А. В. Добров

КОМПЬЮТЕРНЫЙ СЕМАНТИКО-СИНТАКСИЧЕСКИЙ АНАЛИЗ ЯЗЫКОВЫХ ОБОЗНАЧЕНИЙ ДЕЙСТВИЙ ИЛИ ДЕЯТЕЛЬНОСТИ ОРГАНОВ ГОСУДАРСТВЕННОЙ ВЛАСТИ[1]

Аннотация. В статье представлен опыт разработки семантических средств для распознавания обозначений деятельности или действий органов государственной власти с помощью лингвистического процессора AIIRE, включающего в себя также одноименную онтологию. Исследование осуществляется на материале новостных сообщений, полученных совместно с сотрудниками Центра технологий электронного правительства НИУ ИТМО. Данная работа является частью серии исследований «повестки дня», формируемой средствами массовой информации по тематике, связанной с развитием электронного правительства.

Ключевые слова. Автоматическое понимание текстов, онтологическая семантика, концептуальные отношения, лексическая неоднозначность.

Alexey V. Dobrov

SEMANTIC AND SYNTACTIC COMPUTER ANALYSIS OF LINGUISTIC DENOTATIONS OF ACTS OR ACTIVITIES OF PUBLIC AUTHORITIES

Abstract. The article describes the ongoing research aimed at creating semantic tools for recognition of denotations of public authorities acts or their activities with help of AIIRE natural language processor and ontology. The study is carried out on the material of news items, collected in collaboration with ITMO University Center for Electronic Government. This work is a part of a series of studies focused on the "Agenda" through the media resources on topics related to e-government development.

Keywords. NLU, ontological semantics, conceptual relations, lexical disambiguation.

[1] Работа выполняется в рамках проекта «Разработка инструмента опинионмайнинга и его апробирование на задачах обследования общественного мнения о деятельности органов власти» (НИР № 415825, Университет ИТМО).

Введение

В рамках проекта «Разработка инструмента опинион-майнинга и его апробирование на задачах обследования общественного мнения о деятельности органов власти» возникла задача выделения в текстах обозначений действий и различных видов деятельности органов государственной власти (далее ОГВ). Коллективом ООО «АИРЕ» была разработана система мониторинга, которая, совместно с сотрудниками Центра технологий электронного правительства НИУ ИТМО, была настроена для сбора данных о деятельности ОГВ.

Традиционные подходы к решению задачи выделения сущностей, основанные на шаблонах, оказались не вполне адекватными целям данного исследования, поскольку выделение шаблонов обозначений действий или различных видов деятельности ОГВ не даёт возможности в дальнейшем выполнять автоматическое выделение их оценок, и было решено использовать семантические технологии: лингвопроцессор AIIRE и одноимённую онтологию. В ходе компьютерного анализа языковых единиц, обозначающих действия или деятельность ОГВ, возникает необходимость выделения семантических отношений между понятиями указанной предметной области. Такие отношения позволяют строить семантические графы и разрешать неоднозначность различных видов (морфологическую, синтаксическую, лексическую).

В данной работе описываются методы компьютерного анализа языковых выражений, обозначающих действия или деятельность ОГВ, с использованием AIIRE.

Лишь в некоторых из работ, посвящённых исследованиям политической лексики, производится структурно-лингвистический анализ терминологии государственного управления. В этих работах основной акцент видится несколько смещённым в сторону обсуждения того, является ли данная терминология самостоятельной терминосистемой [Нгуен], как эта терминосистема соотносится с терминологией государственного управления, какова структура этой терминосистемы: как представлены её ядро и периферия, каковы особенности политической метафоры [Чудинов] и т. д.

Вопрос о терминологичности единиц политической лексики обусловлен широким распространением не свойственной терминам

синонимии, полисемии и омонимии. Согласно А. С. Герду, наименования учреждений следует относить не к терминам, а к идентификаторам [Герд, с. 4], но «неоднократные призывы филологов различать термины и номены во многом остаются на уровне абстрактных филологических требований, которые невозможно реализовать на практике» [Там же. С. 3].

