ФЕДЕРАЛЬНОЕ АГЕНТСТВО НАУЧНЫХ ОРГАНИЗАЦИЙ

ВСЕРОССИЙСКИЙ ИНСТИТУТ НАУЧНОЙ И ТЕХНИЧЕСКОЙ ИНФОРМАЦИИ
РОССИЙСКОЙ АКАДЕМИИ НАУК
(ВИНИТИ РАН)

НАУЧНО • ТЕХНИЧЕСКАЯ ИНФОРМАЦИЯ

Серия 2. ИНФОРМАЦИОННЫЕ ПРОЦЕССЫ И СИСТЕМЫ

ЕЖЕМЕСЯЧНЫЙ НАУЧНО-ТЕХНИЧЕСКИЙ СБОРНИК

Издается с 1961 г. **№ 6** **Москва 2017**

ОБЩИЙ РАЗДЕЛ

УДК 167/168 : [001.102 : 004 : 001.2]

В.А. Канке

Метанаучные и философские основания определения статуса информатики

Рассмотрены основные положения метанауки и философии науки. В их контексте даны определения науки, отрасли науки и двух смежных по отношению к информатике наук – математики и технологии. Определен статус информатики как формальной отрасли науки, состоящей из теорий, в каждой из которых совершается конструктивный концептуальный переход: принципы алгоритмических вычислений – парадигмы программирования – программы – исполнение программ посредством компьютеров.

Ключевые слова: *естественные, аксиологические и формальные науки, метанаука, философия науки, информатика*

Многочисленные авторы, определяющие статус информатики, так или иначе используют сложившиеся представления о природе метанауки и философии науки. К сожалению, при этом им редко удается проявить в аргументации должную последовательность. С учетом этого обстоятельства мы, прежде всего, обращаемся к основным положениям метанауки и философии науки, которые в достаточно пространном виде изложены во многих наших работах, в частности в монографиях [1–3].

ОСНОВНЫЕ ПОЛОЖЕНИЯ МЕТАНАУКИ И ФИЛОСОФИИ НАУКИ

Положение 1. Базисом научного знания являются теории (*концепции*). Понимание статуса современной науки предполагает использование определенной классификационной системы с соответствующими таксонами. В связи с этим широкую популярность приобрели три таксона, а именно: отрасли науки (англ. *fields of science*), дисциплины и субдисципли-

ны [4, 5]. В России, в соответствии с принятой Всероссийской аттестационной комиссией номенклатурой научных специальностей также используют три основных таксона, заменяя дисциплины и субдисциплины соответственно на группы наук и отдельные науки [6]. Примерами отраслей науки являются математика, физика, экономика, По сути, рассматриваемые классификационные системы тождественны друг другу. Их характерная особенность состоит в том, что исследовательская мысль движется сверху вниз, от более обширных таксонов к их регионам: отрасли науки → дисциплины → субдисциплины. Как показывает анализ, при этом не достигается фундамент научного знания, его почва. Любая субдисциплина из тех, что указаны в [4–6], состоит из многих теорий. Всякое знание укомплектовано в теории и именно они составляют поэтому базу науки. Классификационная система научного знания должна строиться не сверху – вниз, а снизу – вверх, только в этом случае можно определить статус отрасли науки, на высокое звание которой претендует информатика. К сожалению, в современной науке не реализована предлагаемая схема классификации, неизвестно даже число ступеней классификационной лестницы, которая могла бы привести от отдельных теорий к отраслям науки. Это обстоятельство не отменяет Положение 1.

Положение 2. Нет ничего такого, о чем бы можно было бы судить не на основании теорий, а каким-то другим образом. По сути, мы представили принцип теоретической репрезентации: нет ничего такого, что не было бы представлением теорий. Основные ходы мысли, которые ведут к принципу теоретической репрезентации, были продемонстрированы И. Кантом, К. Поппером, А. Эйнштейном, У. Куайном, Н. Хансэном, Т. Куном и П. Фейерабендом. Однако никто из них не представил рассматриваемый принцип в отчетливом виде. В противном случае давно была бы сдана в архив явно устаревшая концепция, согласно которой теориям необходимо противопоставлять реальность и практику. Именно так считают, в частности, марксисты и прагматисты, которые придают теориям всего лишь инструментальную значимость. Тела, практика, ментальность и язык – это представления теории. За пределы теории человеку не суждено выйти. Разумеется, это утверждение истинно лишь при правильном понимании природы теории, а не отождествлении ее всего лишь с ментальностью или языком человека.

Положение 3. Мир природы есть представление естественных теорий; мир человека – представление аксиологических теорий. О природе невозможно судить, игнорируя физические, химические, геологические и биологические теории. Мир человека невозможно определить без технических, агрологических, медицинских, психологических, педагогических, экономических, социологических, политологических, юридических, исторических и искусствоведческих теорий. О природе судят на основании естественных теорий, о природе человека рассуждают, исходя из содержания аксиологических теорий. Концепты и методы всех аксиологических теорий являются предпочтениями людей, т. е. их ценностями. Природные объекты, в том числе высшие млекопитающие, не руководствуются ценностями людей. Мы обратились к естественным и аксиологическим теориям не случайно, а в связи с их особой значимостью в деле интерпретации природы формальных теорий. Именно стартуя от естественных и аксиологических теорий, можно понять природу формальных, например, логических и математических, концепций.

Содержание данного раздела показывает, что неправомерно различение, с одной стороны, теоретических, а с другой стороны, нетеоретических, например, экспериментальных, теорий. Любая теория не может быть нетеоретической. Выражения "теоретическая физика" и «теоретическая информатика» не выдерживают критики. Разумеется, информатика в качестве теории имеет теоретический характер. Нетеоретической информатики не бывает.

Положение 4. Формальные теории выражают изоморфные отношения между концепциями. Наряду с естественными и аксиологическими теориями есть также формальные теории. Постижение их природы оказалось связанным с существенными трудностями. Господствующая позиция состоит в том, что формальные науки имеют дело с объектами, полученными посредством применения к реальным объектам операций абстракций и идеализаций. Эти объекты, образно говоря, режут на части и шлифуют до тех пор, пока из них не получают, точки, плоскости, числа и прочие весьма своеобразные сущие. Например, число 3 есть абстрактное сущее, которое можно приписать любой тройке предметов, например, трем лебедям или трем АЭС. Рассматриваемое воззрение было подвергнуто резкой критике логиком, математиком и философом Г. Фреге. Характеризуя природу математики, он отмечал, что "существует отношение φ, которое взаимно однозначно соотносит предметы, подпадающие под понятие F, с предметами, подпадающими под понятие G" [7, с. 208]. Между тремя лебедями, с одной стороны, и тремя АЭС, с другой стороны, существует изоморфное отношение, только и всего. В данном случае, чтобы его выразить нет никакой необходимости обращаться ни к абстракциям, ни к идеализациям. Строго говоря, операции абстрагирования и идеализации используются и в естественных, и в аксиологических, и в формальных теориях. Они не являются прерогативой формальных наук. Имея дело с изоморфными отношениями, формальные теории выражают определенные классы эквивалентности между различными теориями, причем как естественными, так и аксиологическими, а, при случае, также формальными. Формальные теории не привязаны жестко ни к миру природы, ни к миру человека.

Сторонники понимания объектов формальных наук в качестве абстракций рассуждают крайне непоследовательно. Например, утверждается, что абстрагирование выступает как отбрасывание несущественного. Но что именно является несущественным, не объясняется. Таких исследователей, в частности, любителей математических моделей, обвиняют в том, что они отклоняются от реальности, т.е. искажают действительность. Остается неясным почему именно столь эффективны формальные теории. Знаменитый физик

Е. Вигнер удивлялся, как он выражался, "непостижимой эффективности" математики [8]. Затруднительно объяснить эффективность математики, понимая ее объекты в качестве искажений реальных объектов. Концепция понимания объектов формальных наук в качестве изоморфных отношений с указанными затруднениями не встречается. Они ведь ничего не искажают, а выражают потенциал самих теорий. К формальным концепциям относятся лингвистические, логические, математические и философские теории.

Из содержания данного раздела следует, что нет так называемых общих и общенаучных теорий. Есть в том или ином отношении эквивалентные, но не общие теории. Каждая теория не терпит ничего чужеродного. В физике нет экономического, в экономике отсутствует физическое; но физическое и экономическое может быть эквивалентным, например, там и там используется линейная зависимость между переменными.

Нет ни общей биологии, ни общей психологии, ни общей информатики.

Положение 5. В естественных и аксиологических теориях имеет место четырехзвенная концептуальная трансдукция. Предсказание осуществляется посредством дедукции. Затем в процессе экспериментов, наблюдений и практических поступков имеет место фактуализация. Она происходит под эгидой аддукции (избираются приборы, нивелируются побочные факторы, измеряются переменные). Обработка фактов (данных) происходит посредством индукции, которая предполагает использование дисперсионного, корреляционного, регрессионного и других видов анализа, позволяющих определять средние значения переменных, законы и экстремальные принципы. Обработка данных создает базу для заключительного этапа интратеоретической трансдукции, а именно – обновления дедуктивных принципов. Такое обновление совершается под эгидой абдукции (обновление проводится таким образом, чтобы зафиксированные индуктивные законы и принципы можно было вывести непосредственно из новых дедуктивных принципов). Обновленные дедуктивные принципы позволяют осуществить новый цикл познания.

Положение 6. В формальных теориях отсутствует фактуализация и, следовательно, следующие за ней стадии обработки данных и обновления дедуктивных принципов. Тем не менее, последние обновляются, но лишь в составе предсказания. Если здесь обнаруживаются противоречия, то основания теории, будь то исходные аксиомы или объекты построения, изменяются таким образом, чтобы элиминировать указанные противоречия.

Положение 7. Проблемные ряды теорий в процессе их осмысления преобразуются в лигатеории. Допустим, что прогресс познания выразился в поступи трех теорий (Т): $T_1(p_1) \rightarrow T_2(p_2) \rightarrow T_3$. Проблемы, выявленные в составе первой теории (p_1), были преодолены в теории T_2. Затем были выявлены проблемы второй теории (p_2). Их удалось преодолеть благодаря теории T_3. Ее затруднения пока не выявлены. Исследователям нет никакой необходимости продолжать придерживаться устаревших воззрений $T_1(p_1)$ и $T_2(p_2)$. Они заменяются их усовершенствованными вариантами: $T_1\{T_3\}$ и $T_2\{T_3\}$. Запись $T_1\{T_3\}$ означает, что концептуальное содержание теории T_1интерпретируется с позиций теории T_3. В итоге проблемный ряд теорий заменяется их интерпретационным строем: $T_3 \rightarrow T_2\{T_3\} \rightarrow T_1\{T_3\}$. Такая концептуальная конструкция обладает внутренним единством, ибо былая разобщенность теорий T_1, T_2 и T_3 преодолена. Мы называем ее лигатеорией (от лат. *ligare* – связывать) [3, с. 185]. Примерами лигатеорий являются электродинамика Дирака – Максвелла – Эйнштейна, теория трудовой стоимости Маркса – Риккардо – Смита – Петти, последовательность версий языкаC#: Версия 6.0 – Версия 5.0 – Версия 4.0 – Версия 3.0 – Версия 2.0 – Версия 1.0.

Переход к лигатеориям совершается посредством циклов интертеоретической трансдукции, которая включает три этапа: 1) выявление проблем у частично устаревшей теории, 2) открытие новой теории, которая позволяет преодолеть выявленные проблемы, 3) интерпретацию частично устаревшей теории с позиций самой развитой концепции. Методы этих этапов мы называем соответственно проблематизацией, открытием и интерпретацией. Строго говоря, ученые мыслят не отдельными теориями, а их связками, т.е. лигатеориями.

Положение 8. Между лигатеориями существуют символические связи (отношения). Эти отношения в англоязычной литературе называют интердисциплинарными (англ. *interdisciplinary relationes*). Нет необходимости возражать против такого словоупотребления, но лишь при условии, что дисциплины отождествляются с лигатеориями. Таким образом, строго говоря, речь идет об интерлигатеоретических отношениях. При реализации таких отношений одна лигатеория является акцепторной, а все другие донорными. Допустим, рассматривается отношение между биологической и физической лигатеориями. Если это делается от имени биологии, то биологическая теория является главной теорий. Ее резонно называть акцепторной постольку, поскольку она определенным образом воспринимает достижения физической лигатеории. Исследование может вестись ради развития физики. В таком случае биологическая лигатеория выполняет роль донорной концепции. Существенно, что донорная концепция всегда рассматривается как символ акцепторной концепции. Так, радиобиолог не занимается решением физических уравнений, его интересуют физические факторы лишь постольку, поскольку они имеют значение для характеристики сугубо биологических явлений. Но почему, на наш взгляд, в данном случае целесообразно использовать концепт символа? Мы полагаем, что для характеристики отношений в соответствии с изысканиями Ч.С. Пирса лучше всего подходит аппарат семиотики с ее тремя центральными концептами, а именно – с иконами, симптомами и символами. Соотносительность концептов Ч.С. Пирс характеризовал посредством понятия символа [9, с. 96].

Положение 9. Лигатеории составляют базис всего научного знания. Этот вывод также представляется нам существенным. Отдельные понятия бытуют не иначе, как в составе теорий. Теории входят в состав

лигатеорий. Лигатеории не входят в состав какого-либо другого, более содержательного концептуального образования. Но, разумеется, между ними существуют определенные интерлигатеоретические отношения. Выше отмечалось, что классификационный конус: *отрасли науки – дисциплины – субдисциплины* не доходит до базиса научного знания. Теперь есть возможность констатировать, что этот базис образуют лигатеории, находящиеся друг с другом в определенных отношениях.

Положение 10. В интерлигатеоретических отношениях естественных и аксиологических лигатеорий донорная концепция выступает в качестве внешнего, т.е. экстерналистского фактора. Любая естественная или аксиологическая теория не терпит включения в нее чужеродных ей компонентов. Стартуя в дедуктивном отношении от принципа наименьшего действия, физическая теория не допускает в свои ряды технические концепты, природа которых определяется принципом эффективности. Соответственно, техническая теория с ее принципами эффективности и надежности "не терпит" физические концепты.

Положение 11. Формальные теории в интерлигатеоретических отношениях выступают в качестве не экстерналистского, а потенциально интерналистского фактора. Как отмечалось в Положении 4, формальные теории выражают изоморфные отношения между теориями. При этом они сохраняют теснейшее родственное отношение со своими базовыми теориями. Приведем на этот счет показательный пример. Если физики, развивая аппарат квантовой механики, используют достижения векторной алгебры, с которыми они знакомы хуже, чем математики, то это не означает, что они привносят в физику нечто чуждое ей. Выражаясь несколько неуклюже, можно сказать, что они поручили математикам развить аппарат векторного исчисления, необходимый им для действий над векторами, а затем с благодарностью воспользовались их наработками.

Положение 12. Протеория должна доводиться до стадии метатеории. Далеко не сразу ученые осознали, что следует различать два уровня теории – низший и высший. Высший уровень теории принято называть метатеорией. Греческое μετά означает после. Метатеория является результатом тщательного изучения концептуального и методологического устройства исходной теории. Примерами метатеорий являются математическая теория доказательств Д. Гильберта, физическая программа относительности А. Эйнштейна, концепция парадигм программирования Р. Флойда. В нашем понимании сердцевину любого метатеоретического исследования составляют интратеоретическая и интертеоретическая трансдукции.

Зафиксировав метауровень теории, ученые забыли назвать определенным образом ее исходный уровень. На наш взгляд, его уместно называть протеорией. Греческая приставка πρῶ означает в данном случае как концептуальное предшествование протеории метатеории, так и ее устремленность к последней. Разумеется, не следует отождествлять протеорию с прототеорией. Прототеория по определению не достигла стадии научного познания.

Положение 13. Философия науки представляет собой формальную метанауку. Здесь мы вынуждены отметить, что многие исследователи некритически используют понятие философии науки. Так, широко используется концепт «философия физики». Слово «философия» наводит на вывод, что философия физики является уделом философов. В связи с этим А. Эйнштейн вполне правомерно отмечал, что физик «не может просто уступить философу право критического рассмотрения теоретических основ; он, безусловно, лучше знает и чувствует в чем слабые стороны этой основы» [10, с. 200]. Разумеется, изучение устройства физики – это удел физиков, а не философов. Точно так же изучение устройства информатики является уделом информатиков, а не философов. Информатики имеют дело с метаинформатикой, математики – с метаматематикой, физики – с метанаучной физикой (к сожалению, слово «метафизика» в силу некоторых исторических обстоятельств забронировано философами, которые обозначают им учение о первых принципах всякого бытия). Таким образом, термины философия математики, философия физики, философия техники, философия информатики, философия экономики и подобные им в части использования слова «философия» вызывают иллюзии. От них следовало бы отказаться. Впрочем, очевидно, что в силу инерционности литературных норм этого не произойдет. Тем не менее, необходимо понимать, что не следует отождествлять метатеории с философией науки.

Когда же пробивает час философии науки? Неужели автор, будучи профессиональным философом, намерен отказаться от философии науки? Нет, конечно. Час философии науки наступает тогда, когда выделяются изоморфные отношения между метатеориями. Утверждение, что во всех теориях используется концепция интратеоретической трансдукции является философским. Химик занимается метахимией, информатик – метаинформатикой, экономист – метаэкономикой, а философ определяет присущие метатеориям типы эквивалентности. Подобно всем формальным наукам философия науки, выражаясь языком Е. Вигнера, непостижимо эффективна. Она эффективна постольку, поскольку для всех наук является потенциально интерналистским фактором.

Итак, мы представили основные положения метанауки и философии науки. Далее они будут использованы в связи с определением статуса информатики и смежных с ней отраслей наук.

ОПРЕДЕЛЕНИЯ НАУКИ, ОТРАСЛИ НАУКИ, МАТЕМАТИКИ И ТЕХНИЧЕСКОЙ НАУКИ

Несколько лет тому назад один из ведущих современных знатоков оснований информатики П. Деннинг заметил, что «информатика переживает период ренессанса, поскольку она заново открывает свои научные основания» [11, с. 30]. Следует отметить, что вопрос о научном статусе информатики (англ. *computer science*) всегда вызывал у исследователей значительные затруднения. Мало кто сомневался в том, что развитие информатики стало резуль-

татом взаимного концептуального обогащения математики и технических наук. Однако было неясно, каким именно образом следует интерпретировать результат этого симбиоза. К тому же и воззрения относительно научного статуса как математики, так и технических наук вызывали большие сомнения. Показательно в связи с этим, что в классификациях [5, 6] отраслям науки противопоставляется технология (англ. *technology*). А это означает, что она не признается полноправной отраслью науки. В русском языке часто, причем относительно беззаботно используется выражение «технические науки». Но не реже используется и выражение «наука и техника». Так, говорят о «достижениях науки и техники» и, соответственно, о «научно-техническом прогрессе». К этому приходится добавить, что и концепт науки весьма своеобразно понимается в англоязычной культуре, где под наукой, как правило, имеется в виду совокупность наук о природе (англ. *natural sciences*). Если речь заходит о социальных науках, то англичане и американцы обычно прибегают к различного рода оговоркам. Они, дескать, в отличие от естественных наук субъективны и широко используют качественные методы. В связи с этим возникают сомнения относительно их подлинной научности. С учетом выделенных затруднений обратимся сначала к определению науки, затем отрасли науки – математики, технической науки и, наконец, информатики.

