается оптимальным, а также ориентация на мышление эксперта, что подразумевает учет данных когнитивных исследований; *символьная форма представления информации*, т.е. ориентация на вербально опосредованные способы решения задач человеком [4]. Применение аппарата искусственного интеллекта позволяет разрабатывать интеллектуальные системы в гуманитарных науках, расширяющие исследовательские возможности благодаря использованию специфических средств программирования и моделирования,

создания баз данных, алгоритмов извлечения информации и т.д. Так, в России имеются разработки в области применения методов искусственного интеллекта в социологии [5], политологии [6, 7], науковедении [8, 9]. Благодаря интересу к интеллектуальным системам со стороны отечественной лингвистики разработаны системы автоматического реферирования, автоматической проверки плагиата, генерации и выявления «машинных» текстов, машинного перевода и т.п. [10-12.]. При этом относительно слабо разработанным остается направление, объединяющее интересы психологии, социологии, политологии, культурологии, социальной педагогики, социальной психиатрии и других социо-гуманитарных дисциплин, и связанное с извлечением из сетевого контента информации об *индивидуальных, групповых или популяционных особенностях*, и весьма популярное в мировой компьютерной науке.

О высоком интересе к данной сфере свидетельствуют, в частности, международные исследовательские проекты: SEMIOTIKS (Semantically-Enhanced Information Extraction for Improved Knowledge Superiority), MIMEX (Multivariant Information Management and Exploitation), ITA (International Technology Alliance in Network and Information Sciences), AKT (Advanced Knowledge Technologies), IEXTREME (Extremist Ideological Influences on Group Decision Making). Последний включает в себя разработку таких техник, как анализ культурных структур, онтологии культур, системный анализ культуры (включая составление онтологий и классификацию [13].

С 2009 г. ежегодно проводятся международные соревнования в области идентификации по речевому поведению эмоционального состояния участников сетевой коммуникации, их демографических характеристик, личностных качеств и признаков психологического неблагополучия и психических патологий: INTERSPEECH Emotional Challenge, Speaker State Challenge, Speaker Trait Challenge, Computational Paralinguistic Challenge [14]. Другой пример – соревнования CLPsych (2015, 2016, 2017, 2018), направленные на разработку методов идентификации признаков различных видов психического неблагополучия (депрессия, посттравматическое расстройство, сезонное аффективное расстройство и др.) [15], и аналогичные по тематике CLEF (2013, 2014,

2015, 2016, 2017) [16]. В представленных разработках демонстрируется эффективность таких методов выделения и обработки данных, как поиск ключевых слов по словарям и тезаурусам, методы разметки последовательностей на основе случайных полей [17], методы автоматического выявления значимых символьных n -грамм, компьютерного лексического анализа, анализа параграфических особенностей текста, кластеризации и тематического моделирования [18]; кластеризации Брауна, основанной на алгоритме для вывода скрытой Марковской модели; спектральной кластеризации, основанной на расчете нормализованной матрицы PMI пар слов и спектральной кластеризации векторных представлений слов, построенных с помощью word2vec и GloVe; кластеризации на основе тематического моделирования с помощью LDA, построения линейной модели на основе SVM и логистической регрессии, а также различные ансамбли моделей на подмножествах признаков [19]; метод стохастического градиента с использованием в качестве признаков униграмм, взвешенных с помощью TF-IDF [20]; метаклассификации и обучения на неразмеченном и размеченном корпусе на основе случайного леса деревьев решений [21] и т.д.

Частным, но важным направлением в области автоматического анализа сетевого контента является мониторинг экстремистского сегмента Интернета, в котором пользователи, имеющие психологические проблемы, становятся участниками процессов распространения и навязывания эмоций и мнений, асоциального группирования, формирования негативных программ поведения [22, 23]. Здесь же находят применение и разработки, собственная тематика которых связана с социально-психологическими процессами формирования мнения [24], лидерства [25], влияния [26, 27] и др.

Необходимым средством для проведения подобных исследований являются лингвистические ресурсы, такие как WordNet, WordNet-Affect или SentiWordNet, благодаря которым в сетевом контенте определяется содержание маркированной (чаще всего эмоционально) лексики. Широко используется разработанный под руководством J.W. Pennebaker компьютерный инструмент Linguistic Inquiry and Word Count (LIWC), который позволяет определять частотность аффективно значимых лексем, а также маркеров других тем, таких, как «человек» или «активация», «доминант-

ность», и частоту встречаемости в тексте разных частей речи [28, 29].