Родо-видовые отношения между понятиями, стоящими за лексическими значениями единиц политической лексики, исследуются в работе Т. Н. Юдиной и А. В. Богомоловой [Юдина, Богомолова]. В этой работе рассматриваются теоретические вопросы организации компьютерной онтологии предметной области «Государственное управление».

Существующие в современной научной литературе сведения о семантических свойствах политической лексики, в особенности наименований ОГВ, не отражают фактического употребления различных языковых средств, выражающих действия или виды деятельности ОГВ. Как будет показано ниже, даже сами названия ОГВ часто строятся из таких средств. Среди них частотны отвлечённые процессуальные имена существительные (далее ПС), синтаксические и семантические свойства которых сами по себе на сегодняшний день нуждаются в исследовании.

ПС характеризуются явными особенностями уже на уровне так называемых присловных связей и требуют особого подхода при семантическом анализе синтаксических структур. Такой подход представлен в монографии В. П. Казакова [Казаков]. Вслед за Ю. Д. Апресяном В. П. Казаков отмечает, что ПС, образованные от глаголов (так называемые девербативы), наследуют семантические валентности этих глаголов [Апресян, с. 165]. Вместе с тем даже ПС остаются именами существительными, и потому, в отличие от глаголов, становятся «„опредмеченным“ наименованием — всё же действия, состояния, качества» [Золотова, с. 126]. Концепты, стоящие за значениями глагола и соответствующего ему девербатива, различны, вопреки традиционному в компьютерной лингвистике отождествлению этих концептов (так, концепты ‘бег’ и ‘бежать’ не дифференцируются в большинстве компьютерных онтологий).

В диссертационном исследовании Л. А. Вакарюк [Вакарюк] показано, что далеко не все ПС являются девербативами (ср. *акт*, *процесс*, *реверанс*, *лекция, навигация* и др.). При этом сохраняется

способность ПС непосредственно присоединять наречия времени и места (ср. *лекция вчера*, *навигация повсюду*) и присоединяться к лексемам процессуальной семантики (ср. *прервать спектакль*).

В. В. Богданов квалифицирует функцию девербатива в предложении как «непредметный аргумент», позицию которого могут занимать «отпредикатные существительные, инфинитивы, герундии и т. д.» [Богданов, с. 172]. ПС обозначают свёрнутые пропозиции, а «наличие пропозитивных существительных в некотором предложении свидетельствует о его полипропозициональном строении» [Там же. С. 173]. Следовательно, при наличии в словосочетании ПС это словосочетание может быть развёрнуто в предложение: ср. *продажа Минфином акций — Минфин продаёт акции*. При этом семантические отношения в словосочетании и соответствующем предложении идентичны или имеют взаимно-однозначное соответствие.

Метод межуровневого взаимодействия

Использование статистических эвристик для разрешения неоднозначности при автоматической обработке текста гораздо популярнее применения методов, основанных на правилах. Корпусные подходы (так называемые методы машинного обучения) имеют существенный успех в разрешении морфологической неоднозначности (например, Ян Хаджич и соавторы оценивают качество этих методов в 95% [Serial Combination …]). Корпусные методы более просты и объективны, чем методы, основанные на правилах, так как не требуют создания этих правил. Главный недостаток статистических эвристик состоит в том, что они не дают гарантии отсутствия ложноотрицательных результатов: некорректно исключённая версия морфологического анализа может привести к потере целостности всего синтаксического дерева.

Работа AIIRE основана на методе межуровневого взаимодействия, впервые предложенном Григорием Самуиловичем Цейтиным в 1985 г. [Цейтин]. Несмотря на то что метод был предложен уже тридцать лет назад, он до сих пор не был применён на практике ввиду сложности его алгоритмической реализации и отсутствия соответствующего программного обеспечения. Цель упомянутого метода — избавиться от искусственного разделения уровней лингвистического анализа и анализировать текст одновременно на всех

языковых уровнях. Такой подход позволяет устранять неоднозначность на более низких уровнях, используя правила более высоких уровней ещё до того, как неоднозначность более низких уровней успевает привести к так называемому комбинаторному взрыву.