Определение науки. Наука – это совокупность всех тех теорий, для которых характерны научные методы, интратеоретические и интертеоретические. В связи с этим необходимо определить статус научной теории. Обозначим четыре интратеоретических метода естественных и аксиологических теорий как 4Д-комплекс, ибо речь идет о четырех дукциях, а именно – дедукции, аддукции, индукции и абдукции. Для интертеоретических отношений также характерны дукции, их всего три – проблематизация, открытие и интерпретация. Чтобы не отождествлять интертеоретические дукции с интратеоретическими, обозначим их как $Д^{+}$, а три метода интертеоретической трансдукции – как $3Д^{+}$-комплекс. Итак, естественная или аксиологическая теория признается научной, если для нее характерен 4Д-$3Д^{+}$-методологический комплекс. Но как же обстоят дела с формальными теориями, которые руководствуются всего одним интратеоретическим методом, а именно – дедукцией? Для них характерен 1Д-$3Д^{+}$-методологический комплекс. Правомерно ли считать их научными концепциями? Надо полагать, правомерно. Два аргумента призваны поддержать эту позицию. Во-первых, реализуются систематические переходы между формальными и неформальными теориями, т. е. они не чужды, а родственны друг другу. Во-вторых, формальные теории подобно другим теориям подчиняются $3Д^{+}$-комплексу.

Сравним нашу точку зрения с позицией П. Деннинга. Он выделяет семь идеалов науки [11, с. 32]. Это, во-первых, организация науки ради понимания, использования и освоения широко распространенных явлений; во-вторых, наука имеет дело как с естественными, так и с искусственными явлениями; в-третьих, наука представляет собой кодифицированный структурный объем знаний; в-четвертых, наука обнаруживает приверженность к экспериментальным методам обоснования открытий и валидации (соответствия используемых методик поставленным задачам – *В.К.*); в-пятых, идеалом науки является воспроизводимость результатов; в-шестых, идеалами науки признаются фальсифицируемость гипотез и моделей; в-седьмых, способность науки делать прогнозы, часть из которых является. По сути, в поле зрения П. Деннинга из методов попали только дедукция, посредством которой осуществляются прогнозы, и экспериментальный метод (аддукция). Все остальные так называемые идеалы сами нуждаются в объяснении посредством методов.

Определение отрасли науки. Четкое определение понятия отрасли науки едва ли можно дать, ибо оно не вводится на строгой научной основе. Впрочем, очевидно, что речь должна идти о совокупности таких теорий, которые обладают определенным родовым (англ. *generic*) признаком. Так, совокупность теорий, для которых характерен принцип максимизации нормы прибыли на авансированный капитал, зачисляется в состав экономики как отрасли науки.

Определение математики. Все математические теории являются исчислениями, то же характерно и для логики. Многократно предпринимавшиеся попытки свести либо математику к логике, либо логику к математике закончились провалом. Это обстоятельство укрепило исследователей во мнении, что логика и математика являются различными отраслями науки. Мы уже отмечали, что, вопреки широко распространенному мнению, математика оперирует отношениями эквивалентности, а не абстрактными объектами.

Определение технической науки (англ. *tech science*). Все технические науки руководствуются принципами эффективности, производительности, безопасности и надежности. Мы с уверенностью провозглашаем научный характер огромного числа технических теорий, относящихся, в частности, к радиотехнике, электротехнике и электронике постольку, поскольку все они руководствуются 4Д-$3Д^{+}$-методологическим комплексом. Сомнения в научном характере технических концепций обосновывают в основном двояко, они, дескать, во-первых, имеют не знаниевый, а практический характер и во-вторых, концентрируются на производстве артефактов [12]. Эти сомнения в свете метанауки и философии науки безосновательны. Во всех естественных и аксиологических науках имеет место этап фактуализации. В социальных и технических науках этот этап называется практикой. Но это ничто иное, как фактуализация. Практический характер технических наук никак не свидетельствует об их хотя бы малейшей отчужденности от науки. К тому же следует учитывать, что практика есть этап теории. Его неправомерно противопоставлять теории. Аргумент, что именно техника сводится к производству особых артефактов, объектов, так же бьет мимо цели. Любая естественная теория имеет дело с объектами, а аксиологическая концепция – с субъектами. В результате и в физике, и в химии, и в биологии производят новые объекты, а в аксиологических науках производят даже субъектов. Технические науки имеют аксиологический характер. Они производят новых людей, а технические артефакты являются не более, чем их воплощениями.

ОПРЕДЕЛЕНИЕ СТАТУСА ИНФОРМАТИКИ

Является ли информатика математикой? Как известно, первый этап расцвета информатики был инициирован успехами теории алгоритмов, которые привели к изобретению λ-исчисления А. Чёрча и машины А. Тьюринга, которая неудачно была названа абстрактной. Однако она не является абстрактной, ибо имеет дело с изоморфными отношениями. Выяснилось принципиально новое обстоятельство. Оказалось, что можно осуществлять алгоритмическое решение задач посредством последовательности команд для машины Тьюринга, которая является не техническим устройством, а дедуктивным конструктивным построением. Впрочем, его можно воплотить в технические устройства, каковыми являются, например, компьютеры. С позиций конструктивной математики программирование является вроде бы довольно шаблонной операцией. Но означает ли это, что информатика является разновидностью конструктивной математики? Нет не означает. Дело в том, что в информатике непременно предполагается исполнение программ. А этому требованию не подчиняется ни одна математическая теория. Математики обращаются к достижениям информатики в том же порядке, что и представители всех других наук. Одно это весьма показательно демонстрирует отличие информатики от математики.

Является ли информатика формальной наукой? Да, является. Об этом свидетельствует, в частности, то обстоятельство, что один и тот же язык программирования и его библиотека оказываются актуальными для различных наук. Подобно ситуации с логикой и математикой действие информатики распространяется на все другие науки. Это отчетливый признак именно формальной науки.

Является ли информатика наукой? Да, является, ибо подобно математике и логике руководствуется 1Д-3Д$^{+}$-методологическим комплексом. В соответствии с ним обеспечивается прогресс информационного знания. Благодаря ему весьма действенно так называемое компьютерное моделирование, столь актуальное для всех наук.

Является ли математика технической наукой? Нет, не является. Этот вывод нуждается в объяснении. Во всех известных технических науках характерные для них артефакты фигурируют в качестве изначальных объектов и, следовательно, концептов, ибо объекты являются представлениями соответствующих теорий (см. Положение 2). Показательный пример: атомная энергетика является теорией АЭС, которые производят полезную людям электрическую энергию. Обратимся теперь к информатике. Здесь процессоры в качестве главных составляющих компьютеров являются не изначальным, а конечным звеном конструктивных построений информатиков. Эти построения они начинают с принципов вычисления, затем переходят к парадигмам программирования и к программам. И только после этого наступает очередь процессоров, которые оказываются заключительным звеном конструктивных формальных построений. В качестве таковых процессоры не выводят за их границы. Продолжая приводимое объяснение, еще раз сошлемся на атомную энергетику. В ней широко используются температурные характеристики. Естественно, возникает вопрос об их принадлежности. Является ли, например, температура теплоносителя физической или же технической характеристикой? Если указанная температура объясняется принципами физики, то она должна быть признана физическим параметром. Если же рассматриваемая температура объясняется принципами технической теории, то она является технической характеристикой.

Значение концептов определяется их принадлежностью к некоторой теории. Это правило актуально в случае определения статуса процессоров и, соответственно, компьютеров. Если компьютер рассматривается как электронное устройство, то его природа объясняется посредством технической теории, а именно – электроники. Если же компьютер рассматривается в контексте устройства информатики, то он является не техническим, а информационным устройством. Компьютеры часто называют электронно-вычислительными устройствами. Такое словоупотребление способно ввести в заблуждение, ибо создается впечатление, что есть электронно-вычислительная теория, т. е. своеобразный гибрид технической теории и информатики. Но подлинная наука не знает таких гибридов. Безусловно, существуют интертеоретические отношения между информатикой, с одной стороны, и электроникой – с другой. Но при этом они не сливаются в нечто единое.

После всего изложенного можно отважиться на определение информатики. *Информатика – это формальная отрасль науки, состоящая из теорий, в каждой из которых совершается конструктивный концептуальный переход: принципы алгоритмических вычислений – парадигмы программирования – программы – исполнение программ посредством компьютеров.* При желании каждому этапу концептуальной информационной трансдукции может быть дана детальная характеристика. Наша задача состояла в выделении основного тренда указанной трансдукции, не более того. В противном случае определение информатики не могло бы быть лаконичным и растянулось бы на сотни страниц.

Предлагая некоторое определение информатики, естественно, необходимо иметь в виду его альтернативы. В связи с этим наше внимание привлекла новейшая статья трех весьма известных авторов "Неправильные концепции информатики" [13]. Судя о содержании статьи по названию, следовало ожидать, что ее авторы представят от своего имени правильную концепцию информатики, а затем подвергнут обоснованной критике отклонения от нее. В действительности же, провозгласив информатику наукой, они избегают ее определения. Характеристика информатики проводится исключительно в апофатическом стиле, посредством отрицания всех ее недостаточных, частичных определений. Информатика, утверждают они, это не только программирование, не только разрешение проблем посредством языковых нотаций, не только компьютерное мышление, не только единственный фундаментальный тип мышления, не только естественная наука и математика, но и инженерия; она предполагает знание устройства тех отраслей наук, в которых используется, и учет исто-

рии их развития. В подзаголовке статьи авторы правомерно утверждают, что широко распространенные неправильные концепции информатики затрудняют профессиональный рост исследователей. С этим следует согласиться. Но хотелось бы знать, какой именно наукой является информатика, в частности, естественной, аксиологической, формальной или универсальной.

Вопрос об универсальном характере информатики возник не случайно. Замечено, что представители всех формальных наук, в частности, философы, лингвисты, логики, математики, информатики, демонстрируют свою приверженность установке на универсальную науку в большей степени, чем их коллеги из других областей знания. На наш взгляд, этот феномен объясняется тем обстоятельством, что представители всех естественных и аксиологических отраслей науки не могут обойтись без использования достижений формальных наук. В связи с этим может возникнуть представление, что именно формальные науки имеют универсальный характер. Но в действительности, формальные науки не в состоянии подменить собой все другие науки.

ЗАКЛЮЧЕНИЕ

Бурное развитие информатики предъявляет новые требования к ее осмыслению. Она стала полноправным членом современного научного знания, того огромного целого, которое образует неисчислимое множество лигатеорий. Статус информатики не может быть осмыслен достаточно основательно без должного понимания этого целого. Оно же, в свою очередь, не может быть достигнуто без обращения непосредственно к метанауке и философии науки. К сожалению, приходится отмечать, что многие исследователи, выступающие от имени информатики, крайне робко осваивают ее метауровень. Создание последовательного курса метаинформатики, которую необоснованно называют философией информатики, стало насущной задачей. С другой стороны, ощущается острая потребность в создании последовательной философии науки, от имени которой ученым предлагались бы не устаревшие рецепты неопозитивизма, критического рационализма и аналитической философии, а заключения, которые учитывали бы достижения всех современных наук. К сожалению, и на этом фронте, на этот раз среди философов, наблюдается робость и затишье, вместо энтузиазма и натиска. Нам остается выразить надежду, что даже в чем-то предварительное понимание значимости метанауки и философии науки открывает новые пути к осмыслению настоящего и будущего информатики.

СПИСОК ЛИТЕРАТУРЫ

1. Канке В.А. История, философия и методология техники и информатики. – М.: Юрайт, 2013. – 409 с.
2. Канке В.А. Философские проблемы науки и техники. – М.: Юрайт, 2015. – 288 с.
3. Канке В.А. Взлеты и падения гениев науки: практикум по методологии науки. – М.: Инфра-М, 2017. – 190 с.
4. Proposed International Standard Nomenclature for Fields of Science and Technology // United Nations Educational, Scientific and Cultural Organization. UNESCO/NS/ROU/257 rev. 1 Paris, le 5 décembre 1988. Original: English. – URL: // http://unesdoc.unesco.org/images/0008/000829/082946EB.pdf/ (дата обращения 05.04.2017)
5. Revised Field of Science and Technology (FOS) Classification in the Frascati Manual. – URL: // http://www.oecd.org/science/inno/38235147.pdf/, 2007(дата обращения 05.04.2017).
6. Паспорта научных специальностей – ВАК. – URL: // vak.ed.gov.ru/316 (дата обращения 05.04.2017)
7. Фреге Г. Логико-философские труды. – Новосибирск: Сибирское университетское изд-во, 2008. – 283 с.
8. Вигнер Е. Непостижимая эффективность математики в естественных науках // Успехи физических наук. – 1968. – Т. 94, Вып. 3. – С. 535–546.
9. Пирс Ч.С. Логические основания теории знаков. – СПб.: Лаборатория метафизических исследований при философском факультете СПбГУ, изд-во "Алетейя", 2000. – 352 с.
10. Эйнштейн А. Собрание научных трудов. В 4-х т. Т. 4. – М.: Наука, 1967. – 600 с.
11. Denning P. The Science in Computer Science // Communications of the ACM. – 2013. – Vol. 56, № 5. – P. 30–33.
12. Franssen M., Lokhorst G.-J.,de Poel van I. Philosophy of Technology // The Stanford Encyclopedia of Philosophy / ed. E.N. Zalta. – URL: // https://plato.stanford.edu/archives/fall2015/entries/technology/ (дата обращения 05.04.2017).
13. Denning P.J., Tedre M., Yongpradit P. Misconceptions about Computer Science // Communications of the ACM. – 2017. – Vol. 60, № 3. – P.31–33.

Материал поступил в редакцию 07.04.17.

Сведения об авторе

КАНКЕ Виктор Андреевич – доктор философских наук, профессор кафедры философии НИЯУ МИФИ
e-mail: kanke@obninsk.ru

ИНФОРМАЦИОННЫЙ АНАЛИЗ

УДК 025.4.025 : [004.627 : 81’322.2]

И.В. Селиванова, Б.Я. Рябко, А.Е. Гуськов

Классификация посредством компрессии: применение методов теории информации для определения тематики научных текстов

Предложен метод автоматической классификации научных текстов, основанный на использовании кодирования источников информации (или «сжатия данных»). Метод реализован и исследован на данных, представленных в Архиве научных текстов (arXiv.org), а также в научной электронной библиотеке «Киберленинка» (cyberleninka.ru). Эксперименты показали, что с вероятностью 75-95% метод правильно определяет тематику текстов, при этом его точность зависит от качества исходных данных.

Ключевые слова: *классификация, тематическая классификация текстов, теория информации, компрессия текстов, arXiv.org, cyberleninka*

ВВЕДЕНИЕ

Задача автоматической, т.е. не требующей участия человека, классификации научных текстов (статей, книг и т.п.) представляет большой практический интерес, так как последние десятилетия характеризуются лавинообразным увеличением количества научных публикаций. Во многих областях науки ежегодно появляется так много новых научных текстов, что каждый специалист не может ознакомиться с содержанием их всех. В этих условиях информационная поддержка исследований приобретает особую важность и требует проведения предварительной классификации вновь появляющихся научных публикаций с целью выявления тех из них, которые представляют интерес для конкретного ученого. Существующие методы классификации научных текстов используют труд экспертов, они дороги и малоприменимы, поэтому задача разработки автоматических методов весьма актуальна и привлекает внимание лингвистов, специалистов в области искусственного интеллекта и в таких сравнительно новых дисциплинах, как машинное обучение (*machine learning*), определение значения данных (*data mining*) и анализ Больших Данных (*Big Data*).

Несмотря на многочисленные исследования, задача построения эффективных методов классификации научных текстов пока еще далека от решения. Рассмотрим подробнее, как исследователи подходят к решению этой задачи. Одним из самых популярных для классификации научных текстов является метод построения частотных векторов встречаемости слов [1, 2], а также его модификации. Например, в [3] в качестве компонента вектора встречаемости слов предлагается рассматривать не отдельно взятое слово, а N-грамму.

Такая технология позволяет использовать различные методы: в некоторых случаях между векторами рассчитывается мера близости, а затем применяется классификация, основанная на теории графов [4]; часто применяется наивная байесовская классификация [5], строятся деревья решений [6], нейронные сети [7] *k*-ближайших соседей [8] и др.

В настоящей статье описывается разработанный нами метод автоматической классификации научных текстов, основанный на методах теории информации. Основная идея довольно проста и естественна – в текстах, относящихся к одной области науки, используется много общих понятий, терминов и оборотов, причем, чем уже рассматриваемая научная область, тем “ближе” лексика текстов, к ней относящихся. Это очевидное наблюдение широко используется во многих методах классификации текстов (например, выделение ключевых слов и фраз и т.п.), однако, в отличие от других методов, мы предлагаем оценивать степень “лексической близости” при помощи методов компрессии – сжатия текстов так называемыми “архиваторами” (*archiver*, *data comressor*). Ранее близкий подход применялся при решении задач классификации и определения авторства текстов [9–17]. Однако для классификации научных текстов этот метод не использовался.

ОПИСАНИЕ МЕТОДА

Пусть есть n научных областей: $X_1, X_2, \ldots, X_n$. Для каждой из этих областей определено *ядро* – множество текстов, типичных для данной категории (далее –

обучающая выборка). Обозначим за x_1^1, x_2^1, …, $x_{m_1}^1$ – ядро для первой категории, x_1^2, x_2^2, …, $x_{m_2}^2$ – для второй, x_1^n, x_1^n, …, $x_{m_n}^n$ – для n-й. Пусть дан научный текст y, который относится к одной из этих областей. Предлагаемый метод на основе обучающих выборок должен определить, к какой именно области относится текст y.

Для решения этой задачи мы будем использовать архиватор φ, который может быть применен для сжатия любого множества текстов. Обозначим через K(u_1, ..., u_m) длину "сжатых" методом φ текстов u_1, ..., u_m, представленных в виде компьютерных файлов. Рассмотрим теперь ситуацию, когда некоторый текстовый файл u «сжимается» вместе с u_1, ..., u_m и вычисляется длина сжатых текстов K(u_1, ..., u_m, u). Отметим сразу, что порядок текстов u_1, ..., u_m, u, сжимаемых архиватором, важен – u должен «сжиматься» после u_1, ..., u_m. Понятно, что величина K(u_1, ..., u_m, u) будет больше чем K(u_1, ..., u_m), так как в первом случае количество сжимаемых файлов больше – добавлен u. Обозначим разность между длиной закодированных файлов u_1, ..., u_m, u и u_1, ..., u_m через K(u /u_1, ..., u_m):

$$K(u/u_1, ..., u_m) = K(u_1, ..., u_m, u) - K(u_1, ..., u_m). \quad (1)$$

Здесь файл u кодируется после u_1, ..., u_m и при его «сжатии» архиватор использует сведения о частотах букв и слов, а также о других закономерностях и особенностях в текстовых файлах u_1, ..., u_m. Так как величина K(u /u_1, ..., u_m) зависит от файлов u_1, ..., u_m, то она будет тем меньше, чем больше файл u «похож» на u_1, ..., u_m. Идея метода достаточно проста – относить текстовый файл u, принадлежность которого неизвестна, к той группе текстов, с которой он лучше всего «сжимается». Более формально предлагаемый метод выглядит следующим образом: для текстов из каждой научной области X_i, $i=1,…n$ вычисляем величины K(y/ x_1^i, x_2^i, …, $x_{m_i}^i$) и выбираем j такое, что K(y/ x_1^j, x_2^j, …, $x_{m_j}^j$) имеет минимальное значение. Тогда считаем, что текст y принадлежит области X_j. Формально

$$K(y/x_1^j, x_2^j, …, x_{m_j}^j) = \min_{i=1,…,n} K(y/x_1^i, x_2^i, …, x_{m_i}^i). \quad (2)$$

ИСПОЛЬЗУЕМЫЕ ДАННЫЕ И СХЕМА ЭКСПЕРИМЕНТА

Общее описание данных

Предлагаемый метод был реализован и исследован на данных, представленных в двух репозиториях: англоязычном Архиве научных текстов (arXiv.org) и в научной электронной библиотеке «Киберленинка» (cyberleninka.ru).