В целом, исследования сетевого общения в интересах социо-гуманитарных дисциплин основаны, с одной стороны, на данных о структуре сетевой активности пользователей, с другой – на изучении продуктов этой активности невербального (видео, фото, изображения и т.п.) и вербального характера. Для оценки содержания текстов применяются методы определения ключевых слов, сантмент-анализа и, шире – лексико-частотного анализа, подразумевающего применение средств лексического, грамматического, синтаксического анализа, а также анализа параграфики. Отличительной особенностью предлагаемого нами подхода является характер выделяемых признаков (предикатов и синтаксис), которым соответствует более глубокий, по сравнению с лексическим, уровень организации *речемыслительной деятельности*. В общем, можно сказать, что предлагаемый метод предикатного анализа позволяет работать на уровне «протовысказываний», состоящих из аргумента (которым обычно является образ лица или предмета) и функции, т.е. предиката в широком смысле [30]. Структура «предикат – актанты» является когнитивной и пограничной по отношению к речи, отражая целостный образ ситуации внеязыковой действительности в картине мира человека [31]. Соответственно, в результате анализа предикатных структур, выделяемых в тексте, исследователь получает доступ к когнитивным механизмам, ответственным за функционирование картины мира – ментального образования, не имеющего непосредственной вербальной реализации, что крайне затрудняет его изучение, но при этом определяющего мировосприятие человека и его поведенческие программы. В соответствии со своей общей концептуальной основой предикатный анализ представляет собой реализацию *деятельностного подхода* к анализу текста.

2. Методологические аспекты использования автоматического анализа текста в социо-гуманитарных исследованиях

В рамках развиваемого деятельностного подхода к автоматическому анализу текста последний рассматривается как продукт комму-

никативно-интеллектуальной деятельности автора (ср.: «есть все основания рассматривать значение слова ... как единство обобщения и общения, коммуникаций и мышления» [32]), как продукт, возникающий при решении человеком конкретной задачи его жизнедеятельности, в которой создание текста оказывается необходимым средством при удовлетворении актуальной потребности. Для психологии такой подход к анализу роли текста, а значит и к интерпретации различных его лингвистических характеристик, может развиваться в рамках так называемой субъектной парадигмы. В экспериментальной психологии сенсорно-перцептивных процессов была продемонстрирована возможность двух способов интерпретации получаемых данных: в рамках объектной парадигмы ответы испытуемого рассматриваются как показатели возможностей психического аппарата как «естественного прибора», в рамках субъектной парадигмы, как показатели принятой субъектом стратегии решения задачи. Так, определенная численная характеристика различий в громкости двух звуковых сигналов, которые были испытуемым восприняты как разные, неодинаковые сигналы, может служить для описания дифференциального порога чувствительности слухового анализатора, а может — для описания перехода в процессе решения испытуемым задачи на различение сигналов с ориентации на сравнение их только по громкости к сравнению их и по всем другим доступным характеристикам (высоте, тембру и пр.). В последнем случае испытуемый, после серии неудачных попыток, вдруг демонстрирует способность правильно различать два близких по громкости сигнала. Происходит это не за счет того, что его барабанные перепонки стали лучше работать, а за счет того, что он изменил для себя задачу — не «различить громкость сигналов», а «различить сигналы, подаваемые экспериментатором». Соответственно, при интерпретации данных анализа текста как средства, созданного в определенных условиях и с определенной целью, субъектная парадигма требует искать для характеристик текста их обусловленности не только устройством языка, но и задачами, решаемыми человеком. В лингвистике такой взгляд на текст был принят со времен В. В. Виноградова и получил развитие в школе коммуникативной грамматики Г. А. Золотовой и в школе функционально-стилистического

анализа М. Н. Кожинной. Введение субъектной парадигмы в междисциплинарные исследования текста позволяет ставить задачу расширения возможностей автоматического анализа за счет выявления речевых средств, привлекаемых субъектом речевой деятельности при различных коммуникативных целях и при различных возможностях эмоциональной, когнитивной и поведенческой сфер.

В культурно-историческом подходе язык рассматривается как универсальное орудие, опосредствующее работу всего психического аппарата человека. Высшие психические функции человека, по мысли Л.С. Выготского, отличаются от натуральных познавательных и эмоциональных процессов животных именно включением языка как системы во все виды человеческой деятельности. Решение любой задачи у человека опосредствовано представлениями, сформированными в рамках «второй реальности» языка (термин А.Р. Лурии) и отражающими не только актуальную внешнюю ситуацию, но и широкие надситуативные связи, фиксируемые языком как системой. В связи с таким отношением к роли языка для психологии информационное и математическое моделирование когнитивных процессов человека, осуществляемое в ИИ, будет корректным только в случае учета опосредствованности этих процессов языком как системой. Отношение к тексту как к орудию деятельности лежит и в основе функционально-стилистического подхода: «Функциональная стилистика — это лингвистическая наука, изучающая особенности и закономерности функционирования языка в различных сферах человеческой деятельности и общения, а, следовательно — и «нормы» отбора и сочетания языковых средств...» [33, с. 154]. При этом подчеркивается и специфика человеческого языка как системы и связь этой системности с логикой человеческой деятельности: «...системность и закономерный характер функционирования языковых единиц определяются в первую очередь назначением и спецификой самого вида деятельности...» [33, с. 155]. В рамках этого лингвистического направления реализуется деятельностный подход, развиваемый в общей психологии в школе А.Н. Леонтьева, так что «речь рассматривается как явление социальное; индивидуальное же в речи не может и не должно быть объектом лингвистической стилистики, это область стилистики литерату-