Морфологическая неоднозначность может быть снята по результатам непосредственного синтаксического связывания (или не связывания) первых двух разобранных словоформ, в соответствии с ограничениями грамматики. Синтаксическая неоднозначность может быть снята при невозможности семантического связывания значений элементов синтаксического дерева [Dobrov].

Главный инструмент AIIRE для разрешения неоднозначности — онтология, содержащая в себе концепты (модели понятий), стоящие за значениями лексических единиц, и обеспечивающая возможность вычисления их семантических валентностей на основании тех отношений, которые установлены между концептами.

Концепты онтологии AIIRE и отношения между ними

Концепт в онтологии AIIRE — это набор атрибутов, где каждый атрибут представляет собой пару «отношение — объект». Для экономии вычислительных ресурсов в онтологии хранятся только прямые отношения, а обратные — вычисляются. Виды используемых отношений описаны в работе [Dobrov]. В онтологии AIIRE существуют строгие правила наследования и замещения отношений. Если один концепт наследует другой, то каждый атрибут наследуемого воспроизводится наследующим концептом, однако наследуемый атрибут может быть замещён другим атрибутом, если и отношение, и объект замещаемого атрибута наследуются или совпадают с отношением и объектом замещающего. Например, концепт ‘отрезок’ имеет атрибут <‘иметь размер’, ‘длина’>, в то время как его подкласс ‘период времени’ обладает замещением <‘иметь размер’, ‘длительность’>, что означает, что размер (длина) временного периода — это его длительность (концептом ‘длительность’ также наследуется концепт ‘длина’).

В систему анализа текстов AIIRE входит лингвопроцессор (далее ЛП) — программное средство, осуществляющее семантический анализ текста, то есть представление его семантики в виде семантического графа (далее СГ) на основании данных из онтологии и модулей

грамматики (морфологии и синтаксиса). ЛП производит семантические графы в качестве конечного представления текстовой семантики. Термин «семантический граф» используется в данной статье в широком смысле и не является синонимом термина «концептуальный граф», предложенного Дж. Совой [Sowa]. Рёбра СГ могут быть вершинами и иметь собственные рёбра, при этом вершины и рёбра СГ — концепты онтологии или наследующие их текстовые концепты.

Процедура нормализации СГ, то есть его приведения к единообразному виду, в котором обратные отношения заменяются прямыми, позволяет унифицировать семантические представления различных обозначений одних и тех же ситуаций (ср. *МФ продаёт акции*, *продажа акций МФ*; *продающий акции МФ*, *проданные МФ акции* и т. д.). В процессе нормализации концепты, не являющиеся корнями синонимических рядов (далее СР) — выделенными концептами, используемыми для идентификации СР, — заменяются на корни СР, что позволяет отождествлять, например, *продажа акций МФ*, *продажа акций Минфином*, *сбыт акций МФ* и т. д.

Названия ОГВ в онтологии

Компоненты названий ОГВ

Названия ОГВ, с точки зрения их разбора ЛП, можно разделить на три различных вида: полные названия (например, *Министерство финансов, Федеральная Антимонопльная Служба, Федеральное агентство связи*), слоговые аббревиатуры (*Минобр*, *Минобороны, Минэкономразвития*) и звуковые аббревиатуры (*ФАС*, *ФНС*, *МВД*).

Для корректного разбора ЛП естественно-языковых выражений все концепты, которые могут стоять за лексическими значениями единиц, составляющих эти выражения, должны содержаться в онтологии. Кроме того, в онтологии должны быть установлены отношения между этими концептами, позволяющие, с одной стороны, строить гипотезы о семантических связях между значениями лексических единиц в выражении и, с другой стороны, разрешать семантическую неоднозначность. Эти требования относятся и к названиям ОГВ. Вместе с тем название ОГВ, представленное сложным словосочетанием, обозначает единый концепт и часто является идиоматическим выражением. Онтология должна хранить идиомы и отношения между концептами, обозначаемыми словами, входящими в эти идиомы.