Веб-сайт arXiv.org содержит архив научных текстов по физике, математике, биологии, статистике, компьютерным наукам и финансовой математике. Для каждого из этих разделов определены области науки (например, физика разбивается на астрофизику, физику ускорителей, атомную физику, ядерную и др.). При размещении текста на сайте автор указывает одну или несколько областей, к которым принадлежит его работа. Таким образом, все тексты, помещенные в архив, классифицированы (т.е. отнесены к направлениям и областям) самими авторами, что даёт возможность проверить качество работы нашего метода. Кроме того, размещенные в arXiv.org тексты общедоступны, представлены в стандартизированной форме и практически все на одном языке (английском), что упрощает их обработку. Первую область науки, указываемую автором, будем называть главной, а остальные – второстепенными. Для эксперимента мы выбрали 30 областей науки, принадлежащих физике, математике, биологии или компьютерным наукам. Следует отметить, что одна область науки может относиться к нескольким научным направлениям: например, математическая физика относится как к физике, так и к математике, а теория информации относится к математике и компьютерным наукам.

Классификация статей в «Киберленинке» основана на Государственном рубрикаторе научно-технической информации (ГРНТИ). Из более 80 категорий мы выбрали 20: астрономия, автоматика, биология, экология, экономика, философия, физика, геофизика, география, химия, история, кибернетика, литература, математика, медицина, механика, политика, психология, социология, юридические науки). Это определялось тем, что в указанных категориях было более 300 файлов на русском языке, что позволило провести однородные эксперименты с полученными данными.

Обработка данных

Полные тексты статей были получены в формате .pdf, после чего с помощью программы PDF2Text Pilot (http://www.colorpilot.com/pdf2text.html) из каждого из них был извлечен текстовый слой, а затем были удалены знаки препинания, цифры, символы, которые появились после преобразования формул, а также стоп-слова; для русскоязычных текстов из «Киберленинки» был проведен стэмминг – процесс приведения слов к начальной форме, используя Яндекс mystem (https://tech.yandex.ru/mystem/?ncrnd=4660), и удалены файлы размером менее 10 кб.

Тест на динамику сжатия

Для проверки работы метода определим тест на динамику сжатия (ТДС). Возьмем ядро одной из областей науки и разобьем его на группы файлов в ядре: 0, 1, 2, 3 и т.д., причем каждая предыдущая группа стала подгруппой следующей.

Далее возьмем тестовый файл и последовательно начнем сжимать его с каждой из выделенных групп. В итоге рассмотрим зависимость между количеством файлов в ядре и долей сжатия (наименьшая доля сжатия соответствует наилучшему сжатию). Для правильной работы метода ожидается увидеть убывающую кривую, т.е. чем меньше файлов в ядре, тем доля сжатия больше. Но при наполнении словаря необходимым количеством терминов скорость убывания кривой должна уменьшаться.

Выбор архиватора

Мы отобрали архиватор, наиболее подходящий для наших целей. Для этого с использованием теста на динамику сжатия была проведена серия экспериментов с находящимися в открытом доступе архиваторами WinRAR (http://www.win-rar.ru), 7z (http://www.7-zip.org), PeaZIP (http://www.peazip.org). Известно, что в этих архиваторах применялись разные алгоритмы (PPMd; PPMd и LZMA; Deflate и BWT, соответственно). Все архиваторы рассматривались при различных значениях параметров. По результатам экспериментов был отобран архиватор WinRAR при максимальном значении памяти 128 Мбайт (память – параметр архиватора), и в дальнейшем он использовался при проведении всех экспериментов.

Исследование свойств метода

Опишем эксперимент, показывающий, что предлагаемый метод позволяет правильно определить научное направление, к которому относится текст. Возьмем ядра из 4-х научных направлений Архива научных текстов – arXiv.org: физики, биологии, математики, компьютерных наук; и тестовый файл из направления физика, после чего проведем тест на динамику сжатия.

На рис. 1 показано, что кривая, соответствующая сжатию тестового файла из области физики с ядром из той же области, идет ниже остальных, т.е. тест сжимается лучше всего с направлением, которому он соответствует.

Теперь покажем, что предлагаемый метод работает и для более узких дисциплин. Возьмем тестовый файл из физического направления astro-ph.GA (рис. 2) и проведем тест для других физических направлений. Получаем, что кривая, соответствующая тесту на динамику сжатия тестового файла категории astro-ph.GA и ядра astro-ph.GA, идет ниже остальных. Также видно, что чуть выше идет кривая, соответствующая другой области астрофизики, т. е. метод правильно выделяет область и из двух терминологически близких дисциплин.

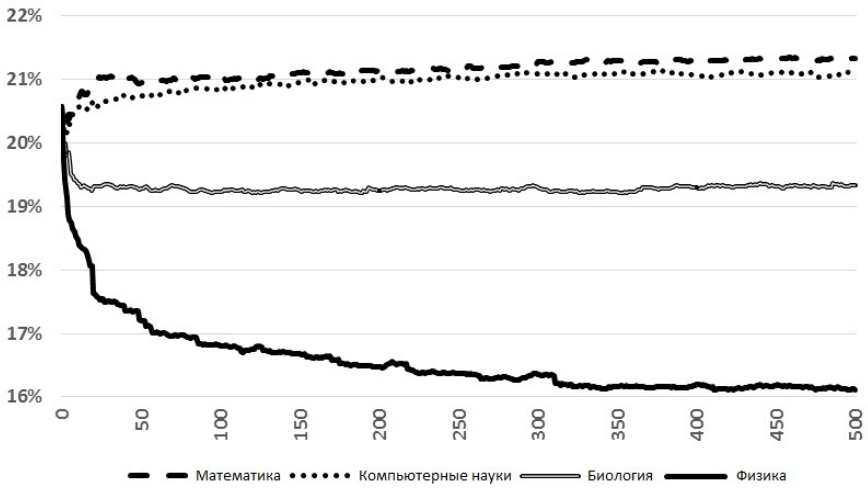


Рис. 1. Зависимость степени сжатия теста из направления «Физика» от количества файлов в ядре и других направлений

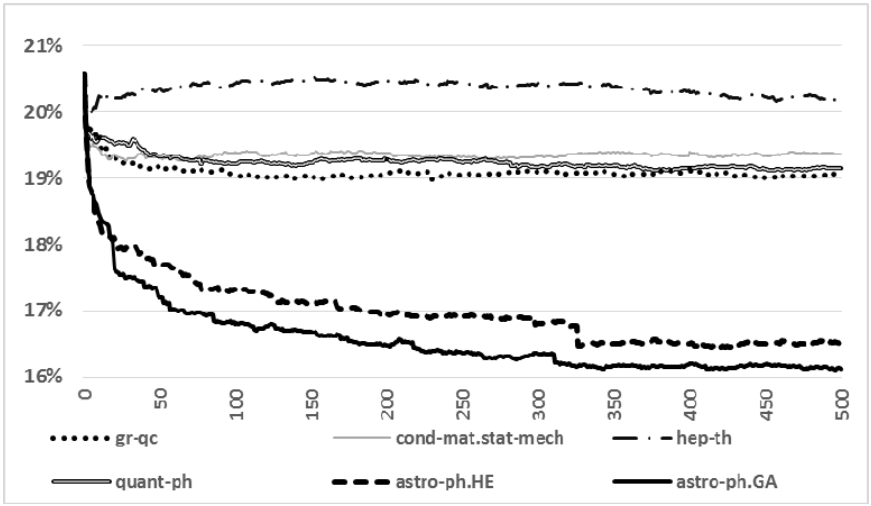


Рис. 2. Зависимость степени сжатия теста из области «astro-ph.GA» от количества файлов в ядре и других физических направлений

Объем обучающей выборки

Степень сжатия текста зависит от размера ядра, что показано на рис. 1 и 2. Далее мы определим число текстов в обучающих выборках, т.е. значение параметра m_i в $x_1^i, x_2^i, ..., x_{m_i}^i$ для $X_1, X_2, ..., X_n$. (см. описание метода). Зависимость количества ошибок от количества файлов в ядре показана на рис. 3. Всего проверка осуществлялась на 450 тестах, т.е. на 15 тестовых файлах каждой категории, не входящих в ядра.

Мы видим, что качество классификации существенно зависит от количества файлов, находящихся в ядре. Причем, при размере обучающей выборки в несколько десятков файлов – ошибка большая, при объеме же ядра в 300-500 файлов дальнейшее увеличение выборки особо не влияет на качество. Поэтому мы рекомендуем выбирать ядро размером не менее 100 текстов, но если мощность вычислительных ресурсов позволяет быстро работать, то его можно увеличить до 1000.

С учетом полученных нами результатов для классификации текстов в дальнейшем будем использовать ядро, состоящее из 100 файлов.

Оптимизация состава обучающей выборки

Мы исследовали зависимость точности классификации от количества файлов, входящих в ядро. Ряд проведенных экспериментов показал, что в зависимости от того, какие файлы находятся в ядре, меняется количество правильно и неправильно определенных тестовых файлов. Оптимальным является такое ядро, которое наиболее полно показывает содержание предметных областей.

Далее мы рассмотрим как состав данных, входящих в обучающую выборку, влияет на вероятность ошибок, и опишем метод ядра, позволяющий уменьшить число неверно определенных тестовых файлов.

Интуитивно понятно, что файлы, включаемые в ядро, неполноценны по своей информативности, и поэтому при произвольном выборе в ядро могут попасть как «наиболее типичные», так и, наоборот, «нетипичные» для данной области тексты.

Для уменьшения вероятности ошибок, возникающих при классификации, был разработан следующий алгоритм, позволяющий формировать «ядро» из «типичных» файлов.

Пусть файлы *A*, *B*, …, *E* принадлежат одной категории. Возьмем файл *A* и начнем последовательно сжимать его с файлами *B*, …, *E*. Долю сжатия будем указывать на пересечении строк и столбцов. Получим матрицу сжатия файлов одной категории, у которой по строкам указано как файл *A* сжимается с каждым из остальных файлов, по столбцам – как другие файлы сжимаются с файлом *A*.

Так как чем меньше доля сжатия, тем лучше, посчитаем среднюю долю сжатия по каждому столбцу и отсортируем получившиеся величины по возрастанию. Далее, удалим из строк и столбцов файл с наименьшей средней долей (этот файл уже точно войдет в ядро), после чего проведем процедуру заново. В итоге, в ядро попадут те 100 файлов, которые в этих «рейтингах» будут в самом начале и с которыми другие файлы этой категории сжимаются лучше всего. Иными словами, в ядре окажутся тексты, несущие наибольший объем информации для других файлов этой категории. Как показали тесты, такой подбор ядра позволяет заметно улучшить результаты классификации.

Следует отметить, что в некоторых случаях (например, если изначальная классификация данных вызывает сомнения) обучающую выборку стоит формировать вручную.

Рис. 3. Зависимость количества ошибок от количества файлов в ядре

РЕЗУЛЬТАТЫ ИССЛЕДОВАНИЯ

Результаты применения рассматриваемого нами метода классификации научных текстов были получены в виде матрицы (табл. 1) сжатия тестовых файлов с файлами каждой из категорий, нормированной по доле сжатия, т.е. из каждой строки матрицы была вычтена минимальная доля сжатия. В итоге, нулевая доля сжатия обозначала, к какой категории относится тест. В случае test2, где в качестве главной категории указана cond-mat.stat-mech, а второстепенной – math-ph**,** видно, что метод определил обе категории правильно (0% сжатия соответствует cond-mat.stat-mech; 0,02% – math-ph).

Такое представление удобно в случае, если статья является междисциплинарной. В табл.1 этот случай представляет test3 (доли сжатия с категориями cond-mat.stat-mech и gr-qc различаются только на 0,01%).

Рассмотрим отдельно результаты, полученные для «Киберленинки» и Архива научных текстов – arXiv.org. В экспериментах с текстами из архива «Киберленинки» произвольным образом были отобраны 400 тестовых файлов, по 20 для каждой категории. В результате при произвольном выборе ядра неправильно было определено 47% тестов. При более детальном изучении текстов, полученных из «Киберленинки», было выявлено, что изначально в некоторых категориях встречались тексты, относящиеся как к близким категориям, так и к совершенно другим областям науки. Например, в категории «Математика» были обнаружены статьи из категорий «История», «Социология» и др. Таким образом, изначально большую долю ошибок можно объяснить тем, что при произвольном выборе в ядра категорий попало большое количество текстов из других областей науки.

При ядрах, подобранных методом, описанным в разделе «Оптимизация состава обучающей выборки», количество ошибок сократилось более чем в 1,5 раза (табл. 2). В двух категориях («История» и «Литература») все тесты были определены полностью правильно. В остальных же случаях чаще всего определялась близкая категория: например, тесты «Математики» были отнесены к «Автоматике» или «Кибернетике», тесты «Геофизики» – к «Географии», тесты «Истории» – к «Политике» или «Юридическим наукам».

Иная ситуация наблюдалась с классификацией на arXiv.org. Для экспериментов было выбрано 600 тестовых файлов (по 20 из каждой категории). Но, в отличие от «Киберленинки», уже при произвольных ядрах количество ошибочно определенных тестов составляло лишь 11% (табл. 3).

Более того, ошибки в классификации текстов из arXiv.org можно разделить на следующие типы:

1) определяется второстепенная категория вместо главной;

2) определяется другая категория из научного направления;

3) определяется другое научное направление.

Таким образом количество ошибок сократилось примерно на 3%. Второстепенная категория вместо главной определяется в 3%; другая категория – в 4%. Другое научное направление определяется лишь в 1%, причем, например, в научной области cs.CR определяется math.PR, но в качестве второстепенной категории у этого теста указана другая математическая категория, которой нет в исследуемом списке.

Таблица 1

Матрица сжатия тестовых файлов категории «cond-mat.stat-mech» с другими категориями

Категория	Тестовый файл		
	test1 (cond-mat.stat-mech)	test2 (cond-mat.stat-mech, math-ph)	test3 (cond-mat.stat-mech)
cond-mat.stat-mech	**0%**	**0%**	**0%**
CS.IT	1,48	2,46	1,60
CS.LO	2,73	3,63	2,63
gr-qc	0,44	1,66	**0,01**
math-ph	0,48	**0,02**	1,60
nucl-ex	2,00	2,83	1,56
physics.acc-ph	2,75	3,30	1,71
physics.atom-ph	1,43	2,46	0,98
physics.optics	1,59	2,50	1,03
physics.soc-ph	1,06	1,80	0,74
q-bio.BM	1,18	1,68	0,38
quant-ph	0,76	1,37	0,79

Таблица 2

Результаты классификации научных текстов из архива «Киберленинки»

	Произвольные ядра	**Подобранные ядра**
Автоматика	7	7
Астрономия	4	1
Биология	12	13
Экономика	17	5
Экология	16	6
Философия	8	4
Физика	8	9
Геофизика	7	4
География	8	10
История	2	0
Кибернетика	13	14
Литература	1	0
Математика	13	10
Медицина	7	2
Механика	15	5
Химия	7	2
Политика	5	8
Психология	8	1
Социология	17	9
Юридические науки	12	2
Всего	187 (47%)	114 (28%)

Таблица 3

Результаты классификации научных текстов из arXiv.org *

Направление	**Научная область**	**Произвольные ядра**	**Подобранные ядра**			
		Кол-во ошибок	**Кол-во ошибок**	**Тип 1**	**Тип 2**	**Тип 3**
Физика	astro-ph.CO	2	2	2	0	0
Физика	astro-ph.GA	3	3	1	2	0
Физика	astro-ph.HE	2	2	1	1	0
Физика	cond-mat.dis-nn	2	4	3	1	0
Физика	cond-mat.stat-mech	3	1	0	1	0
Комп. науки	cs.AI	3	6	1	4	1
Комп. науки	cs.CR	0	2	0	1	1
Комп. науки	cs.IT	2	2	1	0	1
Комп. науки	cs.LO	2	0	0	0	0
Комп. науки	cs.SE	2	0	0	0	0
Физика	gr-qc	2	0	0	0	0
Физика	hep-ex	1	4	0	4	0
Физика	hep-th	1	1	1	0	0
Математика	math.AG	0	0	0	0	0
Математика	math.CO	1	1	0	1	0
Математика	math.DG	0	2	0	2	0
Математика	math.FA	0	0	0	0	0
Математика	math.GR	0	0	0	0	0
Математика	math.PR	1	2	1	1	0
Математика	math.ST	0	0	0	0	0
Математика	math-ph	18	4	1	3	0
Физика	nucl-ex	4	0	0	0	0
Физика	nucl-th	1	2	2	0	0

Направление	Научная область	Произвольные ядра	Подобранные ядра			
		Кол-во ошибок	Кол-во ошибок	Тип 1	Тип 2	Тип 3
Физика	physics.acc-ph	0	1	0	1	0
Физика	physics.atom-ph	3	1	0	1	0
Физика	physics.ins-det	7	3	1	2	0
Физика	physics.optics	0	1	1	0	0
Физика	physics.soc-ph	0	0	0	0	0
Биология	q-bio.BM	2	2	0	0	2
Физика	quant-ph	1	1	0	1	0
	Всего	63	47	16	26	5
	Доля, %	11	8	3	4	1

* *Примечание:* Виды ошибок: тип 1 – определяется второстепенная категория вместо главной; тип 2 – определяется другая категория из научного направления; тип 3 – определяется другое научное направление)

ЗАКЛЮЧЕНИЕ

В настоящей работе предлагается метод автоматической классификации научных текстов, точность которого исследуется на примере двух репозиториев: англоязычном Архиве научных текстов arXiv.org и русскоязычной электронной библиотеке «Киберленинка».

Как и следовало ожидать, точность метода зависит от качества «исходной» классификации, используемой для первоначального обучения. В случае текстов из arXiv.org, где изначальная классификация проводится самими авторами и имеет низкую долю ошибок, наш метод работает довольно точно, показывая более чем 90%-ю эффективность

В случае текстов из «Киберленинки» было выявлено, что изначально некоторые статьи классифицированы неправильно: например, в категории «Математика» были обнаружены статьи по истории, социологии и др. Это осложняет формирование обучающих выборок и ухудшает эффективность метода в 2-5 раз.

Описанное нами исследование показало, что для качественных данных доля ошибок не превышает 5-8%, причем метод правильно определяет не только общее научное направление, но и более узкую научную категорию. Достоинством метода является возможность оценивать «близость» текста к другим предметным областям.

Предлагаемый метод может применяться для классификации больших объемов научных текстов при введении новой системы классификаторов или верификации существующей, а также в задачах обнаружения «близких» по содержанию текстов.