роведческой» [33, с. 155]. Подчеркивается, что функциональную стилистику интересуют те закономерности речи, которые складываются «в результате функционирования языка в сфере сугубо социальной, например, публицистической, научной, деловой, художественной и пр.» [33, с. 156]. Это означает, что первой задачей, которую следует решать при анализе текста, является определение мотива той деятельности, в рамках которой был создан данный текст. Именно такой взгляд «сверху», ориентированный на выделение характеристик текста, обусловленных объективной социальной природой данного высказывания, позволяет выявить в тексте структуру речевых средств, использованных субъектом речевой деятельности. Эта структура будет обусловлена спецификой национального языка, особенностями языковой картины мира, характером концептосферы, обсуждаемой в тексте сферы действительности, языковой компетенцией автора текста, особенностями его личностной и когнитивной сфер, актуальным психическим статусом и пр. Все эти лингвистические и психологические механизмы оказываются «действующими факторами», влияние которых можно изучать, если мы смотрим на текст как на орудие решения социальной по происхождению задачи, на орудие деятельности.

Другими словами, если проводить экспериментальное исследование текста, то в качестве независимой переменной нужно брать деятельность, в рамках которой этот текст возник, а в роли зависимых переменных тогда будут выступать самые разные механизмы, отвечающие за порождение текста (лингвистические, психологические, социологические, исторические и проч.). В таком случае у нас возникает поле широких междисциплинарных исследований, в которых анализ текстов, созданных в рамках конкретной деятельности, служит задаче выявления сопряженного действия механизмов, изучаемых разными социо-гуманитарными дисциплинами. Если же мы фиксируем в исследовании еще одну переменную в качестве второй независимой, например, психиатрический статус авторов текстов или их политические убеждения, то у нас появляется возможность провести внутридисциплинарное исследование, опирающееся на доступные нам средства анализа текста.

В развиваемом нами подходе в качестве основного средства анализа текста используется реляционно-ситуационный анализ (РСА). Метод РСА [34], опирается на синтаксемный анализ Г.А. Золотовой [35] и на концепцию неоднородных семантических сетей Г.С. Осипова [4]. «Синтаксемой называется минимальная, далее неделимая синтактико-семантическая единица русского языка, выступающая одновременно как носитель элементарного смысла и как конструктивный компонент более сложных синтаксических построений, характеризующаяся, следовательно, определенным набором синтаксических функций» [36, с. 4]. В работе [4] значение этого термина раскрывается таким образом: «Так, например, в локативных формах *в лесу, за лесом, над лесом, из леса, из-за леса, возле леса* актуализируется значение места. ... В рассмотренных примерах синтаксемами (с локативным значением) являются «в+ предл. падеж» и «за+твор. падеж». С другой стороны, в предложении «*Старик (мальчик, рыбак) ловил неводом рыбу*», имя *старик* невозможно заменить, на имя, например, *свисток* – это разрушит предложение. Слова *старик, мальчик, рыбак* имеют общее категориальное значение – значение лица, *свисток* же имеет категориальное значение предмета и не может выполнять функцию субъекта. Таким образом, кроме морфологической формы слова решающую роль играет обобщенное значение, т.е. категориально-семантический класс слова». Значение синтаксемы оказывается и шире значения отдельного слова (локативом могут служить самые разные предметы) и одновременно – уже, конкретнее, поскольку вычленяет в предмете только одну из его многочисленных ролей: «Существительное *президент* как личное существительное в разных синтаксических конструкциях может выражать значения: субъекта (*Президент подписал указ*), объекта (*доверять президенту*), адресата (*обращение к президенту*), делиберата (*говорить о президенте*), комитатива (*Патриарх с президентом зашли в храм*), предиката (*Иванов – президент страны*)» [4]. Таким образом, синтаксема, с одной стороны, представляет значение не одного предмета, а целого класса, но, с другой, – уточняет значение конкретного предмета в конкретной ситуации.

Для психологии понятие синтаксемы соотносится с представлением Л.С. Выготского о

роли значения как единицы анализа речевого мышления (см. цитату выше). А.Н. Леонтьев, развивая идеи Л.С. Выготского, определяет значение как тот компонент сознательного представления, в котором отражен общественно-исторический опыт использования предмета. Можно сказать, что важнейшее для культурно-исторического подхода представление о языковом опосредствовании высших психических функций человека нуждается именно в представлении о синтаксеме как о единице анализа языка. Ведь, в отличие от значения отдельного слова, лишь указывающего на предмет, синтаксема отражает ситуацию бытования предмета в культуре, она представляет обстоятельства его использования в человеческом обществе, характеризует тот класс предметов, которые являются эквивалентными по значению для исполняемого действия. В этом плане скорее значение, представленное в синтаксеме, чем словарное значение слова, является единицей речевого мышления - обобщению подвергается сам момент действия с предметом и предмет предстает в той роли, которую он играет в конкретной ситуации. Именно такой, ориентированный на вычленение принятых в культуре действий с предметами, «ролевой» лингвистический анализ текста адекватен задачам психологического анализа продукта речевой деятельности.

Для анализа предложения важно не только знание значений встречающихся в нем синтаксем, но и их сочетаемость друг с другом. Эта сочетаемость определяется семейством бинарных отношений на множестве синтаксем. Реляционно-ситуационный анализ позволяет выявлять семантику текста, ставя в соответствие синтаксемную структуру предложения логической структуре действий, описанных в этом предложении. Главную роль здесь играют глаголы, имеющие, как правило, центральное положение в семантической структуре предложения и оказывающие решающее влияние на именные словосочетания и предложения, составляющие синтаксемные группы.

Машина РСА, построенная на так организованном автоматическом анализе, позволяет исследователям-гуманитариям решать целый ряд задач, связанных с экспликацией представлений и отношений автора текста к реалиям, о которых он пишет:

- на основе анализа одного текста выяснять в каких семантических ролях и с какими предикатами встречается в тексте интересующий исследователя объект;

- на основе исследования коллекции текстов автора представить его картину мира в виде системы значимых для него объектов и действий с ними;

- на основе изучения коллекций текстов разных авторов на заданную тему представить мнения авторов как пространство убеждений/отношений.

Выявление синтаксемно-предикатной структуры высказывания позволяет операционализировать интуитивные представления исследователя-гуманитария о смыслах, придаваемых автором текста, обсуждаемым им темам.

Использование тех или иных языковых средств автором не является случайным событием, речевой процесс не является по своей природе вероятностным, «мысль не выражается, но совершается в слове» [32], а значит факторы, повлиявшие на коммуникативные намерения автора, и механизмы, обусловившие появление у него некоторой мысли, некоторого содержимого коммуникативного процесса, могут стать предметом изучения при анализе текста. Следовательно, анализ текста в социогуманитарных исследованиях не может сводиться к подсчету тех лексических, синтаксических и морфологических характеристик текста, которые стали доступны благодаря автоматизации. Процедура исследования должна быть обратной – от гипотезы о действующих в коммуникативной ситуации автора механизмах (психологических, социологических, политологических и пр.) к проверке того, какие языковые средства оказываются им релевантны. Так, при изучении средств, используемых авторами научных публикаций, необходимо научно обоснованное представление о действиях, характерных для научной деятельности (например, [37]), тогда методы интеллектуального анализа текста позволяют выделить операциональный состав речемыслительных действий, специфичных для разных типов научных работ и разных научных традиций [38]. При изучении же особенностей текстов сетевых дискуссий с точки зрения их фрустрационного потенциала требуется, соответственно, исследование связей между типами реакции на фрустрацию, выделяемыми в психологии, и

данными, получаемыми с помощью методов ИИ. В исследованиях, посвященных реакциям населения на обострение социальных проблем, когда эти реакции уже нашли выражение в текстах - фокус-групп, комментариев в сетевых СМИ или, если объектом исследования стали исторические реалии XX в., в «письмах в редакцию» или в дневниковых записях. Такой теоретической основой могут стать представления о стрессе, развиваемые в медицине, психологии и экспериментальной физиологии. В целом, именно хорошая теория из области социальных и гуманитарных наук должна лежать в основе тех процедур разметки корпуса текстов и последующего машинного обучения, которые применяются в работах по интеллектуальному анализу текста.