СПИСОК ЛИТЕРАТУРЫ

1. Baghel R., Dhir R. A Frequent Concepts Based Document Clustering Algorithm // International Journal of Computer Applications. – 2010. – Vol. 4, № 5. – P. 6 – 12
2. Beil F., Ester M., Xu X. Frequent Term-Based Text Clustering // Proc. 8th Int. Conf. on Knowledge Discovery and Data Mining (KDD ‘2002). – Edmonton, Alberta, Canada, 2002. – P. 436-442.
3. Miao Y., Keselj V., Milios E. Document clustering using character n-grams: a comparative evaluation with term-based and word-based clustering // In CIKM ’05: Proceedings of the 14th ACM international conference on Information and knowledge management. – NY, USA, 2005. – P. 357–358.
4. Schaeffer S.E. Graph clustering // Computer Science Review. – 2007. – Vol.1, №1. – P. 27–64.
5. Kim S., Han K., Rim H., Myaeng S.H. Some effective techniques for naïve bayes text classification.//IEEE Transactions on Knowledge and Data Engineering. – 2006. – Vol. 18, № 11. – P. 1457-1466.
6. Шевелев О.Г., Петраков А.В. Классификация текстов с помощью деревьев решений и нейронных сетей прямого распространения // Вестник Томского государственного университета. – 2006. – Т. 290. – С. 300-307.
7. Wang Z., He Y., Jiang M. A comparison among three neural networks for text classification // In proceedings of the IEEE 8th international conference on Signal Processing. – 2006. – № 3. – P. 1883-1886.
8. Матяско А.А., Хаустов В.А. Классификация документов в векторном пространстве. Сравнение методов Роккио и метода k-ближайших соседей // Информационные технологии и системы 2012 (ИТС 2012) : материалы международной научной конференции (г. Минск, Беларусь, 24 октября 2012 г.) = Information Technologies and Systems 2012 (ITS 2012) : Proceeding of The International Conference, BSUIR, Minsk, 24th October 2012 / ред.кол. : Л. Ю. Шилин и др. – Минск : БГУИР, 2012. – С. 140–141.
9. Li M., Vit´anyi P.M.B. An Introduction to Kolmogorov Complexity and Its Applications. 2nd ed. – New York: Springer-Verlag, 1997. – P. 637.
10. Cilibrasi R., Vitanyi P.M.B. Clustering by Compression // IEEE Transactions on Information Theory. – 2005. – Vol. 51, № 4. – P. 1523-1545.

11. Cilibrasi R., Vitanyi P.M.B., de Wolf R. Algorithmic clustering of music based on string compression // Comp. Music J. – 2004. – Vol. 28, № 4. – P. 49–67.
12. Li M., Chen X., Li X., Ma B., Vitanyi P.M.B. The similarity metric // IEEE Transactions on Information Theory. – 2004. – Vol. 50, № 12. – P. 3250-3264.
13. Кукушкина О.В., Поликарпов А.А., Хмелев Д.В. Определение авторства текста с использованием буквенной и грамматической информации // Пробл. передачи информ. – 2001. – Т. 37, №. 2. – С. 96–109.
14. Хмелёв Д.В. Сложностной подход к задаче определения авторства текста // Труды и материалы Международного конгресса «Русский язык: исторические судьбы и современность» (13-16 марта 2001 года). – М.: МГУ. – 2001. – С. 426-427.
15. Malyutov M.B. Authorship Attribution of texts: a review // Springer Lect. Notes in Comp. Sci. 4123 / eds. R. Ahlswede et al. – 2007. – P. 362–380.
16. Malyutov M.B., Wickramasinghe C.I., Li S. Conditional Complexity of Compression for Authorship Attribution. SFB 649 Discussion Paper No. 57. – Berlin: Humboldt University, 2007. – P. 38
17. Ryabko B., Astola J., Malyutov M. Compression-Based Methods of Statistical Analysis and Prediction of Time Series. – Springer, 2016.

Материал поступил в редакцию 03.02.17.

Сведения об авторах

СЕЛИВАНОВА Ирина Вячеславовна – младший научный сотрудник Государственной публичной научно-технической библиотеки (ГПНТБ) СО РАН, Новосибирск; аспирант Новосибирского государственного университета
e-mail: selivanova@ict.sbras.ru

РЯБКО Борис Яковлевич – доктор технических наук, профессор ГПНТБ СО РАН; главный научный сотрудник Института вычислительных технологий СО РАН, Новосибирск
e-mail: boris@ryabko.net

ГУСЬКОВ Андрей Евгеньевич – кандидат технических наук, директор ГПНТБ СО РАН; старший научный сотрудник Института вычислительных технологий СО РАН, Новосибирск
e-mail: guskov@spsl.nsc.ru

АВТОМАТИЗАЦИЯ ОБРАБОТКИ ТЕКСТА

УДК 81’322’373.42

В.А. Яцко, Т.С. Яцко

Особенности структуры лингвистической онтологии*

Описывается методика разработки лингвистической онтологии как компонента системы автоматического анализа мнений покупателей о коммерческих продуктах. Обосновываются фундаментальные принципы построения онтологий данного типа, к которым относятся: связь онтологии с грамматикой; выделение в её структуре параметрических и оценочных терминов и разделение оценочных терминов на синтаксические и семантические; бинарное отношение между синтаксическими и семантическими терминами; градационная шкала интенсивности оценок. Впервые на материале русского языка анализируются случаи омонимии и синонимии оценочных терминов.

Ключевые слова: *автоматический анализ мнений покупателей, лингвистическая онтология, параметрические термины, оценочные термины, бинарное отношение, омонимия и синонимия*

ВВЕДЕНИЕ

Понятие онтологии, которое первоначально было разработано и применялось в философии, в настоящее время часто используется в информатике для обозначения одного из интенсивно развивающихся направлений автоматической обработки данных. В общем смысле под онтологией понимается сложно структурированный словарь, моделирующий некоторую предметную область и включающий термины с заданными отношениями между ними. Под сложной структурой понимается многоуровневая иерархия, конечными элементами которой выступают конкретные термины (инстанциации), а на верхних уровнях находятся термины, отражающие структуру предметной области. Отношения между терминами описываются аксиомами [1, с. 79-82; 2; 3]. Онтологии используются в качестве основного компонента систем интеллектуального анализа данных (data mining), предназначенных для поддержки принятия решений [4].

Вслед за авторами [5] мы считаем необходимым разграничивать лингвистические и формальные онтологии. Формальные онтологии представляются в табличном формате и обрабатываются средствами СУБД. Например, руководство фирмы может установить зависимость между покупками определённых товаров и такими данными о покупателях, как наличие в собственности дома, марка автомобиля, возраст, профессия, доход, расстояние между домом и магазином. В результате, может быть принято решение о таком формировании ассортимента товаров, который повлияет на увеличение количества продаж. Основная трудность при создании формальных онтологий состоит в сборе данных о покупателях. Обычная практика – анкетирование, за участие в котором покупателям предоставляются скидки, бонусы или просто выплачиваются деньги. Одно из направлений разработки формальных онтологий – создание онтологий, отражающих структуру научной дисциплины, которые авторы [6] относят к информационным онтологиям. В [7] описывается онтология археологии и этнографии Сибири, предназначенная для адекватного представления научной дисциплины.

В предлагаемой нами интерпретации особенность лингвистических онтологий состоит в том, что они используются для поддержки функционирования систем интеллектуального анализа текста (text mining) и разрабатываются в рамках лингвистической информатики. На входе у таких систем, так же, как и у другого лингвистического программного обеспечения, – текст на естественной языке, а на выходе – информация, имплицитно содержащаяся в тексте. К такой информации, могут относиться числовые коэффициенты, а также термины, которые отсутствуют во входном тексте и генерируются путём логического вывода. Этим системы интеллектуального анализа текста отличаются от других систем автоматической обработки текста, в частности от информационно-поисковых и автомати-

* Работа выполнена при поддержке Проекта по разработке системы автоматического анализа мнений покупателей грантом РФФИ 16-07-00014

ческого реферирования, которые на выходе предоставляют пользователю информацию, непосредственно содержащуюся во входном тексте. Вместе с тем, возможность получать новую информацию путём логического вывода, обусловливает сходство систем интеллектуального анализа текста с экспертными системами. Существенной особенностью лингвистической онтологии является связь с грамматикой, правила которой позволяют распознавать термины онтологии во входном тексте и генерировать на выходе фразы, отражающие его содержание, в том числе и имплицитное. Соответственно, с позиций теоретической лингвистики, термины онтологии можно определить как единицы языка, а выходные фразы – как единицы речи, что обусловливает высокий уровень сложности систем интеллектуального анализа текста, разработка которых требует применения разноуровневых лингвистических алгоритмов: морфологических, лексических, синтаксических, дискурсивных.

В настоящее время развиваются два основных направления разработки систем интеллектуального анализа текста: системы информационной поддержки сотрудников и системы анализа мнений покупателей.

В качестве примера систем информационной поддержки можно привести проект немецких специалистов [8]. Их целью было создание системы, которая позволила бы инженерам, обслуживающим системы электроизоляции высоковольтных ротационных устройств, получать информацию о методах диагностики неполадок, видах неполадок, их признаках и причинах. Из существующих диагностических отчётов был создан корпус текстов, который автоматически проаннотирован тегами когнитивных ролей (knowledge roles), такими, как *Observed Object*, *Symptom*, *Cause*, и в итоге была разработана поисковая система, позволяющая получать информацию о признаках неполадки конкретного объекта, её причинах и способах устранения. Благодаря этому повышалась эффективность работы персонала, в первую очередь новых сотрудников, которым не нужно было тратить месяцы на знакомство с оборудованием и приобретение опыта его отладки. Схожие системы информационного обеспечения и обмена опытом широко используются в медицине, где создаются базы данных на основе историй болезней, что позволяет врачу узнавать, какие ставились диагнозы, какое лечение назначалось больным с одинаковыми симптомами, а также каковы результаты этого лечения [9].

Системы автоматического анализа мнений покупателей позволяют выявлять наиболее общие недостатки или достоинства товаров, что оказывает существенное влияние на маркетинговую политику и рекламные кампании фирм-производителей, на принятие решений о продвижении тех или иных продуктов. Исходным материалом для систем этого типа служат тексты с отзывами о коммерческих продуктах, которые размещаются пользователями социальных сетей, крупнейшей из которых в России является Яндекс Маркет[1]. На выходе эти системы выдают пользователю: 1) обобщенный коэффициент интенсивности положительных оценок продуктов; 2) обобщенный коэффициент интенсивности отрицательных оценок; 3) обобщенный коэффициент для данного текста/группы текстов в целом; 4) оценки конкретного продукта, указанного в запросе пользователя, а также термины текста, выражающие эти оценки. Пользователь имеет возможность отследить динамику мнений о продукте (продуктах) в течение определенного периода времени. Некоторые системы предлагают в качестве дополнительного сервиса отслеживание новостей и уведомление об упоминании продукта. Кроме того, в настоящее время проводятся исследования, направленные на распознавание в тексте дополнительной демографической информации о пользователях (пол, возраст, уровень образования), что позволяет устанавливать распределение оценок и мнений по категориям пользователей/покупателей и соответственно ориентировать продажи конкретного продукта.

В настоящей статье описывается опыт создания лингвистической онтологии в рамках реализации первого этапа проекта по разработке системы автоматического анализа мнений покупателей. На входе у системы, локализованной для русского языка, – текст отзыва и запрос с названием продукта. На выходе: общие коэффициенты, отражающие интенсивность положительных и отрицательных оценок этого продукта во входном тексте; обобщенный коэффициент, вычисляемый по среднему арифметическому отрицательных и положительных оценок; конкретные термины (словосочетания и предложения), выражающие отрицательную или положительную оценку с приписанным каждому термину числовым коэффициентом, отражающим степень интенсивности оценки. Созданная нами онтология включает три тематических категории: «Телефоны», «Гостиницы», «Фильмы», что позволяет охватить три разнородных сферы, связанные с искусством, электроникой, услугами, и сделать систему более универсальной.

ЭТАПЫ СОЗДАНИЯ ОНТОЛОГИИ В ПОЛУАВТОМАТИЧЕСКОМ РЕЖИМЕ

1. Составление контрольного корпуса текстов, необходимого для взвешивания терминов с помощью разработанного нами приложения TF*IDF Ranker[2] .

В корпус были включены тексты по техническим и естественнонаучным дисциплинам: медицине, физике, строительству, программированию, экономике, кибернетике, спорту, военному делу. Также были добавлены три классических художественных произведения двух авторов – И.С. Тургенева и А.К. Толстого. Все тексты были взяты из электронной библиотеки Мошкова[3]. Обращаясь к этой библиотеке, мы учитывали, что размещенные в ней тексты модерируются и оцениваются на предмет нарушения авторских прав. При составлении этого текстового корпуса мы сходили из того, что в процессе взвешивания терминов будет применяться контрастивное сопоставление, т.е. тексты, относящиеся к указанным трём тематическим категориям, будут сопоставляться с текстами других жанров. Как предполагалось, это

[1] https://market.yandex.ru/

[2] http://yatsko.zohosites.com/tf-idf-ranker1.html

[3] www.lib.ru

позволит выявить термины, специфичные для указанных тематических категорий.

Общий объем контрольного текстового корпуса составил 1423528 токенов в 31 файле. Определяя его размер, мы учитывали, что размер эталонного текста должен быть не менее миллиона токенов, поскольку именно на текстах такого размера выполняется закон Ципфа. Статистические данные были получены с помощью конкорданса AntConc 3.4.3w, который распространяется бесплатно по лицензии *freeware* и поддерживает обработку русских текстов в кодировке UTF-8[4].

2. Составление эталонного текстового корпуса.

Было составлено три текстовых файла, соответствующих выбранным тематическим категориям. Файл для категории «Телефоны» был составлен на основе отзывов покупателей на Яндекс-Маркете, представляющем собой модерируемый ресурс, на котором для каждого товара приводятся отзывы, сгруппированные по общей оценке, выставленной покупателем – от 1 до 5 (*ужасно, плохо, обычно, хорошо, отлично*). Это позволило отдельно выбирать отрицательные, положительные и средние отзывы, что обеспечило необходимый диапазон оценочной терминологии. Файл для категории «Гостиницы» был составлен на основе отзывов на портале *Tripadvisor*[5] Это также модерируемый ресурс с пятибалльной шкалой оценок и с такой же группировкой отзывов по оценкам (оценка в три балла=*неплохо*). Были выбраны отзывы с оценками морских курортов в Таиланде, Турции и России, а также отзывы о российских горнолыжных базах отдыха и гостиницах в крупных некурортных городах, что обеспечило достаточную, как мы полагаем, репрезентативность. Выбирались только отзывы, написанные русскоязычными авторами. Файл для категории «Фильмы» был создан на основе отзывов на сайте Кинопоиск[6], на котором применяется трёхбалльная система оценок фильмов (положительная, отрицательная, нейтральная); отзывы размещаются только после проверки, при этом к ним предъявляются требования по объёму и структуре.

Отметим, что мы не использовали ресурсы агрегаторов отзывов, таких как "Отзовик" и «IRecommend», поскольку они, в отличие от указанных выше ресурсов, оплачивают отзывы, что создаёт вероятность размещения заказных отзывов и вызывает много нареканий[7].

Статистические данные созданных корпусов текстов приводятся в табл. 1.

3. Взвешивание терминов эталонного корпуса.

Для терминов в каждом из трёх текстов эталонного корпуса были получены весовые коэффициенты с помощью приложения TF*IDF Ranker; применялась модифицированная формула TF*IDF [10]. Оправдалось наше предположение о том, что при контрастивном сопоставлении наиболее высокие коэффициенты получат термины, которые либо не используются в контрольном корпусе, либо используются очень редко. Первые места в ранжированном списке занимают термины, специфичные для отзывов покупателей.

Использование TF*IDF Ranker позволило существенно ускорить разработку онтологии.

4. Создание списков параметрических терминов.

Анализ результатов взвешивания позволил выявить два основных вида терминов: параметрические и оценочные. Параметрические термины обозначают оцениваемые параметры, компоненты объекта; оценочные – указывают на степень интенсивности отрицательной или положительной оценки.

Все термины обрабатывались разработанным нами стеммером [11], и в онтологию включались стеммы, для того, чтобы разрабатываемая система в процессе поиска могла отождествлять словоформы с одной основой. Для русского языка это особенно актуально, учитывая его морфологическую развитость (Ср.: основу *плох*, с которой соотносятся 16 словоформ: *плохой, плохая, плохие, плохого, плохому, плохом, плохим, плохую, плохой, плохих, плохими, плохое, плох, плохо, плохи, плоха*).

Для категории «Телефоны» было выявлено 188 параметрических терминов (стемм), распределение которых имеет отчётливо выраженную иерархическую структуру с глубиной до пяти уровней: на первом уровне находятся слова-гиперонимы по отношению к ядерному слову *телефон*: *девайс, устройств, аппарат*; на втором уровне – ядерное слово *телефон* и кореферентные термины (*смартфон, смарт*). Отметим, что имеется в виду контекстуальная кореферентность, поскольку в лексико-семантическом плане эти термины не являются кореферентными; на третьем уровне – гипонимы к термину *телефон*; на четвертом – гипонимы к терминам третьего уровня, например, *девайс – телефон – производительность – тест*. Наиболее многочисленной является группа терминов на третьем уровне, которые и обозначают основные параметры телефона (*связ, ОС, эргономик, энергосбережен, экран, чехол, характеристик, фонарик, софт, сим, сборк, разъем, процессор, плат, настройк, модел, корпус, конструкц, комплект, карт памят, производительн, кабел, камер, зарядк, докстанц, дизайн, датчик, гарнитур, аккумулятор, автономн, адаптер, аксессуар*). На четвертом уровне больше всего терминов-гипонимов к параметру *связ*, что вполне естественно, поскольку это основная функция телефона.

Выделение кореферентных терминов имеет существенное значение для последующей работы над проектом, чтобы адекватно соотносить оценочные термины с именем оцениваемого объекта в процессе разрешения анафоры. Выделение гипонимов и гиперонимов важно для реализации функции интеллектуального анализа. Например, в отзыве покупателя могут оцениваться отдельные параметры связи, такие как *прием, вай-фай, блютус*. Соотнеся эти параметры с термином-гиперонимом, предлагаемая система может автоматически сгенерировать обобщенную оценку в виде фразы *качественная связь*, которая в самом отзыве не содержится.

Для категории «Гостиницы» было найдено 297 параметрических терминов, при этом наиболее многочисленными подкатегориями являются «Номер» (64 термина) и «Питание» (49 терминов) с глубиной

[4] http://www.laurenceanthony.net/software.html
[5] https://www.tripadvisor.ru
[6] https://www.kinopoisk.ru
[7] http://анти-мошенник.рф/vkcom/data/irecommend.ru/

иерархии до 5 уровней. Появились новые подкатегории, такие как «Персонал», «Отдых», «Контингент», а также существенно увеличилось количество терминов в категории «Сервис». Это вполне объяснимо, так как постояльцы гостиниц намного больше контактируют с обслуживающим персоналом.

Для категории «Фильмы» выделение параметрических терминов имеет особое значение, так как оценочные термины в текстах отзывов рецензий могут соотноситься не с фильмом, а с описанием персонажей, событий, сюжета (Ср.: *Между бандитами и полицейскими завязалась* ***жестокая*** *перестрелка* и *Фильм пропагандирует* ***жестокость****, нетерпимость, насилие*). Соответственно, эти оценки не должны учитываться. На следующем этапе реализации проекта при разработке грамматики будут созданы специальные правила, позволяющие разграничить эти два вида оценок.