Заключение

Информационные системы находят широкое применение в мировой социо-гуманитарной исследовательской практике. В отличие от большинства предлагаемых сейчас гуманитариям моделей анализа текста и средств извлечения информации о состоянии его автора, метод РСА, выявляющий предикатно-синтаксическую структуру высказываний, позволяет преодолеть ориентацию на относительно легко формализуемые признаки (лексические, грамматические и морфологические). С психологической точки зрения предикатно-синтаксические структуры отражают глубинные, пограничные для доречевых ментальных процессов и речевой деятельности, структуры, соотносимые с языковой картиной мира личности, что дает доступ к ее смысловой сфере, изучение которой традиционно представляется как проблематичное. Благодаря выходу на смысловой уровень анализа, реализующий РСА инструмент – «машина РСА», описанию которой будет посвящена вторая часть настоящей работы, – делает возможным постановку новых задач в области социо-гуманитарных исследований.

Литература

1. Голенок М.П., Осипова Н.О. Digital humanities: проблемное поле и перспективы развития // Научное обозрение. 2018. № 1. URL: <https://cyberleninka.ru/article/n/digital-humanities-problemnoe-pole-i-perspektivy-razvitiya>
2. Базенков Н.И., Губанов Д.А. Обзор информационных систем анализа социальных сетей // Управление большими системами: сборник трудов. 2013. № 41. С. 357-394.
3. Van Royen, K., Poels, K., Vandebosch, H., Adam, P. Thinking before posting? Reducing cyber harassment on social networking sites through a reflective message // Comput. Hum. Behav., 2017, 66: 345–352.
4. Осипов Г.С. Приобретение знаний интеллектуальными системами: Основы теории и технологии. М.: Наука, Физматлит, 1997.
5. Михеенкова М.А. Принципы и логические средства интеллектуального анализа социологических данных. Дис. ... док. техн. наук. М., 2012.
6. Бурковская Ж.И., Михеенкова М.А., Финн В.К. Об интеллектуальной системе для анализа электорального поведения / IX Национальная конференция с международным участием «Искусственный интеллект-2004», Тверь, сентябрь 8-11, 2004, Труды конференции в 2 т. Т.1. С. 120-128.
7. Ясницкий Л.Н. О возможностях применения методов искусственного интеллекта в политологии // Вестн. Перм. ун-та. Сер.: Политология. 2008. № 2. С. 147-155.
8. Тихомиров И.А., Смирнов И.В., Соченков И.В., Девяткин Д.А., Шелманов А.О., Зубарев Д.В., Швец А.В., Лешкин А.В., Суворов Р.Е. Exactus Expert: Поисково-аналитическая система поддержки научно-технической деятельности / Труды тринадцатой национальной конференции по искусственному интеллекту с международным участием КИИ-2012. Б.: БГТУ, 2012. Т. 4. С. 100-108.
9. Тихомиров И.А., Соченков И.В., Швец А.В. Наукометрия и полнотекстовая аналитика в российских реалиях // Научно-технические исследования. № 1. С. 197-212.
10. Роганов В.Р., Роганова С.М., Новосельцева М.Е. Методы искусственного интеллекта для машинного перевода текстов. Пенза, 2007.
11. Шелманов А.О., Каменская М.А., М.И. Ананьева М.И., Смирнов И.В. Семантико-синтаксический анализ текстов в задачах вопросно-ответного поиска и извлечения определений // Искусственный интеллект и принятие решений. 2016. № 4. С. 47-61.
12. Sochenkov, I., Zubarev, D., Tikhomirov, I., Smirnov, I., Shelmanov, A., Suvorov, R., & Osipov, G. Exactus Like: Plagiarism Detection in Scientific Texts // Advances in Information Retrieval. Springer International Publishing, 2016: 837-840.
13. <https://www.ecs.soton.ac.uk/research/projects/746>
14. <http://compare.openaudio.eu/>
15. http://clpsych.org/shared_task/
16. <http://clef2017.clef-initiative.eu/index.php>
17. Yang, H., Willis, A., de Roeck, A., & Nuseibeh, B. A Hybrid Model for Automatic Emotion Recognition in Suicide Notes // Biomedical Informatics Insights, 2012, 5 (Suppl. 1): 17–30.
18. Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. Clpsych 2015 shared task: depression and ptsd on twitter // Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado, 2015: 31–39.
19. Preotiuc-Pietro, D., Sap, M., Schwartz, H. A., & Ungar, L. Mental illness detection at the world well-being project for the clpsych 2015 shared task // Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psy-

- chology: From Linguistic Signal to Clinical Reality, Denver, Colorado, 2015: 40–45.
20. Mac Kim, S., Wang, Y., Wan, S., & Paris, C.. Data61-csiro systems at the clpsych 2016 shared task // Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, 2016: 128–132.
 21. Malmasi, S., Zampieri, M., & Dras, M. Predicting post severity in mental health forums // Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology, San Diego, CA, USA, 2016: 133–137.
 22. Rieger, D., Frischlich, L., & Bente, G. Propaganda 2.0: Psychological effects of right-wing and Islamic extremist Internet videos, 2013. URL: <https://www.researchgate.net/publication/264829254>
 23. Berger, J.M. The Metronome of Apocalyptic Time: Social Media as Carrier Wave for Millenarian Contagion // Perspectives on Terrorism, 2015, 9(4): 61–71.
 24. Yang Shan-lin, Zhu Ke-yu, Fu Chao, & Lu Guang-yan. Simulation of the group decision conformity based on cellular automata model // System Engineering-Theory & Practice, 2009, (9): 115–124.
 25. Ma Yu-tao, Cai Shu-qin, & Wang Rui Study on the method of identifying opinion leaders based on online customer reviews // Conference: Management Science and Engineering (ICMSE), 2011. URL: https://www.researchgate.net/publication/261275952_Study_on_the_method_of_identifying_opinion_leaders_based_on_online_customer_reviews
 26. Barbieri, N., Bonchi, F., & Manco, G. Who to follow and why: Link Prediction with explanations // KDD'14, August 24–27, 2014, New York, NY, USA.
 27. Vollenbroek, W., de Vries, S., Constantinides, E., & Kommers, P. Identification of influence in social mediacommunities // Int. J. Web Based Communities, 2014, 10(3).
 28. Choudhury, M.D, Gamon, M., Counts, S., & Horvitz, E. Predicting Depression via Social Media // ICWSM, 2013, 2. URL: <http://research.microsoft.com/en-us/um/people/counts/pubs/depressionicws2013.pdf>
 29. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. The Development and Psychometric Properties of LIWC2007. URL: http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/liwc2007_languageanal.pdf
 30. Леонтьев А.А. Основы психолингвистики. М.: Смысл, 1997.
 31. Курдюмов В.А. Идея и форма. Основы предикационной концепции языка. М.: Военный ун-т, 1999.
 32. Выготский Л.С. Мышление и речь. М.: Лабиринт, 1999.
 33. Кожина М.Н. Речеведение. Теория функциональной стилистики. М.: Флинта: Наука, 2014.
 34. Золотова Г.А., Онипенко Н.К., Сидорова М.Ю. Коммуникативная грамматика русского языка. М.: Ин-т рус. яз. РАН им. В.В. Виноградова, 2004.
 35. Осипов Г.С., Смирнов И.В., Тихомиров И.А. Реляционно-ситуационный метод поиска и анализа текстов и его приложения // Искусственный интеллект и принятие решений. 2008. № 2. С. 3–10.
 36. Золотова Г.А. Синтаксический словарь: Репертуар элементарных единиц русского синтаксиса. М.: Наука, 1988.
 37. Салимовский В.А. Жанры речи в функционально-стилистическом освещении: Русский научный академический текст. Дисс. ... док. филол. наук. Екатеринбург, 2002.
 38. Девяткин Д.А., Кузнецова Ю.М., Чудова Н.В. Методы автоматического выявления ментальных действий в текстах научных публикаций // Искусственный интеллект и принятие решений. 2018. № 2. С. 36–46.

Creating a text analysis tool for socio-humanitarian research

P.1. Methodical and methodological aspects

S.N. Enikolopov¹, Y.M. Kuznetsova¹, I.V. Smirnov¹, M.A. Stankevich¹, N.V. Chudova¹

¹Mental Health Research Center of Russian Academy of Sciences, Moscow, Russia

¹Federal Research Center "Computer Science and Control" of Russian Academy of Sciences, Moscow, Russia

Abstract. The first part of the work conceptually examines applications of artificial intelligence in socio-humanitarian studies. The world trend of the last decades is focused on the involvement of social media data for research interests, which allows solving the tasks of monitoring, analysis, forecasting and management in relation to such parameters of network communication as opinions, attitudes, emotions, and behavior patterns of communicants. In order to develop tools for network content analysis, we propose to utilize the method for relational-situational analysis that based on the syntactic analysis approach by G.A. Zolotova and the concept of heterogeneous semantic networks proposed by G.S. Osipov. The proposed approach makes it possible to consider texts as the product of the author's speech-thinking activity, which allows applying Russian psychological research background to study cognitive processes, personality, motivation, emotional states, individual differences, and the general worldview.