Глубина иерархии для этой тематической категории также составляет 5 уровней, причем на первых двух уровнях находятся термины-гиперонимы: *культура – кинематограф=киноиндустрия*. Наиболее многочисленными являются класс «Персонаж», который включает 29 терминов, и класс «Сюжет», включающий 23 термина. Именно эти классы важны для разграничения оценок, относящихся к фильму и к событиям в фильме. Всего было выделено 230 параметрических терминов для этой тематической категории. В табл. 2 приводятся примеры параметрических терминов.

Таблица 1

Статистические данные текстовых корпусов

Категория	Уникальные слова	Токены	Размер (кб)
Телефон	13950	51684	621
Гостиницы	11370	51688	607
Фильмы	10384	51781	652
Контрольный корпус (31 файл)	109069	1423528	22759

Таблица 2

Примеры параметрических терминов категории «Гостиницы»
(знак равенства указывает на кореферентные термины)

Уровень иерархии				
Термины				
Отел= гостиниц =hotel= курорт	Персонал	Руководство	Администратор =манагер	
		Сотрудник	Аниматор	
			Бармен	
			Водител	
			Горничн	
			Девоч= девушк= женщин	
			Носильщик	
			Официант	
			Охранник	
			Повар	
			Уборщиц	
			Шеф-повар	
	Номер	Комнат	тумбочк	
			Розетк	
			Батар	
			Вентилятор	
			Окн	Штор
			Телевизор	Пульт
				Видео
			Кондиционер	
			Чайник	
			Шкаф	
			Вешалк	
			Шкаф	
			Вешалк	
			Кроват	Бель
				Матрас
				Подушк

5. Создание списка оценочных терминов.

Из каждого из эталонных текстов были удалены выявленные ранее параметрические термины (использовалась соответствующая функция конкорданса AntConc). Далее для оставшихся текстов были получены коэффициенты терминов с помощью приложения TF*IDF Ranker. Сопоставление проводилось с контрольным текстовым корпусом, включающим 31 файл (см. табл.1). При удалении параметрических терминов оценочные термины получили высокие коэффициенты и сгруппировались в верхней части списка.

Вначале был получен ранжированный список для тематической категории «Фильмы» (Список 1), из которого были удалены нейтральные слова и найдены оценочные термины. Затем были получены ранжированные списки для двух других тематических категорий (Список 2, Список 3).

С целью удаления дубликатов было произведено (с помощью формул MS Excel) последовательное пересечение и вычитание терминов в трёх списках, после чего был проведён стемминг. Получившийся общий список стемм был отредактирован, удалены повторяющиеся стеммы.

В этом списке были выделены семантические и синтаксические оценочные термины. Семантические термины непосредственно выражают оценочную семантику, в то время как синтаксические термины могут изменять интенсивность положительной или отрицательной оценки, выражаемой семантическими терминами (Ср.: *очень хорошо* и *очень плохо*, где синтаксический термин *очень* изменяет интенсивность семантического термина *хорошо*).

6. Семантическим терминам онтологии (стеммам) были начислены коэффициенты в зависимости от интенсивности выражаемой положительной или отрицательной оценки. Присваивание оценочных коэффициентов проводилось участниками проекта вручную по принципу консенсуса. Получилась шкала из семи уровней: от 1 до 7 и от -1 до -7. Общее количество оценочных семантических терминов в настоящее время – 1618 стеммы. В табл. 3 представлены данные об этом компоненте онтологии.

Подчеркнём, что речь идёт о количестве стемм, каждой из которых соответствует множество производных словоформ с суффиксами и падежными окончаниями. Мы включили в этот словарь лишь несколько словоформ в степенях сравнения, которые встречались в текстах. На следующем этапе будет написано специальное правило для распознавания терминов со степенями сравнения, что позволит существенно увеличить количество терминов с шестыми и седьмыми коэффициентами.

В процессе создания онтологии возникли трудности, связанные с морфологической, лексической и грамматической омонимией и синонимией терминов.

Под лексической омонимией мы понимаем использование словоформ одной лексемы как в нейтральном значении, так и в оценочном значении. Типичный пример – использование одних и тех же форм глагола:

(1) *Галустян* ***достал*** *его оттуда зубами*.

(2) *Ты* ***достал меня*** *уже со своей Эфиопией!!*

(3) *Наполен* ***захватил*** *Москву*.

(4) *Фильм* ***захватывает*** *с первого кадра*.

(5) *Сценарий. Настолько* ***захватывающий*** *и детально продуманный, учтена каждая мелочь*.

(6) *Капкан состоит из 8 дуг,* ***захватывающих*** *конечность жертвы*.

Разрешение лексической омонимии предусматривает анализ контекста. В (1) глагол используется в качестве нейтрального термина, в то время как в (2) такая же глагольная форма выражает отрицательное значение. Оценочная семантика глагола актуализируется в составе устойчивого сочетания с личным местоимением, соответственно, для разграничения нейтрального и оценочного использования форм глагола *доставать* достаточно написать правило, указав список местоимений. Вообще, изменение семантики с нейтральной на оценочную типично для фразеологизмов и идиом (Ср.: *срубить дерево* и *срубить бабки*). В этих случаях для распознавания оценочного термина можно указать термины, с которыми он контекстуально связан, и описать специальным правилом их позицию в клаузе.

Таблица 3

Словарь оценочных семантических терминов (фрагмент)

№	Коэффициент	Количество терминов	Пример стеммы	Пример соответствующей словоформы
1	1	150	Авторитет	Авторитетный
2	-1	292	Гримас	Гримаса
3	2	120	Аромат	Аромат
4	-2	236	Безразлич	Безразличие
5	3	124	Искрен	Искренний
6	-3	222	Безвкус	Безвкусный
7	4	118	Достойн	Достойный
8	-4	171	Безобраз	Безобразнй
9	5	50	Безупречн	Безупречный
10	-5	78	Абсурд	Абсурд
11	6	11	Бесподобн	Бесподобный
12	-6	31	Испохаб	Испохабить
13	7	7	Совершен	Совершенный
14	-7	8	Блев	Блевать
Всего		**1618**		

В (3) *захватил* используется как нейтральный термин в сочетании с одушевленным существительным, в то время как в (4) он сочетается с параметрическим термином, выраженным неодушевлённым существительным. Контекстуальная оппозиция *одушевленное – неодушевленное* существительное типична разграничения нейтральных и оценочных терминов (Ср.: *сюжет предсказуем – экстрасенс предсказал; мы еле тащимся – мы тащимся от этого фильма; законченное произведение – законченный негодяй*). Для разрешения омонимии в данном случае требуется грамматика, разработать которую мы предполагаем на следующем этапе на основе аннотирования частями речи.

Грамматика понадобится и для разграничения более сложных случаев, когда нейтральные и глагольные формы характеризуются одинаковой сочетаемостью (Ср.: *Галустян достал* в (1) и *Авторы достали своей "исторической правдой"*). В этом случае нейтральная и оценочная формы сочетаются с одушевленным существительным. Для разрешения омонимии требуется выполнить разрешение кореференции, благодаря чему можно установить, что под авторами имеются в виду авторы фильма.

В (5) явно выражена оценочная семантика термина, который сочетается с параметрическим термином *сценарий*. В данном случае, однако, не требуется разработки отдельных правил, поскольку наиболее частотным является употребление этого термина в оценочном значении. В некоторых справочниках указывается, что это словоупотребление относится к прилагательному, а не к причастию, т. е. в данном случае речь должна идти о морфологической омонимии. Это вполне правомерно, поскольку прилагательное может использоваться в качестве именной части составного сказуемого, а причастие – как часть составного глагольного сказуемого (Ср.: *Этот фильм – просто захватывающий*; *Этот город должен быть захвачен*).

Под морфологической омонимией мы понимаем термины с одной основой (стеммой), относящиеся к разным частям речи. В процессе стемминга они отождествляются по одной стемме:

(7) *Зато в некоторых сценах было* ***откровенно*** *скучно…*

(8) *Не представляю, как самым маленьким зрителям может это понравиться из-за* ***откровенного*** *бреда, происходящего на экране.*

(9) *Неприметная, нельзя ее назвать красоткой, в стандартом понимании, с довольно хорошим чувством юмора и саркастическим взглядом на жизнь и конечно же с предельной* ***откровенностью.***

В этих примерах наречие и прилагательное используются в качестве синтаксических терминов, т.е. они усиливают отрицательную оценку, которая выражается семантическими терминами (*скучно, бреда*). В (9) существительное с той же основой используется в качестве оценочного термина, причем с положительным значением. По-видимому, наиболее частотными являются случаи омонимии причастий (нейтральных терминов) и прилагательных (оценочных терминов). Однако это различие наблюдается и в других частях речи, возможны разные варианты:

- существительное – прилагательное (ср.: *знак – знаковый*);
- существительное – глагол (ср.: *впечатление – впечатлять*);
- глагол – прилагательное (ср.: *выражать – выразительный*).

Разрешение омонимии в данном случае возможно за счёт аннотирования тегами частей речи.

Возможны и случаи морфологической синонимии, под которой понимается использование словоформ, выражающих разную степень интенсивности оценки и соотносящихся с одной стеммой:

(10) *Ремейки классики, которые* ***совершенствовали*** *оригинал.*

(11) *Смартфон ZTE Axon Mini:* ***совершенство*** *без компромиссов.*

(12) *Все персонажи —* ***совершенно*** *ужасные.*

(13) ***Совершенно*** *потрясающе передана атмосфера той самой Одессы.*

В (10) и (11) глагол и существительное являются оценочными терминами, однако выражают разную степень интенсивности положительной оценки; в (12) и (13) – наречие используется как синтаксический термин, причем эти примеры показывают, что он повышает интенсивность как положительной, так и отрицательной оценки. Эта проблема решается выделением наречия в отдельный термин, который не подлежит стеммингу, и указанием различных стемм для глагола (*совершенствов*), существительного (*совершентств*) и прилагательного (*совершенн*).

Особо сложные случаи – использование синонимичных терминов, которые могут приобретать как положительную так и отрицательную семантику:

(14) *Я офигел от звука. Настолько он четкий.*

(15) *Вы что офигели?*

(16) *Офигенный кальян-бар.*

В (14) и (16) просторечное слово «офигеть» и соответствующее прилагательное выражают положительную семантику, а в (15) – отрицательную. Анализ показывает, что прилагательное обычно используется в положительном значении, в то время как значение глагола зависит от контекста. В первом случае проблема решается аннотированием тегами частей речи, а второй требует контекстуального анализа, методы которого еще предстоит разработать.

Можно предположить, что существуют и случаи лексической синонимии, когда словоформы одной лексемы выражают разные степени оценки, однако они нам пока не встречались.

Проведённый нами анализ позволил установить, что степень интенсивности оценочной семантики определяется следующими факторами:

- семантикой самого оценочного слова (*неплохо – хорошо – отлично*). Нами была создана шкала, включающая семь степеней оценки от 1 до 7 и от -1 до -7;
- использованием степеней сравнения (*глупый – глупее – глупейший*). В группу терминов онтологии со степенью интенсивности 7 и -7 в основном вошли термины с превосходной степенью сравнения;
- использованием сленговых слов и жаргонизмов, которые актуализируют отрицательную семантику по сравнению со стилистически нейтральным эквивалентом (*деньги – бабло – баблосы*);

▪ использованием уменьшительных форм, которые: (а) снижают интенсивность отрицательной оценки (*мрачный – мрачноватый*), (б) придают отрицательную семантику семантически нейтральным словам (*фильм – фильмец*), (в) меняют положительную семантику термина на отрицательную (*легенда – легендочка*).

В созданной нами онтологии приводятся стеммы наиболее частотных сленговых слов, а также терминов в степенях сравнения, однако на следующем этапе представляется необходимым разработать дополнительные правила образования уменьшительных форм существительных и степеней сравнения прилагательных путем добавления соответствующих суффиксов. Это может быть реализовано на основе предварительного аннотирования тегами частей речи посредством теггера, разработка которого запланирована на следующем этапе.

К синтаксическим относятся термины, которые сами по себе не имеют оценочного значения, но могут изменять интенсивность оценки, выражаемой семантическими терминами. Отношение между синтаксическими и семантическими терминами мы рассматриваем как бинарное, выделяя подгруппы с симметричным, асимметричным, несимметричным и обратным отношениями между ними.

Симметричное отношение реализуется следующими группами синтаксических терминов:

термины, которые увеличивают интенсивность как положительной, так и отрицательной оценки (Ср.: *очень хорошо* и *очень плохо.* В этих фразах синтаксический термин *очень* увеличивает интенсивность как положительной, так и отрицательной оценки, выражаемой семантическими терминами *хорошо* и *плохо*);

термины, которые снижают интенсивность как положительной, так и отрицательной оценки (Ср.: *почти идеальный* и *почти бессмысленный.* В этих фразах синтаксический термин *почти* снижает интенсивность как положительной, так и отрицательной оценки, выражаемой семантическими терминами *идеальный* и *бессмысленный*).

Асимметричное отношение реализуется следующими группами синтаксических терминов:

термины, которые усиливают интенсивность отрицательной оценки и снижают интенсивность положительной оценки (Ср.: *слишком дорогой* и *слишком прекрасный*. В таких фразах синтаксический термин *слишком* снижает интенсивность положительной оценки, выражаемой семантическим термином *прекрасный*, и усиливает интенсивность отрицательной оценки, выражаемой семантическим термином *дорогой*);

термины, которые усиливают интенсивность положительной оценки и снижают интенсивность отрицательной оценки. В процессе выполнения проекта мы не нашли примеров таких терминов, однако можно предположить, что они обнаружатся в дальнейших исследованиях.

Обратное отношение реализуется терминами, изменяющими оценочное значение семантического термина на противоположное: отрицательное значение – на положительное, а положительное – на отрицательное (Ср.: *без волокиты* и *без достоинств.* В этих фразах синтаксический термин *без* меняет положительную семантику термина *достоинств* на отрицательную, а отрицательную семантику термина *волокиты* – на положительную. Ср.: также, *не порадовал* и *не разочаровал*).

Несимметричное отношение реализуется двумя группами терминов: терминами, которые усиливают положительную оценку (позитивно-синтаксическими), и терминами, которые усиливают отрицательную оценку (негативно-синтаксическими).

(17) *Номер Люкс за 4000 руб. в сутки – было откровенно ужасно.*

(18) *Зимний отдых в Сочи также сказочно хорош.*

(19) *"Форпост" - это фильм-откровение.*

В (17) синтаксический термин *откровенно* усиливает отрицательную семантику семантического термина *ужасно*; при этом вряд ли допустимо использование этого синтаксического термина в сочетании с термином, выражающим положительную оценку (Ср.: ?*откровенно прекрасно*; *откровенно плохой* и ?*откровенно хороший*). В (18) *сказочно* усиливает положительную семантику *хорош*, однако *сказочно* сам по себе выражает положительную семантику, т.е. является семантическим термином (Ср.: *сказочный пейзаж*). Нам пока не удалось найти чисто синтаксических терминов, которые бы усиливали интенсивность положительной оценки. Хотя, исходя из логики бинарных оппозиций, они должны существовать. Возможно, их удастся найти в дальнейшем.

Возможны и гибридные варианты бинарных отношений между синтаксическими и семантическими терминами. Например, несимметрично-обратное отношение может реализовываться глагольными формами *разоблачать, раскрывать, показывать, выявлять*:

(20) *В фильме открыто* ***разоблачаются*** *преступления власти.*

(21) *Книга* ***раскрывает*** *их* ***высокие профессиональные качества.***

(22) *В фильме* ***раскрывается*** *механизм* ***криминального беспредела.***

В (20) отрицательная семантика термина *преступления* снимается глаголом *разоблачаются* и предикативная конструкция в целом приобретает положительную семантику. При этом использование терминов с положительной семантикой в сочетании с этим глаголом вряд ли допустимо (Ср.: ?*разоблачается доброта*). В (22) также имеет место нейтрализация отрицательной семантики термина *беспредел*, однако, глагол раскрывать может сочетаться и с терминами с положительной семантикой (21).

Синтаксические термины распределены по разным частям речи и, соответственно, характеризуются различной сочетаемостью, что будет необходимо учитывать при разработке грамматических правил. В связи с этим целесообразно выделить два вида терминов по критерию морфологической изменяемости. Некоторые синтаксические термины используются только в одной форме, в то время как другие могут использоваться в качестве разных частей речи, сохраняя свои свойства. Примеры (23), (24) и (25) пока-

зывают, что наречие *совершенно* используется как синтаксический термин, а прилагательное – как семантический термин:

(23) *Я купила телефон, а он* ***совершенно отвратительно*** *работает.*

(24) *Персонал* ***совершенно непрофессиональный.***

(25) *G3:* ***совершенный*** *смартфон нового поколения.*

Следует отметить, что наречие может использоваться и как семантический термин (ср.: *в этом фильме всё совершенно*), однако такие случаи встречаются редко и в составленных нами корпусах не обнаружены, что позволяет ими пренебречь. В качестве синтаксического термина может использоваться форма превосходной степени сравнения: *совершеннейшая глупость*. Эта форма включена в онтологию. Вообще, следует учитывать наличие степеней сравнения у синтаксических терминов (Ср.: *огромный недостаток* и *огромнейший недостаток*).

Более сложный случай – совпадение форм синтаксического и семантического терминов, которое наблюдается в случае с термином *откровенный, откровенно*, поскольку и формы прилагательного, и формы наречия могут использоваться и для выражения оценки, и для её усиления. Эта проблема решается правилом грамматики, в котором указывается сочетаемость термина: если он используется в одной клаузе в позиции перед семантическим термином, то его следует рассматривать в качестве синтаксического.

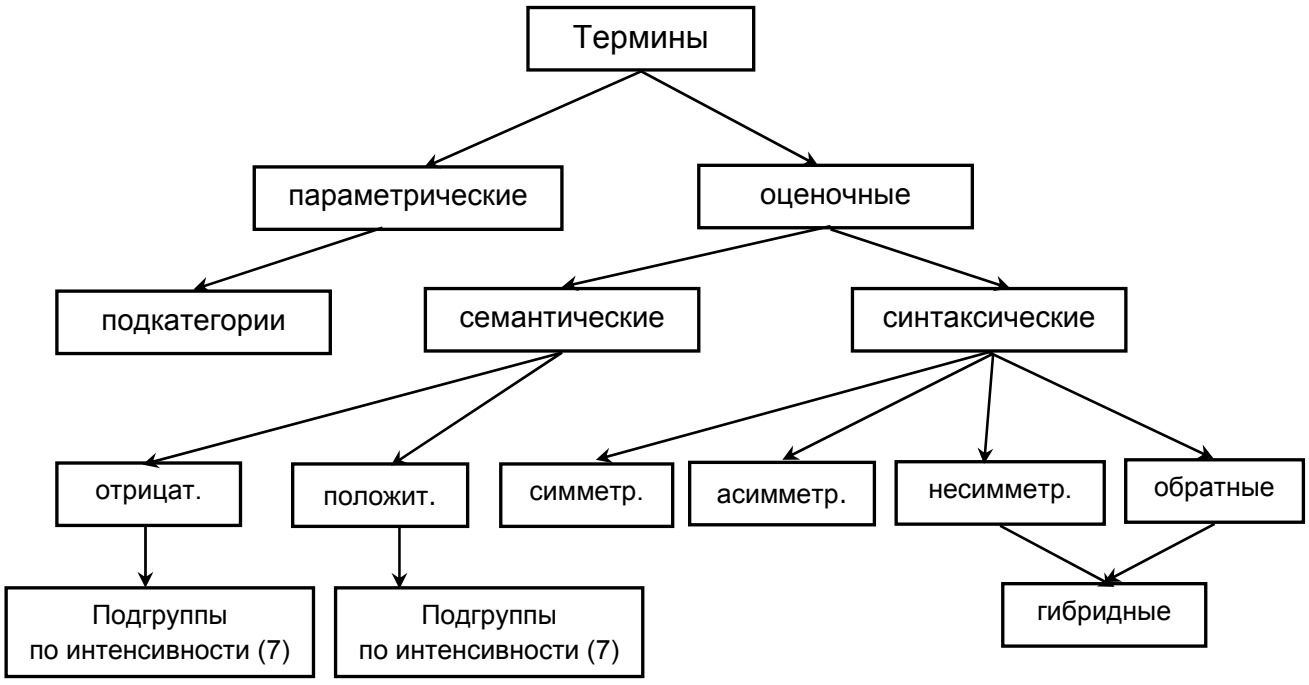
Обобщенная структура онтологии представлена на рисунке.

ЗАКЛЮЧЕНИЕ

В обзоре работ по автоматическому анализу оценочной лексики и тональности текстов Н.В. Лукашевич и И.И. Четвёркин [12] ограничились констатацией факта отсутствия разработок, локализованных для русского языка. Как мы полагаем, для научного исследования в области лингвистической информатики важно выявить универсальные принципы построения таких систем, которые должны быть реализованы независимо от специфики конкретного языка. Наша работа позволила впервые установить следующие фундаментальные принципы создания лингвистической онтологии для рассматриваемой предметной области.

1. Лингвистическая онтология представляет собой сложноструктурированный словарь, моделирующий структуру заданной предметной области, связанный с грамматикой. Грамматика выполняет функции распознавания единиц онтологии в текстах на естественном языке и генерации выходных фраз. Лингвистическая онтология и грамматика составляют две подсистемы систем интеллектуального анализа текста.

2. Лингвистическая онтология включает параметрические и оценочные термины. Параметрические термины обозначают компоненты и параметры оцениваемого объекта; оценочные термины указывают на степень интенсивности положительной или отрицательной оценки.



Обобщенная структура разработанной лингвистической онтологии

Выделение параметрических терминов в качестве отдельного класса позволяет:

- выполнять адекватное соотнесение оценочных терминов с именем оцениваемого объекта;
- повышать степень интеллектуальности системы благодаря возможности генерировать обобщенные выводы, не содержащиеся в самом тексте отзыва;
- выполнять разрешение кореференции;
- снимать омонимию.

Именно параметрические термины выполняют функцию моделирования предметной области, которая считается одним из свойств онтологии. Другой класс терминов онтологии – оценочные термины – не имеет чёткой привязки к какой-либо предметной области.

3. Оценочные термины делятся на два вида: синтаксические и семантические термины. Лексическое значение семантических терминов содержит оценочный компонент; синтаксические термины не выражают оценки, однако могут изменять интенсивность оценки, выражаемой семантическими терминами. Отношение между синтаксическими и семантическими терминами описывается следующими аксиоматическими выражениями:

- отношение представляет собой отношение между зависимой и независимой переменной. Синтаксические термины учитываются, только если они встречаются в одной клаузе с семантическими терминами. Соответственно, программная реализация алгоритма функционирования системы предусматривает, что вначале находится семантический термин, а затем в той же клаузе ищется синтаксический термин. В том случае, если он находится, числовой коэффициент, отражающий интенсивность оценки, выражаемой семантическим термином, изменяется в соответствии с правилами, установленными для данной системы;
- отношение является бинарным и может быть симметричным, асимметричным, несимметричным, обратным, либо гибридным. В случае симметричного отношения синтаксический термин увеличивает или уменьшает интенсивность положительной или отрицательной оценки, выражаемой семантическим термином. В случае асимметричного отношения синтаксический термин увеличивает интенсивность положительной оценки и уменьшает интенсивность отрицательной, либо наоборот. В случае несимметричного отношения синтаксический термин увеличивает или уменьшает либо отрицательной, либо положительной оценки. В случае обратного отношения синтаксический термин меняет ориентацию семантического термина на противоположную. Гибридные варианты представляют собой комбинации основных видов бинарного отношения.

Неточным является представление о том, что слова-модификаторы лишь увеличивают или уменьшают эмоциональный вес соседнего слова [13]. Заметим, что положение о бинарном отношении между синтаксическими и семантическими терминами, сформулированное нами еще в 2011 г. [14], было проигнорировано авторами [13].

4. Семантические оценочные термины делятся на подгруппы в зависимости от интенсивности выражаемой отрицательной или положительной семантики. Градационная шкала определяется особенностями каждого реализуемого проекта и может быть различной. В нашем проекте она включает 7 уровней, что, как мы предполагаем, является оптимальным для русского языка.

В ходе реализации проекта нами были впервые выявлены, проанализированы на материале русского языка и классифицированы случаи омонимии и синонимии оценочных терминов, а также предложены подходы к решению проблем, связанных с их многозначностью, на основе разработки грамматики, что является задачей реализации следующего этапа нашего проекта.

СПИСОК ЛИТЕРАТУРЫ

1. Hoekstra R. Ontology representation: design patterns and ontologies that make sense. – Amsterdam: IOS Press, 2009. – 248 p.
2. Gruber T. Towards principles for the design of ontologies used for knowledge sharing // International journal of human and computer studies. – 1995. – Vol.43, № 5/6. – P. 907-928.
3. Pretorius A.J. Ontologies - introduction and overview. – 2004. – URL: http://starlab.vub.ac.be/teaching/Ontologies_Intr_Overv.pdf
4. Mansingh G., Rao L. Enhancing the decision making process: An ontology-based approach // CONF-IRM Proceedings. – 2014. – URL: http://aisel.aisnet.org/cgi/viewcontent.cgi?article=1018&context=confirm2014
5. Добров Б.В., Лукашевич Н.В. Лингвистическая онтология по естественным наукам и технологиям для приложений в сфере информационного поиска // Web journal of formal, computational and cognitive linguistics. – 2006. – URL: http://fccl.ksu.ru/fcclpap.htm .
6. Марчук А.Г., Целищев В.В. О расхождении информационных онтологий с концептуализациями внешнего мира // Философия науки. – 2013. – № 4. – С. 67-78. – URL: http://www.sibran.ru/upload/iblock/ce6/ce63797450b84ed6a7e203ab8a5b8fe0.pdf
7. Марчук А.Г., Холюшкин Ю.П., Загорулько Ю.А. и др. Интеллектуальный интернет-портал знаний для доступа к информационным ресурсам по археологии и этнографии // Информационные технологии в гуманитарных исследованиях. – 2004. – Вып. 8. – С. 7-13. – URL: http://www.prometeus.nsc.ru/elibrary/infohum/2004-008.pdf
8. Mustafaraj E., Hoof V., Freisleben D. Mining diagnostic text reports by learning to annotate knowledge roles // Natural language processing and text mining / eds. A. Kao, S. R. Poteet. – London: Springer, 2007. – P. 45-68.
9. Razika, D., Houda B., Nawel K. Domain ontology building process based on text mining from medical structured corpus // The proceedings of the international conference on digital information processing, data mining, and wireless communications. – Dubai, 2015. – P. 128-139. – URL: http://www.academia.edu/10352217/Domain_Ontology_Building_Process_Based_on_Text_Mining_from_Medical_Structured_ Corpus

10. Яцко В.А. Достоинства и недостатки взвешивания терминов по формуле Tf*Idf // В мире научных открытий. – 2013. – № 6. – С. 229-244.
11. Яцко В.А. Особенности разработки стеммера // Символ науки. – 2016. – № 10, Ч. 2. – С. 103-105.
12. Лукашевич Н.В., Четвёркин И.И. Открытое тестирование систем анализа тональности на материале русского языка // Искусственный интеллект и принятие решений. – 2014. – № 1. – С. 25-33. – URL: https://istina.msu.ru/publications/article/6927077/
13. Клековкина М.В., Котельников Е.В. Метод автоматической классификации текстов по тональности, основанный на словаре эмоциональной лексики // Труды 14-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции». – Переславль-Залесский, 2012. – С. 81-86. – URL: http://ceur-ws.org/Vol-934/paper15.pdf
14. Яцко В.А., Стариков М.С. Опыт разработки онтологии для автоматического анализа мнений пользователей о коммерческих продуктах // Научно-техническая информация. Сер.2. – 2011. – № 7. – С. 9-14; Yatsko V.A., Starikov M.S. On the Experience of Designing an Ontology for Automatic Analysis of User Sentiments about Commercial Products // Automatic Documentation and Mahematical Linguistics. – 2011. – Vol. 45, №4. – P.163-168.

Материал поступил в редакцию 10.02.17.

Сведения об авторах

ЯЦКО Вячеслав Александрович – доктор филологических наук, профессор, Хакасский государственный университет им. Н.Ф. Катанова, г. Абакан
e-mail: viatcheslav-yatsko@rambler.ru

ЯЦКО Татьяна Сергеевна – кандидат филологических наук, доцент, Хакасский государственный университет им. Н.Ф. Катанова, г. Абакан
e-mail: tatiana-yatsko@rambler.ru

УДК 81’ 322.4

А. В. Новикова, Л. А. Мыльников

Вопросы реализации машинного перевода текстов деловой коммуникации для языковой пары «русский язык – английский язык»

На примере текстов деловой коммуникации рассмотрены практические аспекты, связанные с проблемой искажения смыслов при переводе с одного языка на другой с использованием существующих систем машинного перевода и лежащего в их основе подхода, основанного на пословном переводе. Предлагается комплексный функциональный метод перевода текстов деловой коммуникации на основе анализа семантических и морфологических признаков фактического содержания текста, а также аксиологических и эпистемических семантических признаков, актуализирующих субъективную модальность. На основе предложенного метода разработан алгоритм машинного перевода текстов деловой коммуникации, позволяющий решить проблему пословного перевода и передавать смыслы небольших текстов. Приведены примеры апробации предложенной методики и полученного алгоритма на языковой паре «Русский язык – Английский язык».

Ключевые слова: *машинный перевод, деловая коммуникация, передача смысла, субъективно ориентированное сообщение, модальный окрас, смысловое содержание, фактическое содержание, семантические признаки, метаданные языка общения, структура предложения*

ВВЕДЕНИЕ

За последние десятилетия интерес к системам машинного перевода растет; по данным мировых аналитических агентств (IDC, Acute Market Reports, Business Wire) рынок таких систем к 2020 г. достигнет $983,3 млн. Причиной этого являются механизмы глобализации, которые охватили все сферы деятельности – от экономики, культуры и индустрии до общения в социальных сетях и путешествий. В связи с этим увеличивается и спрос на качество систем машинного перевода (СМП). Если проследить историю развития таких систем, то можно выделить несколько этапов в становлении этой области. Впервые идея автоматизации перевода возникла в XVII в. и только в 1930-х гг. американские, российские и французские ученые впервые всерьез занялись этой проблемой [1]. George Artsrouni (1933) и Petr Smirnov-Troyanskii (1933) негласно считаются основателями машинного перевода, хотя впервые о возможности использовать компьютер для осуществления перевода с одного языка на другой объявила организация Warren Weaver of the Rockefeller Foundation [2]. В период с 1950 по 1960 гг. известность в области систем машинного перевода получила корпорация RAND. С тех пор идея машинного перевода стремительно развивалась, пройдя этапы от прямого структурно-грамматического подхода до методов корпусной лингвистики и использования нейронных сетей для поиска адекватной генерации смысла.

Следуя классификации, предложенной в работе [1], можно выделить три подхода к созданию систем машинного перевода: 1) прямой подход, разрабатываемый исключительно для одной языковой пары; 2) подход «Interlingva», базирующийся на общих грамматических и семантических «примитивных» конструкциях, характерных для разных систем языка; 3) гибридный подход, охватывающий различные комбинации способов анализа текста и методов. Большинство современных систем машинного перевода, таких как SYSTRAN, Babylon, Interlingva, Google Translate, Yandex Translate, PROMT и т.д. стараются охватить как можно большее число языковых пар и являются образцами гибридной модели [3], использующими: а) структурно-грамматические методы, изначально разработанные для систем GAT, COMIT, METAL, ESPERANTO; б) синтаксические подходы, предложенные специалистами в области машинного перевода (Paul Garvin, Anthony Brown, Ariadne Lukjanow и др.); в) семантические подходы, характерные для систем ЭТАП-1,2,3, DLT, Rosetta, KANT. В настоящее время функционирование таких систем нельзя представить без достижений корпусной лингвистики и данных статистического анализа [4]. Google Translate широко использует возможности нейронных сетей [5]. Наряду с этим есть системы, работающие только с одной языковой парой. В частности, при проектировании канадской системы METEO, выполняющей перевод метеопрогнозов с французского языка на английский и обратно, разра-

ботчики еще в 60-х годах XX в. основывались на предположении, что автоматизированный машинный перевод возможен только в условиях искусственно ограниченного языка, и добились успеха.

1. ОБЗОР ВОЗМОЖНОСТЕЙ СУЩЕСТВУЮЩИХ СИСТЕМ МАШИННОГО ПЕРЕВОДА

В практическом аспекте научно-исследовательские школы машинного перевода, на сегодняшний день, достигли высокого уровня перевода простых предложений, используя различные языковые пары, и они могут осуществлять перевод технических корпусов текстов. Лидерами являются афро-американская корпорация SYSTRAN, корпорации Google с продуктом Google Translate, Yandex и их система Yandex Translate, на российском рынке корпорация PROMT. В 2002г. был проведен эксперимент, направленный на выявление уровня точности передачи фактического содержания в языковых парах (ЯП) немецкий язык – английский язык и французский язык – английский язык; СМП SYSTRAN показал 50 % по точности перевода и 90% по уровню общего понимания; практические исследования МП ЯП испанский язык – английский язык в 2007 г. достиг у SYSTRAN 70% точности [6]. Несмотря на то, что с момента эксперимента прошло уже 15 лет, уровень точности перевода увеличился незначительно, при этом точность передачи фактического содержания не стабильна у разных ЯП и требует дополнительной проработки[6].Наибольшую сложность вызывают языковые пары, связанные с переводом на русский язык. Косвенно эту тенденцию можно проследить статистически: по аналитическим данным компании PROMT – 30% российских пользователей обращаются к ЯП русский – английский для перевода личной и деловой коммуникации и только 10% хотели бы узнать, что им ответили[7].

Проблема передачи фактического содержания в МП при условии небольшого количества слов (до 30) практически решена, но даже на таких небольших текстах не решена проблема смысловой передачи содержания и адекватной передачи идиоматических языковых средств. Исследователи университетов Georgetown University – IBM (Leon E. Dostert, Paul Garvin), University of London (Andrew D. Booth, Leonard Brandwood), the University of Texas (Eugene Pendergraft) придерживаются мнения, что машина не способна передавать смысловые нюансы. Однако в настоящее время появляется все больше и больше работ, посвященных роли субъективного содержания высказываний в МП и решению проблем с наличием «искусственного шаблонного стиля», который появляется при МП[5, 8]. Переход на стиль более субъективно ориентированных сообщений приводит к тому, что системы почти «хором» начинают давать сбой и производить комические сообщения, способные полностью исказить смысл, закладываемый адресантом. Появляется необходимость постобработки текстов, что при оперативном общении (например, деловая переписка), когда необходимо ответить адресату прямо «здесь и сейчас» и когда передача смысловых оттенков от выражения опасений, надежд и степеней уверенности имеет принципиальное значение, становится существенным препятствием.

По данным российского форума переводчиков (2016) и аналитическим данным агентства Common Sense Advisory (2015) до 20 % всей переписки составляет деловая коммуникация. Данная область имеет специфику, связанную с точностью и адекватностью перевода фактического и смыслового содержания, которая и обусловливает межкультурное взаимопонимание и, как следствие, ведет к принятию правильных решений и соблюдению норм делового общения.

Проверка качества смыслового содержания при переводе системами SYSTRAN, PROMT, Google Translate и Yandex Translate на 50 выражениях для ЯП русский язык – английский язык показывает их несостоятельность; при этом, на сегодняшний день, для деловой переписки человеку необходимо знать оба языка. Примеры типичных ошибок, связанные с модальным окрасом и идиоматическими соответствиями, ведущими к искажению смысла, приведены в табл. 1.

Таблица 1

Анализ МП с точки зрения модального окраса и идиоматических соответствий (подчеркиванием выделены смысловые ошибки в переводе)

Высказывания для перевода на русском языке/ правильный перевод	***SYSTRAN***	***PROMT***	***Google Translate***	***Yandex Translate***
Я хотел бы вернуться к теме совещания на прошлой неделе = I would like to touch base on the meeting last week	I would want to return to the theme of conference last week.	I would like to return to a meeting subject last week.	I would like to return to the theme of the meeting last week.	I would like to return to the topic of the meeting last week.
Я просмотрел доклад прошлым вечером, он еще совсем сырой. = I looked over the report last night, it is far from done.	I examined the report to the past in the evening, it also entirely damp.	I have checked the report last evening, it still absolutely crude	I looked through the report last night, he is still very raw.	I watched the report last night, he's still quite raw

Высказывания для перевода на русском языке/ правильный перевод	*SYSTRAN*	*PROMT*	*Google Translate*	*Yandex Translate*
Я покажу, где у нас подводные камни = I will point out where our weak spots are.	I will show, where we have the underwater stones.	I will show where we have reefs.	I'll show you where we pitfalls.	I'll show you where we have pitfalls.
Вкратце, у нас будут потери = The bottom line is we will not break even.	Briefly, we will have the losses.	In brief, we will have losses.	In short we are losing.	In short, we will have losses.

2. МЕТОДОЛОГИЯ ПЕРЕВОДА ТЕКСТОВ ПРИ ДЕЛОВОЙ КОММУНИКАЦИИ

Данная проблема может быть решена путем применения функциональной схемы, состоящей из системы упрощения текста, анализа семантических признаков и фактического содержания (см. рис. 1).

Данная схема демонстрирует гибридную модель СМП применительно к деловым текстам и ориентирована на режим онлайн переписки, когда максимальное количество слов для одного акта коммуникации достигает 25-30 слов. Такое ограничение способствует увеличению качества перевода и большей однозначности выражения мысли за счет использования ограниченного набора структур предложений при деловой коммуникации. В структурно-грамматическом и семантическом аспектах данная схема базируется на теории «смысл-текст», которая применительно к МП была разработана И.А. Мельчуком в сотрудничестве с А.К. Жолковским, Ю.Д. Апресяном, А.В. Гладким и Л.Н. Иорданской [9] на основе достижений структурной семантики (А. Богуславский 1966, Charles J. Fillmore 1968, Lakoff 1966, Weinreich 1956, A. Wierzbicka1969 и др.).