Keywords: socio-humanitarian research, artificial intelligence, automatic text analysis, relational-situational analysis, predicate-argument structure

DOI 10.14357/20718594190203

References

- Golenok M.P., Osipova N.O. Digital humanities: problemnoe pole i perspektivy razvitiya [Digital Humanities: a problematic sphere and development tendency] // Nauchnoe obozrenie [Scientific Review]. 2018. № 1. URL: <https://cyberleninka.ru/article/n/digital-humanities-problemnoe-pole-i-perspektivy-razvitiya>
- Bazhenkov N.I., Gubanov D.A. Obzor informacionnyh sistem analiza social'nyh setej [Digital humanities: a problematic sphere and development tendency] // Upravlenie bol'shimimi sistemami: sbornik trudov [Management of large systems: a collection of works.]. 2013. № 41. S. 357-394.
- Van Royen, K., Poels, K., Vandebosch, H., Adam, P. Thinking before posting? Reducing cyber harassment on social networking sites through a reflective message // Comput. Hum. Behav., 2017, 66: 345–352.
- Osipov G.S. Priobretenie znaniy intellektual'nymi sistemami: Osnovy teorii i tekhnologii [Acquisition of knowledge by intellectual systems: Fundamentals of theory and technology]. M.: Nauka, Fizmatlit [Science, Fizmatlit], 1997.
- Miheenkova M.A. Principy i logicheskie sredstva intellektual'nogo analiza sociologicheskikh dannyh [Principles and logical tools for the intellectual analysis of sociological data]. Dis. ... dokt. tekhn. nauk. [Doctoral thesis] M., 2012.
- Burkovskaya ZH.I., Miheenkova M.A., Finn V.K. Ob intellektual'noj sisteme dlya analiza ehlektoral'nogo povedeniya [Intellectual system for electoral behavior analysis] // IX Nacional'naya konferenciya s mezhdunarodnym uchastiem "Iskusstvennyj intellekt-2004" [IX National Conference with International Participation "Artificial Intelligence-2004"], 8-11, 2004, Trudy konferencii [Proceedings of the Conference] v. 2 t. T.1. S. 120-128.
- YAsnickij L.N. O vozmozhnostyah primeneniya metodov iskusstvennogo intellekta v politologii [the opportunities of application of artificial intellectual methods in political science] // Vestn. Perm. un-ta. Ser.: Politologiya [Political science]. 2008. № 2. S. 147-155.
- Tihomirov I.A., Smirnov I.V., Sochenkov I.V., Devyatkin D.A., Shelmanov A.O., Zubarev D.V., Shvec A.V., Leshkin A.V., Suvorov R.E. Exactus Expert: Poiskovo-analiticheskaya sistema podderzhki nauchno-tekhnicheskoy deyatel'nosti [Exactus Expert: Search-analytical system for supporting scientific and technical activities]// Trudy trinahtatoj nacional'noj konferencii po iskusstvennomu intellektu s mezhdunarodnym uchastiem KII-2012. [Proceedings of the Thirteenth National Conference on Artificial Intelligence with International Participation KII-2012] B.: BGTU, 2012. T. 4. S. 100-108.
- Tihomirov I.A., Sochenkov I.B., Shvec A.V. Naukometriya i polnotekstovaya analitika v rossijskikh realiyah [Naukometriya and full-text analytics in the Russian realities] // Zhurnal "Naukovedcheskie issledovaniya" [Journal "Science research"]. № 1. S. 197-212.
- Roganov V.R., Roganova S.M., Novosel'ceva M.E. Metody iskusstvennogo intellekta dlya mashinnogo perevoda tekstov [Methods of artificial intelligence for machine translation of texts]. Penza, 2007.
- Shelmanov A.O., Kamenskaya M.A., M.I. Anan'eva M.I., Smirnov I.V. Cementiko-sintaksicheskij analiz tekstov v zadachah voprosno-otvetnogo poiska i izvlecheniya opredelenij [Semantic-syntactic analysis for question answering and definition extraction] //Iskusstvennyj intellekt i prinyatie reshenij. [Artificial Intelligence and Decision Making] M.: ISA RAN. 2016. № 4.S. 47-61.
- Sochenkov, I., Zubarev, D., Tikhomirov, I., Smirnov, I., Shelmanov, A., Suvorov, R., & Osipov, G. Exactus Like: Plagiarism Detection in Scientific Texts //Advances in Information Retrieval. Springer International Publishing, 2016: 837-840.
- <https://www.ecs.soton.ac.uk/research/projects/746>
- <http://compare.openaudio.eu/>
- http://clpsych.org/shared_task/
- <http://clef2017.clef-initiative.eu/index.php>
- Yang, H., Willis, A., de Roeck, A., & Nuseibeh, B. A Hybrid Model for Automatic Emotion Recognition in Suicide Notes // Biomedical Informatics Insights, 2012, 5 (Suppl. 1): 17–30.
- Coppersmith, G., Dredze, M., Harman, C., Hollingshead, K., & Mitchell, M. Clpsych 2015 shared task: depression and ptsd on twitter //Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado, 2015: 31–39.
- Preotiuc-Pietro, D., Sap, M., Schwartz, H. A., & Ungar, L. Mental illness detection at the world well-being project for the clpsych 2015 shared task // Proceedings of the 2nd Workshop on Computational Linguistics and Clinical Psychology: From Linguistic Signal to Clinical Reality, Denver, Colorado, 2015: 40–45.
- Mac Kim, S., Wang, Y., Wan, S., & Paris, C.. Data61-csro systems at the clpsych 2016 shared task // Proceedings of the Third Workshop on Computational Linguistics and Clinical Psychology, 2016: 128–132.
- Malmasi, S., Zampieri, M., & Dras, M. Predicting post severity in mental health forums // Proceedings of the 3rd Workshop on Computational Linguistics and Clinical Psychology, San Diego, CA, USA, 2016: 133–137.
- Rieger, D., Frischlich, L., & Bente, G. Propaganda 2.0: Psychological effects of right-wing and Islamic extremist Internet videos, 2013. URL: <https://www.researchgate.net/publication/264829254>
- Berger, J.M. The Metronome of Apocalyptic Time: Social Media as Carrier Wave for Millenarian Contagion // Perspectives on Terrorism, 2015, 9(4): 61-71.
- Yang Shan-lin, Zhu Ke-yu, Fu Chao, & Lu Guang-yan. Simulation of the group decision conformity based on cellular automata model // System Engineering-Theory & Practice, 2009, (9): 115-124.
- Ma Yu-tao, Cai Shu-qin, & Wang Rui Study on the method of identifying opinion leaders based on online customer reviews // Conference: Management Science and Engineering (ICMSE), 2011. URL:https://www.researchgate.net/publication/261275952_Study_on_the_method_of_identifying_opinion_leaders_based_on_online_customer_reviews
- Barbieri, N., Bonchi, F., & Manco, G. Who to follow and why: Link Prediction with explanations // KDD'14, August 24–27, 2014, New York, NY, USA.
- Vollenbroek, W., de Vries, S., Constantinides, E., & Kommers, P. Identification of influence in social media-communities // Int. J. Web Based Communities, 2014, 10(3).