В функциональной модели, разработанной И.А. Мельчуком и его последователями, указывается на то, что переход от смыслов к текстам и обратно может быть выполнен за счет моделирования языковых познаний говорящих (linguistic competence в терминологии Н. Хомского), а не действительных актов речевого общения (linguistic performance) [9]; см. также [10, 11]. При этом языковые познания говорящих определяются границами их языка в терминологии Л. Витгенштейна [12].Данное утверждение подтверждает тот факт, что для машинного перевода априори характерен механизм формализации семантических значений и, как следствие, формирования ограниченного набора разновидностей моделей языковых структур, максимально описывающих языковые познания говорящих на определенный момент времени в виде глубинных и поверхностных языковых структур. Использование данных корпусной лингвистики и статистического анализа позволяет постепенно увеличивать «охват» контекстов употребления языковых структур. В этом смысле МП всегда будет отставать от темпа развития языка в действительности, ведь согласно В. Гумбольдту, язык – это живой организм, который развивается по своим внутренним законам. В представленной на рис. 1 схеме перевода текстов деловой коммуникации предпринимается попытка «оживить» искусственность шаблонного стиля МП с помощью комплексного использования семантических признаков, выражающих субъективную [13-15] или прагматическую установку субъекта к содержанию сообщения [11].

2.1. Вопросы алгоритмизации применения классификации по семантическим признакам

В настоящее время сформулированы и формализованы метаданные семантических (онтологических, эпистемических и аксиологических) признаков, выражающих отношение автора к своему тексту см. [16].

В силу языковой асимметрии одно и то же языковое средство может выступать в качестве модального показателя разных признаков, что объясняет их синкретизм при реализации в тексте. Онтологические признаки лежат в основе, и на них последовательно накладываются эпистемические и аксиологические признаки. Данный механизм означает, что одно и тоже языковое средство может реализовывать одновременно несколько признаков [17, 18].

Выделяемые семантические значения реализуются определенным набором языковых средств. Данные языковые средства можно обозначить числовыми данными. Если обозначить действительное положение дел «1», а значение нереализованности – «-1», то сфера эпистемических признаков окажется в промежутке от «0» до «1», если речь идет о возможном положении дел, и в промежутке от «0» до «-1», если актуализируется значение нереализованности. В связи с тем, что эпистемические признаки наслаиваются на онтологические признаки, данные показатели низкой, средней и высокой степеней уверенности будут соответствовать показателям «0,1»–«0,5»–«0,9» относительно действительного положения дел и «-0,1»-«-0,5»-«-0,9» относительно значения нереализованности. Аксиологические признаки относятся к будущему положению дел и соответствуют показателю «1». При графическом представлении возможно построение диаграмм, показывающих развитие модального содержания текста [16].

Например, для немецкого языка, который также как и английский имеет жесткую структуру предложений, – «Knecht **haette** seine Ernennung, diese letzte und hoechste seiner Berufung, **recht wohl** auch selbst **erraten** oder mindestens als **moeglich**, vielleicht als-**wahrscheinlich erkennen koennen**; **doch** ueberraschte, **ja** erschreckte sie ihn auch diesmal. Er **haette** es **sich denken koennen**, sagte er sich nachtraeglich» [19, с. 298], получим на основе анализа метаданных следующий модальный окрас фрагмента из произведения Германа Гессе (рис. 2).

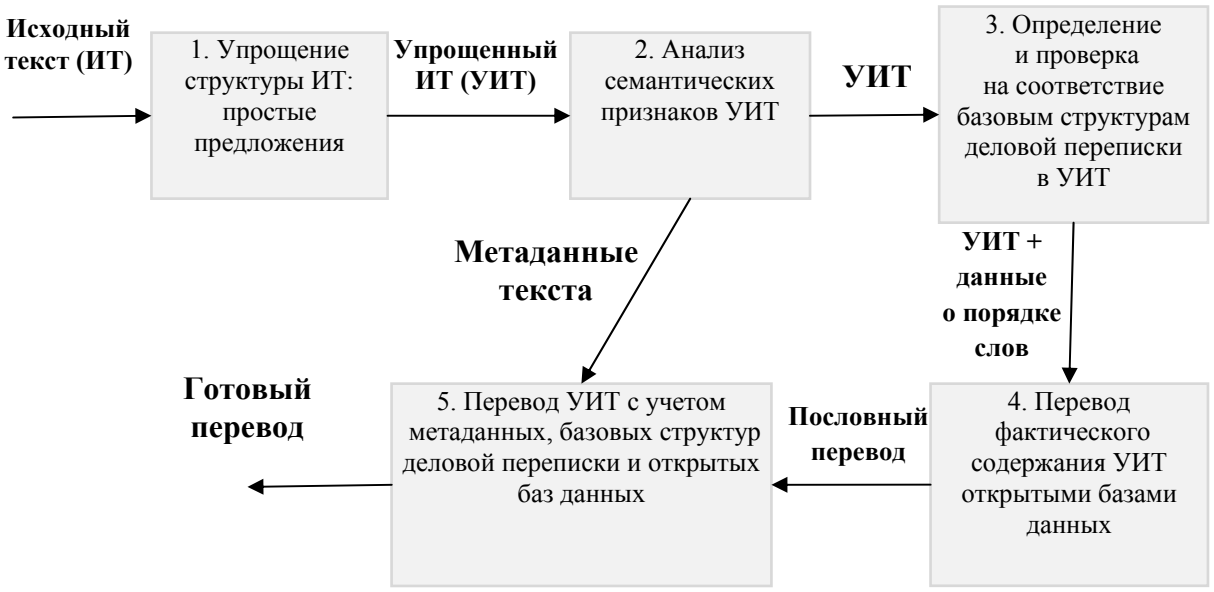


Рис. 1. Функциональная схема перевода текстов деловой коммуникации

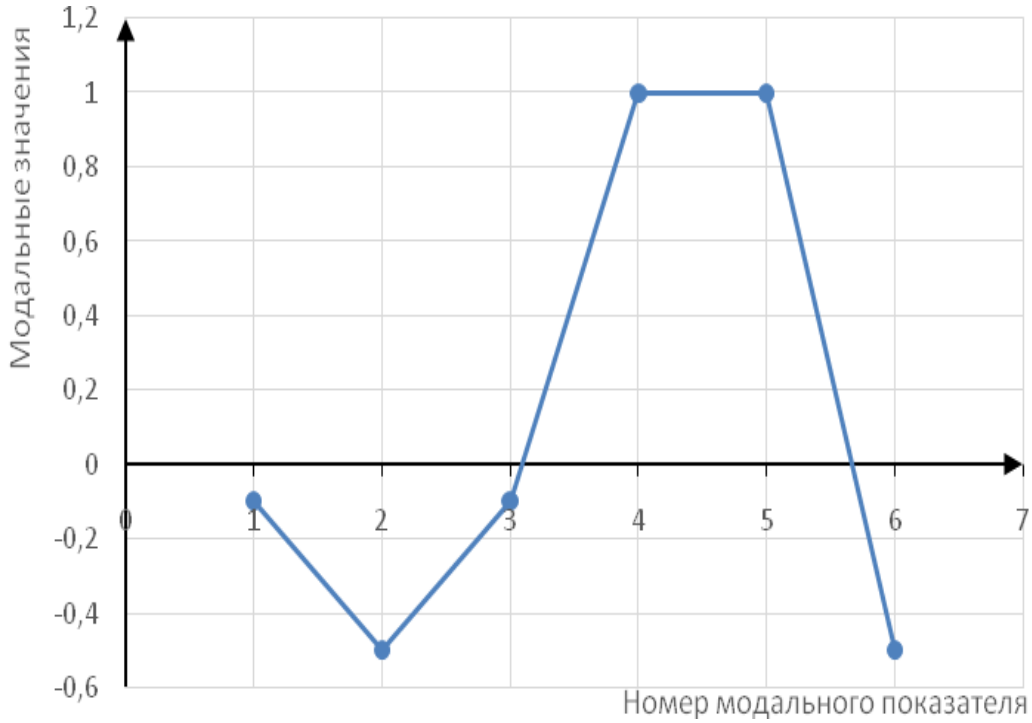


Рис. 2. Модальный окрас фрагмента из произведения Германа Гессе

Примечание: Соответствия номеров модальных показателей: 1 – haette recht wohl erraten, 2 – haette moeglich erkennen koennen, 3 – haette wahrscheinlich erkennen koennen, 4 – doch ueberraschte, 5 – ja erschreckte, 6 – haette sich denken koennen

Для приведенных семантических признаков с помощью метода частотного анализа были выявлены маркеры соответствующих семантических признаков. Эти данные были формализованы и уточнены для английских текстов. Метаданные являются общей закономерностью для русского и английского языков, см. табл. 2.

Исходя из данных таблицы, можно сделать вывод о том, что семантические признаки предположения, желания и рекомендации выражаются ограниченным набором фреймов, который характерен для русского и английского языков. А именно, значение предположения выражается 2-я фреймами, при этом маркерами выступают модальные слова и смысловые глаголы предположительной семантики; значение желания актуализируется 4-я фреймами, в качестве маркеров выступают смысловые глаголы в сослагательном наклонении в личной и безличной формах, смысловые глаголы со значением желания и глагольные формы побудительного наклонения; и наконец, значение рекомендации выражается 2-я фреймами, на которые указывают маркеры смысловых глаголов со значением рекомендаций и форма вопросительного предложения.

Таблица 2

Фреймы актуализации семантических признаков для русского и английского языков

Вид семантических признаков / фреймы для русского и английского языков					
Предположение					
Фрейм 1 Русский язык					
N1	Маркер семантического признака – модальное слово		Vf	N4	
Менеджер	0.9 скорее всего, точно; во что бы то ни стало		выполнит	заказ	
Менеджер	0.5 возможно		выполнит	заказ	
Менеджер	0.1 вряд ли, едва ли		выполнит	заказ	
Фрейм 1 Английский язык					
N1	Маркер семантического признака – модальное слово		Vf +V	N4	
The manager	0.9 most likely		will fill	the order	
The manager	0.5 probably		will fill	the order	
The manager	0.1 will un likely		fill	the order	
Фрейм 2 Русский язык					
N1	Маркер семантического признака – смысловой глагол Vf	Conj.	N1	Vf	
Менеджер	0.9 уверен	что	директор	заплатит	
Менеджер	0.5 предполагает	что	директор	заплатит	
Менеджер	0.1 сомневается	что	директор	заплатит	
Фрейм 2 Английский язык					
N1	Маркер семантического признака – смысловой глагол Vf	Conj.	N1	Vf+V	
The manager	0.9 is sure	that	the director	will pay.	
The manager	0.5 suggests	that	the director	will pay.	
The manager	0.1 doubts	that	the director	will pay.	
Желание					
Фрейм 1 Русский язык					
N1 (Ppron)	Маркер семантического признака – смысловой глагол Vf (subjunctive)	Conj.	N1 (Ppron)	Vf	N4
Я	хотел бы	чтобы	он	сделал	заказ
Фрейм 1 Английский язык					
N1 (Ppron)	Маркер семантического признака – смысловой глагол Vf (subjunctive)	Conj.	N 4 (Ppron)	Vf	N4
I	would like	___	him	to make	the order
Фрейм 2 Русский язык					
Маркер семантического признака-смысловой глагол V (subjunctive)	Short Adj.	Conj.	N 1 (Ppron)	Vf	N4
Было бы	здорово	если бы	он	сделал	заказ

Вид семантических признаков / фреймы для русского и английского языков					
Фрейм 2 Английский язык					
N1 + Маркер семантического признака- смысловой глагол V (subjunctive)	Short Adj.	Conj.	N 1 (Ppron)	Vf	N4
It would be	good	if	he	made	the order.
Фрейм 3 Русский язык					
N1 (Ppron)	Маркер семантического признака – смысловой глагол Vf	Conj.	N 1 (Ppron)	Vf	N4
Я	надеюсь	что	он	сделает	заказ.
Фрейм 3 Английский язык					
N1 (Ppron)	Маркер семантического признака – смысловой глагол Vf	Conj.	N 1 (Ppron)	Vf	N4
I	hope	that	he	will make	the order.
Фрейм 4 Русский язык					
Маркер семантического признака - смысловой глагол V (imperative)	V	N4			
Давайте	сделаем	заказ			
Фрейм 4 Английский язык					
Маркер семантического признака - смысловой глагол V (imperative) + Ppron (N4)	V	N4			
Let's	make	the order.			
Рекомендация					
Фрейм 1 Русский язык					
Ppron	Маркер семантического признака – смысловой глагол Vf	V		N4	
Я	рекомендую / советую	сделать		заказ	
Фрейм 1 Английский язык					
Ppron.	Маркер семантического признака – смысловой глагол Vf	Gerund (Ving)		N4	
I	Recommend/suggest	making		the order.	
Фрейм 2 Русский язык					
Маркер семантического признака –Interrog.question + subjunctive	Ppron.(N3)	Negation +V		N4	
Почему бы	вам	не сделать		заказ?	
Фрейм 2 Английский язык					
Маркер семантического признака –Interrog.question + Vf + Negation	Ppron.(N3)	V		N4	
Why don't	you	make		the order?	

Список сокращений: V – глагол; Vf – спрягаемый глагол; Gerund (Ving) – герундий; N – существительное, N1 – существительное в именительном падеже; N2 – существительное в родительном падеже, N3 – существительное в дательном падеже, N4 – существительное в винительном падеже, N 5 – существительное в творительном падеже, N 6 – существительное в предложном падеже; Prep. – предлог; Ppron. –личное местоимение; Negation – отрицание; Interrog.sentence – вопросительное предложение; imperative – побудительное наклонение; subjunctive - сослагательное наклонение; ShortAdj. – краткое прилагательное

2.2. Упрощение текста

В основе упрощения сложных предложений лежит механизм развертывания фразы для английского и русского языков; он сводится к обнаружению основных маркеров (союзов), позволяющих разбивать сложное предложение на простые предложения. В табл. 3 приведен фрагмент таблицы о соответствиях сочинительных и подчинительных союзов в русском и английском языках.

Для улучшения качества перевода будем использовать алгоритм перевода простых предложений в направлении от глагола, традиционно постулируемого в лингвистике как ядерная часть предложения, которая определяет валентностные характеристики и закладывает основу разворачивания мысли. Для деловой переписки было выделено 350 базовых смысловых глаголов, достаточных для обеспечения деловой коммуникации, а также установлено 50 фреймов для русского языка и выявлены соответствия для перевода данных фреймов на английский язык. Фрагмент соответствий приведен в табл. 4.

Таблица 3

Соответствие соединительных сочинительных союзов в русском языке союзам в английском языке

Соединительные сочинительные союзы: и, а, но, да (=и), или, либо, ни–ни, то–то	
Предложения на русском языке	Предложения на английском языке
Мы разрабатываем модели **и** готовим смету на продукт.	We are developing a model **and** preparing an estimate for the product.
Мы разрабатываем модели, **а** наши эксперты готовят смету затрат.	We are developing a model, **and** our experts are preparing a cost estimate.
Мы разработали новую модель, **но** еще необходимо продумать схему выхода на рынок продаж.	We have developed a new model, **but** it is still necessary to devise a scheme for market sales.
Мы выпустили в продажу новую модель, **да** создали пару новых модификаций по старой модели.	We have released a new model, ___ we create a couple of new modifications according to the old model.
Мы должны выпустить новую модель на рынок к началу года, **или** мы можем начинать продажи прямо с завтрашнего дня.	We should release the new market model by the beginning of the year, **or** we could start selling first thing tomorrow.
Мы можем ориентироваться на крупного российского клиента **либо** выходить на международный рынок.	We can focus on a major Russian customer **or** enter the international market.
В этой ситуации мы не сможем **ни** цену понизить, **ни** качество значительно улучшить.	In this situation, we will not be able to lower the price **nor** significantly improve the quality.
Цены **то** растут, **то** падают.	The prices rise **and then** fall.

Таблица 4

Фреймы базовых смысловых глаголов деловой переписки для русского языка и их соответствие для английского языка

V + N4	
Применять что-то	To apply something
Просмотреть что-то	To look over something
V + Prep.N5 + Prep.N4	
Извиняться перед кем-то за что-то	To apologize to someone for something
V + N4 + Prep.N4+ Prep.N4	
Выставлять цену на что-то для кого-то	To set the price of something for someone
V + Prep.N2	
Идти в направлении чего-то	To go to
V + Prep.N2	
Расти за счет чего-то	To grow at the expense of something
V +N4 + N3 + Prep.N4	
Выставлять счет кому-то за что-то	To bill someone for something
V+ Prep.N6	
Появляться в чем-то	To appear somewhere
Присутствовать на чем-то	To attend something
Учиться в чем-то	To study at
Участвовать в чем-то	To participate in something

V + N4	
Переживать что-то	To survive something
V + N4 + Prep.N6	
Основывать что-то на чем-то	To base something on something
V + Prep.N6	
Договориться о чем-то	To agree on something
Сожалеть о чем-то	To regret something
Нуждаться в чем-то	To need something
V + N4 + Prep.N6	
Просить кого-то о чем-то	To ask someone about something
V + Prep.N4	
Прибывать в/ на что-то	To arrive at the meeting
Верить во что-то	To believe in something
Приходить куда-то	To come to something
Надеяться на что-то	To hope for something
Указывать на что-то	To specify something
V + Prep.N6	
Пребывать где-то/ на чем-то	To be at the meeting, in the office
Происходить где-то/ на чем-то	To happen in/at/on something
V + Prep.N5	
Колебаться между чем-то	To fluctuate between something
V + Prep.N3	
Вернуться к теме	To touch base on something
V + N3 + N 5 + Prep.N2	
Помогать кому-то чем-то/ с помощью чего-то	To help someone with something
V + N4 + Prep.N5	
Ассоциировать кого-то с кем-то/ чем-то	To associate someone with someone/ something

3. ПРАКТИЧЕСКОЕ ИСПОЛЬЗОВАНИЕ МЕТОДА МП ТЕКСТОВ ДЕЛОВОЙ КОММУНИКАЦИИ

Использование описанного метода сводится к последовательности 8 шагов. Для пояснения рассмотрим их применение на примере предложений из табл. 1: «Я просмотрел доклад, который ваши специалисты выслали мне на прошлой неделе, но он еще совсем сырой для обсуждения, поэтому советую вернуться к теме совещания на прошлой неделе».

Шаг 1. *Определение типа предложения.*

Для этого ищем знаки, способные показать тип предложения (вопросительные знаки). Наличие или отсутствие вопросительных знаков позволяет определить тип предложения – утвердительный или вопросительный – и, как результат, линейный (при утверждении) и нелинейный порядок слов в предложении (вопрос как тип инверсии в целевом языке).

В приведенном примере утвердительный тип предложения и, следовательно, линейный порядок слов.

Шаг 2. *Поиск и выявление простых предложений.*

2.1. Ищем союзы (см. табл. 3) и находим соответствия в целевом языке. Союзы выступают маркерами границ простых предложений в составе сложного предложения. В рассматриваемом примере 2 союза: *но (сочинительный противительный союз)*, *поэтому (следственно-присоединительный союз)* – и, следовательно, 3 простых предложения:

а) Я просмотрел доклад, который ваши специалисты выслали мне на прошлой неделе;

б) предложение с сочинительным противительным союзом: **но** он еще совсем сырой для обсуждения;

в) предложение со следственно-присоединительным союзом: **поэтому** советую вернуться к теме совещания на прошлой неделе».

2.2. Далее в полученных структурах предложений ищем союзные слова, определяем тип связи и находим соответствия в целевом языке.

В рассматриваемом примере 1 союзное слово, которое выступает маркером придаточного определительного, а именно: я просмотрел доклад, **который** ваши специалисты выслали мне на прошлой неделе.

Следовательно, при переводе получаем следующую схему (схема 1).

Шаг 3. *Определение синтаксических позиций в простых предложениях (находим подлежащее и сказуемое, второстепенные члены) и находим соответствующую синтаксическую структуру в целевом языке.*

Для примера приводим синтаксический разбор предложения со следственно-присоединительным союзом *поэтому* (схема 2).

Шаг 4. *Морфологический анализ основных членов предложения: сказуемого и подлежащего – и поиск соответствий в целевом языке.*

Для примера рассмотрим подлежащее и сказуемое в двух первых простых предложениях:

а) Я просмотрел доклад,

б) который ваши специалисты выслали мне на прошлой неделе.

4.1. Находим словоформы подлежащего в двух предложениях и проводим морфологический анализ данных словоформ: находим реляционные морфемы (окончание) для определения числа (схема 3).

4.2. Находим словоформы сказуемого в двух предложениях и ищем частицу «не». Наличие или отсутствие частицы «не» определяет тип глагольной формы – утвердительный или отрицательный. При отрицании в целевом языке появляется вспомогательный глагол.

4.3. Проводим морфемный анализ словоформ сказуемого, находим реляционные морфемы: суффиксы и окончания – и определяем их грамматические значения; при этом категория числа словоформы сказуемого зависит от категории числа словоформы подлежащего (выявлена на предыдущем этапе), а именно (см. схему 4).

4.4. Согласуем грамматические значения реляционных морфем подлежащего и сказуемого в языке-источнике с грамматическими значениями в целевом языке, а именно (см. схему 5).

4.5. Ищем словоформы глаголов в базе данных табл. 4 и находим соответствующую модель управления в целевом языке (см. схему 6).

4.6. Переводим на целевой язык главных членов предложения, учитывая полученные данные шагов с 1-го по 4-й (схема 7).

Шаг 5. Поиск маркеров семантических признаков предположения, желания, рекомендации по данным таблицы 2 и определение соответствующего маркера в целевом языке.

В рассматриваемом примере есть маркер семантического признака рекомендации – смысловой глагол «советую».

В целевом языке этот признак соответствует структуре «suggest+Ving».

Следовательно, при переводе получаем следующее словосочетание (схема 8):

Шаг 6. Ищем устойчивые выражения из базы фреймов деловой коммуникации по данным табл. 4 и находим соответствующие фреймы в целевом языке (схема 9).

Выявление фрейма из базы фреймов деловой коммуникации обеспечивает смысловое соответствие этого фрейма при переводе на английский язык; база фреймов выступает фильтром ввода информации, а именно устанавливает ограничения на использование смысловых фреймов деловой переписки и, как следствие, повышает уровень смыслового соответствия на языке перевода.

Схема 1

Язык – источник	Целевой язык
Я просмотрел доклад, **который** ваши специалисты выслали мне на прошлой неделе	Я просмотрел доклад, **that** ваши специалисты выслали мне на прошлой неделе
но он еще совсем сырой для обсуждения	**but** он еще совсем сырой для обсуждения
поэтому советую вернуться к теме совещания на прошлой неделе	**so** советую вернуться к теме совещания на прошлой неделе

Схема 2

Порядок слов характерный для русского языка						
1	2	3	4	5	6	7
союз	имплицитно актуализированное подлежащее	сказуемое	предлог	дополнение	предлог	обстоятельство
поэтому	**----**	**советую вернуться**	**к**	**теме совещания**	**на**	**прошлой неделе**
Порядок слов характерный для английского языка						
союз	подлежащее	сказуемое	предлог	дополнение	отсутствие предлога	обстоятельство

Схема 3

Подлежащее 1 Словоформа «я»	Реляционные морфемы	Нулевое окончание
	Грамматическое значение в языке-источнике	1 лицо единственного числа
Подлежащее 2 Словоформы «специалисты»	Реляционные морфемы	Окончание «ы»
	Грамматическое значение в языке-источнике	3 лицо множественного числа

Схема 4

Сказуемое 1 Словоформа **«просмотрел»**	Реляционные морфемы	Суффикс «л»	Нулевое окончание
	Грамматическое значение в языке-источнике	Прошедшее время	1 лицо единственного числа
Сказуемое 2 Словоформа **«выслали»**	Реляционные морфемы	Суффикс «л»	Окончание «и»
	Грамматическое значение в языке-источнике	Прошедшее время	3 лицо множественного числа

Схема 5

Язык-источник: реляционная морфема/ грамматическое значение		Целевой язык реляционная морфема/ грамматическое значение	
Подлежащее 1			
Нулевое окончание	1 лицо единственного числа	Нулевое окончание	1 лицо единственного числа
Подлежащее 2			
Окончание «ы»	3 лицо множественного числа	Окончание «s»/ «es»*	3 лицо множественного числа
Сказуемое 1			
Суффикс «л»	Прошедшее время	Суффикс «ed»**	Простое прошедшее время: 2 форма глагола (прошедшее неопределенное)
Нулевое окончание	1 лицо единственного числа	Нулевое окончание	1 лицо единственного числа
Сказуемое 2			
Суффикс «л»	Прошедшее время	Суффикс «ed»**	Простое прошедшее время: 2 форма глагола (прошедшее неопределенное)
Окончание «и»	3 лицо множественного числа	Нулевое окончание	3 лицо множественного числа

* кроме супплетивных и аблативных форм образования множественного числа.
** кроме образования временных форм у неправильных глаголов

Схема 6

Словоформа в языке-источнике	Язык-источник	Целевой язык
просмотрел	V +N4	look over something
выслали	V +N4	send something

Схема 7

Словоформа в языке-источнике	Перевод в целевом языке
Я просмотрел	I looked over
специалисты выслали	experts sent

Схема 8

Язык-источник	Целевой язык
советую	suggest +Ving

Схема 9

Язык-источник	Целевой язык
Я **просмотрел доклад**	looked over the report
который ваши **специалисты выслали мне** на прошлой неделе	experts sent me
но **он еще совсем сырой для обсуждения**	it is far from done
	for discussion
поэтому **советую вернуться к теме совещания** на прошлой неделе	suggest touching base
	on the meeting

Шаг 7. *Поиск и перевод слов, не вошедших в структуру предложений деловой коммуникации в виде маркеров.*

На данном этапе выполняется перевод фактического содержания, который не был учтен на предыдущих этапах, для чего используются электронные словари. Слова для перевода электронными словарями выделены жирным шрифтом:

«**Я** просмотрел доклад, который **ваши** специалисты выслали мне **на прошлой неделе**, но он еще совсем сырой для обсуждения, поэтому советую вернуться к теме совещания **на прошлой неделе**».

Шаг 8. *Дополнение перевода предложения переведенными словами фактического содержания в соответствии с данными шагов 3 и 4.*

В результате получаем следующий перевод на целевой язык:

I looked over the report that your experts sent me last week, but it is far from done, so I suggest touching base on the meeting last week.

В приведенной последовательности 8 шагов можно выделить обязательные и инвариантные шаги, работа которых проявляется в случае наличия

проверяемых на них признаков, а именно союзов и союзных слов на шаге 2 и маркеров семантических признаков предположения, желания и рекомендации на шаге 5.

ЗАКЛЮЧЕНИЕ

В настоящей статье были рассмотрены вопросы практической реализации использования базы данных фреймов семантических метаданных, базовых смысловых глаголов и идиоматических выражений для языковых пар русский язык – английский язык и вопросы, связанные с решением проблемы пословного перевода при автоматизированном переводе деловой коммуникации. Анализ последовательности шагов предлагаемого авторами алгоритма для машинного перевода показывает, что уровень смыслового соответствия на языке перевода улучшается по сравнению с переводами существующих систем машинного перевода.

Авторы предполагают, что применение описанного метода на обширном корпусе текстов позволит расширить область применения метода и проводить исследования по его адаптации для других языковых пар. При этом недостатком метода является необходимость построения для каждой языковой пары базы соответствий и структур предложений; устранением данного недостатка может стать получение методов их автоматизированного наполнения и структурирования.

СПИСОК ЛИТЕРАТУРЫ

1. Hutchins W.J. Concise history of the language sciences: from the Sumerians to the cognitivists: Machine translation: a brief history. – Oxford: Pergamon Press, 1995. – C. 431–445.
2. Henisz-Dostert B., Macdonald R.R., Zarechnak M. Machine translation. – New York: Mouton, 1979. – 265 C.
3. Costa-Jussa M.R., Fonollosa J.A.R. Latest trends in hybrid machine translation and its applications // Computer Speech & Language. – 2015. – Vol. 32, № 1. – P. 3–10.
4. Costa-Jussa M.R., Farrus M. Statistical machine translation enhancements through linguistic levels: A survey // ACM Computing Surveys. – 2014. – Vol. 46, № 3. – P. 1–28.
5. Wu Y. et al. Google's Neural Machine Translation System: Bridging the Gap between Human and Machine Translation. – Google.– 2016.
6. Aiken M. et al. An Evaluation of the Accuracy of Online Translation Systems // Communications of the IIMA. – 2009. – Vol. 09, № 04. – P. 66-84.
7. Соколова С. Российский рынок прекрасен, но его недостаточно. – 2016. – URL: https://www.dp.ru/a/2016/02/09/Svetlana_Sokolova/
8. Shachar M. et al. Motivating Personality-aware Machine Translation // The 2015 Conference on Empirical Methods on Natural Language Processing (EMNLP). - Haifa, Israel. - 2015. – P. 1102–1108.
9. Мельчук И.А. Опыт теории лингвистических моделей "смысл-текст": семантика, синтаксис. – М.: Языки русской культуры, 1999. – 345 с.
10. Morgan J.L. Conversational Postulates Revisited // Language. – 1977. – Vol. 53, № 2.– 277 p.
11. Austin J.L., Urmson J.O. How to do things with words. – Cambridge, Massachusetts: Harvard Univ. Press, 2009. – 168 p.
12. Wittgenstein L. Logisch-philosophische Abhandlung: Tractatus logico-philosophicus. – Frankfurt am Main: Suhrkamp, 2004. – 114 p. 13. Логический анализ языка: избранное 1988-1995. – М.: Indrik, 2003. – 695 с.
14. Doignon J.-P., Falmagne J.C. Knowledge spaces. – Berlin, New York: Springer, 1999. – 333 p.
15. Hintikka J. Knowledge and belief: an introduction to the logic of the two notions. – London: King's College London Publications, 2005. – 133 p.
16. Новикова А.В. Лингвистический анализ реализации возможных миров в художественном тексте: дис. канд. филол. наук. Пермский национальный исследовательский политехнический университет. – Пермь, 2010. – 170 с.
17. Ryan M.L. The Text as World Versus the Text as Game: Possible Worlds Semantics and Narrative Theory // Journal of Literary Semantics. – 1998. – Vol.3, № 27. – P. 137–163.
18. Goodman N., Elgin C.Z. Reconceptions in philosophy and other arts and sciences. – Indianapolis: Hackett Pub.Co, 1988. – 174 p.
19. Hesse H. Ausgewählte Werke. T.04. – Frankfurt am Main: Suhrkamp, 1994. – 660 p.

Материал поступил в редакцию 17.02.17.

Сведения об авторах

НОВИКОВА Анна Вячеславовна – кандидат филологических наук, доцент кафедры «Иностранные языки, лингвистика и перевод», Пермский национальный исследовательский политехнический университет
e-mail: novikova@yandex.ru

МЫЛЬНИКОВ Леонид Александрович – кандидат технических наук, доцент кафедры «Микропроцессорные средства автоматизации», Пермский национальный исследовательский политехнический университет
e-mail: leonid.mylnikov@pstu.ru

ИНФОРМАЦИОННОЕ ПИСЬМО И ПРИГЛАШЕНИЕ

МЕЖДУНАРОДНАЯ КОНФЕРЕНЦИЯ К 65-ЛЕТИЮ ВИНИТИ РАН

«ИНФОРМАЦИЯ В СОВРЕМЕННОМ МИРЕ»

Москва, 25-26 октября 2017 г.

подробная информация на сайте: http://www.viniti.ru

Главный организатор:

Всероссийский институт научной и технической информации Российской академии наук (ВИНИТИ РАН)

Соорганизаторы:

Российская академия наук
Федеральное агентство научных организаций
Российский фонд фундаментальных исследований
Министерство образования и науки РФ

Проблемно-тематическое направление конференции: современный издательский процесс, интеллектуальная собственность, научные библиотеки, информационное обеспечение научной и инновационной деятельности, информационные технологии для научной и библиотечной отрасли, информационная безопасность, международное сотрудничество и информационный обмен, инфометрия, классификации, стандартизация, образование для отрасли, экономика информации

Основные вопросы, предлагаемые к обсуждению:

● Популяризация научных знаний: Новые модели распространения научной информации
● Редакционно-издательская деятельность в цифровой среде: продукты и сервисы
● Издательские стандарты и технологии
● Перспективы развития книжного дела. Проекты и программы
● Взаимодействие цифровых и печатных ресурсов в научно-технической библиотеке
● Информационно-библиотечное обслуживание: сервисный подход
● Управление данными и навигация в современной научной библиотеке
● Научные библиотечные консорциумы – основные подписчики на научную литературу
● Перспективы развития национальных систем научно-технической информации
● Государственные проекты и программы поддержки информационного обеспечения научно-образовательной деятельности
● Тенденции развития региональных аналитических центров
● Информационное обеспечение экспертной деятельности. Использование информационно-аналитических систем для управления наукой и образованием
● Формальные и неформальные каналы развития современных научных коммуникаций

● Современные агрегаторы научной литературы открытого доступа как источник научной информации
● Машинная обработка данных и аналитические исследования: Приоритеты и сотрудничество
● Использование специальных сервисов компании CrossRef для идентификации научных публикаций
● Роль поисковых систем в современном издательском процессе
● Защита данных от несанкционированного использования. Маркеры безопасности. Политика безопасности открытых систем
● Вопросы достоверности и доверенности при обработке информационного потока
● Межгосударственный обмен научно-технической информацией на евразийском пространстве
● Информационное взаимодействие в рамках СНГ
● Международное партнерство при хранении и обработке больших массивов данных
● Современное состояние систем классификации знаний как инструмента индексирования и поиска данных по перспективным направлениям науки и критическим технологиям
● Современные библиометрические методы определения научных лидеров: Новые математические модели
● Анализ читательской аудитории научной литературы путем вебометрического анализа
● Подготовка специалистов в сфере научно-информационной деятельности
● Мастер-класс по работе с классификационными системами (УДК, ГРНТИ)
● Информация как источник цифрового капитала и фактор социальных изменений
● Информационная деятельность как фактор национальной экономики
● Новейшие бизнес-модели для публикаций открытого и закрытого доступа

На конференции планируются доклады представителей ведущих информационных центров и научно-технических библиотек России, СНГ и дальнего зарубежья.

В рамках юбилейной конференции состоится научно-практический семинар по классификационным системам «Перспективные направления научных исследований и критические технологии в классификационных системах». Предполагается проведение специализированных обучающих мероприятий по УДК индексированию. Запланировано заседание методического совета пользователей ГРНТИ и УДК. Участники конференции получат свидетельства о повышении квалификации.

Материалы конференции будут опубликованы в сборнике Трудов и на CD-ROM, основные – в сборнике **«Научно-техническая информация»**.

Доклады

Принимаются оригинальные работы, имеющие научное и прикладное значение, соответствующие тематическим направлениям конференции и НЕ ОПУБЛИКОВАННЫЕ ГДЕ-ЛИБО РАНЕЕ.

Предлагаемый доклад должен отвечать следующим требованиям:

1. Необходимо указать название доклада, фамилию, имя, отчество (полностью) авторов/соавторов, название организации, город, страну, выделить автора, который будет представлять доклад.

2. Необходимо наличие аннотации, раскрывающей содержание доклада. Размер аннотации - не более 850 знаков (включая пробелы).

3. Доклады принимаются только в электронной форме; тексты – в формате MS Word; схемы, диаграммы, фотографии, сканированные виды экранов и т. п. - в формате JPG. Объем доклада вместе с аннотацией, рисунками, приложениями и т.п. не более 10 страниц формата А4.

4. Доклад необходимо выслать по электронной почте до 11 сентября 2017 г. в адрес оргкомитета: conf@viniti.ru

Доклады, не соответствующие вышеуказанным требованиям,
НЕ РАССМАТРИВАЮТСЯ.

Программный комитет оставляет за собой право определять статус доклада (пленарный доклад, доклад, стендовый доклад), включать принятые доклады в те или иные секции.

Время для выступления: пленарные доклады – 15–20 мин., доклады на отдельных мероприятиях – до 10 мин. Доклады включаются в Труды на основании решения экспертов оргкомитета.

Контакты: 125190, Москва, ул. Усиевича, 20, ВИНИТИ РАН
Телефоны: 8 (499) 152 61 13, 8 (499) 155 42 52, 8 (499) 151 02 61. Факс 8 (499) 943 00 60
Интернет-сайт: http://www.viniti.ru Эл. почта: conf@viniti.ru

База данных (БД) ВИНИТИ РАН

Федеральная база отечественных и зарубежных публикаций по естественным, точным и техническим наукам, генерируется с 1981 г., обновляется ежемесячно, пополнение составляет около 1 млн документов в год. Тематическое наполнение соответствует реферативному журналу ВИНИТИ. Для поиска одновременно по всем или нескольким тематическим фрагментам генерируется единая Политематическая БД.

БД ВИНИТИ РАН в сети INTERNET

Сервер ВИНИТИ - http://www.viniti.ru – обеспечивает on-line доступ к Базе данных ВИНИТИ РАН круглосуточно без выходных.

На основе БД ВИНИТИ РАН предоставляются следующие услуги:

- Диалоговый поиск научно-технической информации **в режиме on-line**;
- **Демо-версия**, позволяющая ознакомиться с основными функциями поисковой системы, составом данных, формами представления документов и получить навыки работы с системой;
- **Поисковые эксперты** ВИНИТИ выполнят тематический поиск по разовым или постоянным запросам, а также окажут **консультационные услуги**.

БД ВИНИТИ РАН на CD-ROM

Любые наборы тематических фрагментов БД ВИНИТИ или их разделов за любой период с 1981 г., а также **проблемно-ориентированные выборки** из БД ВИНИТИ по актуальным направлениям научных исследований могут быть предоставлены на договорной основе **в поисковой системе (ИПС) "Сокол"**, работающей под управлением Microsoft Windows и обеспечивающей следующие возможности:

- **Чтение** документов в режиме последовательного просмотра или выборочно по оглавлению за весь период заказанной ретроспективы
- **Поиск** документов по автору, заглавию, источнику, ключевым словам или словосочетаниям, реферату, рубрикам, году издания, стране, языку и т.д. (всего более 20 признаков)
- **Словарь** системы поможет правильно подобрать термины для поиска и выбрать глубину их усечения.
- **Для уточнения поиска** можно дополнительно использовать год издания документа, язык текста документа, рубрики, шифры тематических разделов БД.
- Выполненные **запросы можно сохранять** для их последующего использования и/или редактирования.

125190, г. Москва, ул. Усиевича, 20, БД ВИНИТИ РАН.
Отдел взаимодействия с потребителями – (499) 155-45-25, (499) 152-58-81
E-mail: csbd@viniti.ru, sales@viniti.ru

WWW: http://www.viniti.ru