28. Choudhury, M.D, Gamon, M., Counts, S., & Horvitz, E. Predicting Depression via Social Media // ICWSM, 2013, 2. URL: <http://research.microsoft.com/en-us/people/counts/pubs/depressionicwsm2013.pdf>
29. Pennebaker, J.W., Chung, C.K., Ireland, M., Gonzales, A., & Booth, R.J. The Development and Psychometric Properties of LIWC2007. URL: http://homepage.psy.utexas.edu/homepage/faculty/pennebaker/reprints/liwc2007_languageanal.pdf
30. Leont'ev A.A. Osnovy psiholingvistiki [Fundamentals of psycholinguistics]. M.: Smysl, 1997.
31. Kurdyumov V.A. Ideya i forma. Osnovy predikacionnoj koncepcii yazyka [Idea and form. Fundamentals of the predicate concept of language]. M.: Voennyj un-t, 1999.
32. Vygotskij L.S. Myshlenie i rech' [Thinking and speaking]. M.: Labirint [Maze], 1999.
33. Kozhina M.N. Rechevedenie. Teoriya funkcional'noj stilistiki [Speech. Theory of functional stylistics.]. M.: FLINTA: NAUKA, 2014.
34. Zolotova G.A., Onipenko N.K., Sidorova M.YU. Kommunikativnaya grammatika russkogo yazyka [Communicative grammar of the Russian language]. M.: In-t rus. yaz. RAN im. V.V. Vinogradova [Institute of Russian language of RAS of V.V. Vinogradov], 2004.
35. Osipov G.S., Smirnov I.V., Tihomirov I.A. Relyacionno-situacionnyj metod poiska i analiza tekstov i ego prilozheniya [Relational-situational method of search and analysis of texts and its applications] // Iskusstvennyj intellekt i prinyatie reshenij [Artificial intelligence and decision making]. 2008. № 2. S. 3-10.
36. Zolotova G.A. Sintaksicheskij slovar': Repertuar ehlemen-tarnyh edinic russkogo sintaksisa [Syntax dictionary: The repertoire of elementary units of Russian syntax]. M.: Nauka [Science], 1988.
37. Salimovskij V.A. Zhanry rechi v funkcional'no-stilisticheskom osveshchenie: Russkij nauchnyj akademicheskij tekst [Genres of speech in functional-stylistic lighting: Russian academic academic text]. Diss. ... dokt .filol .nauk. [Doctoral thesis], 2002.
38. Devyatkin D.A., Kuznecova YU.M., CHudova N.V. Metody avtomaticheskogo vyyavleniya mental'nyh dejstv-vij v tekstah nauchnyh publikacij [Methods of automatic detection of mental actions in the texts of scientific publications] // Iskusstvennyj intellekt i prinyatie reshenij [Artificial intelligence and decision making]. 2018. № 2. C. 36-46.