Например, название *Министерство здравоохранения* может быть разобрано ЛП только в том случае, если в онтологии обработаны концепты ‘министерство’ и ‘здравоохранение’ и определены концептуальные отношения, которые при разборе позволят ЛП связать эти концепты между собой. Для решения данной задачи в онтологии AIIRE для концепта конкретного ОГВ создается отношение ‘(о субъекте) осуществлять деятельность в предметной области’, направленное на концепт, представляющий собой вид деятельности ОГВ. Это отношение является частным случаем абстрактного отношения ‘(об объекте или процессе) принадлежать объекту или процессу’, обозначаемого сочетанием именной группы (далее ИГ) с несогласованным определением в родительном падеже (далее ИГРП). При этом концепт ‘министерство здравоохранения’ является подклассом концепта ‘министерство’ и наследует все его отношения, в том числе отношение ‘осуществлять деятельность в области…’, направленное на концепт ‘здравоохранение’, что позволяет ЛП разбирать словосочетание *министерство здравоохранения* как ‘министерство, осуществляющее деятельность в области здравоохранения’.

Аббревиатуры и проблема рода

В разборе ЛП аббревиатур имеются трудности, связанные с необходимостью определять род, который может не совпадать с родом, употребляемым в тексте. Склонение звуковых инициальных аббревиатур зависит от их опорного слова, то есть, например, если опорное слово — мужского рода, то и вся аббревиатура приобретает мужской род. В то же время если опорное слово — женского рода, но вся аббревиатура имеет морфологический облик слова, принадлежащего к мужскому роду, то указанное правило нарушается. К таким случаям относятся аббревиатуры, имеющие опорное слово женского рода и оканчивающиеся на согласный, например *ФАС* (*Федеральная Антимонопольная Служба*), *ЦИК* (*Центральная избирательная комиссия*). Опорное слово в таких случаях имеет женский род, но, так как аббревиатура оканчивается на согласный, она может употребляться в тексте как слово мужского рода (*ФАС произвёл проверку*, *ЦИК опубликовал результаты*).

Фактическое употребление аббревиатур в современном медиадискурсе вариативно, поэтому в подобных случаях в морфологическом словаре ЛП необходимо хранить два варианта рода аббревиатур.

Иногда AIIRE хранит в морфологическом словаре аббревиатуры в трёх экземплярах: мужского, женского и среднего рода. Данная ситуация типична для случаев омонимии: ср. *ФАС* (*Федеральная Антимонопольная Служба*) и *ФАС* (*Федеральное агентство связи*).

Названия ОГВ

Анализ синтаксических конструкций более 360 названий федеральных ОГВ показал, что в системе этих наименований используются всего два типа синтаксических структур: ИГРП (например, *Министерство финансов*, *Министерство иностранных дел*, *Минобр России*) и модифицированная предложным определением ИГ (далее ИГПО) (*Агентство по труду и занятости*).

Для разбора ИГРП ЛП осуществляет запрос, то есть процедуру поиска атрибута по абстрактному отношению, обозначаемому родительным падежом (далее РП), или любого подкласса этого отношения у концепта, обозначаемого ИГ в РП. При наличии такового проверяется, является ли концепт, обозначаемый ИГ в препозиции, подклассом или надклассом концепта, на который направлено отношение. Если связь найдена, то концепты, обозначаемые ИГ, семантически связаны найденным отношением, а концепт, обозначаемый ИГ в препозиции, замещён объектом отношения. Например, при разборе *министерство здравоохранения* ЛП обнаруживает отношение ‘(о предметной области) быть областью, в которой осуществляет деятельность субъект’, направленное на концепт ‘министерство здравоохранения’.

Аналогично ИГРП, ЛП обрабатывает ИГПО путём запроса отношения ‘обладание аргументом предложного отношения’.

В ЛП AIIRE используется внутренняя вспомогательная база данных идиом и соответствующих им СГ. Для её построения производится анализ каждой идиомы, и результат анализа сохраняется в базе идиом. В ходе разбора текста, если СГ содержит часть, соответствующую СГ идиомы, производится замена этой части на соответствующий концепт из базы идиом. Некоторые идиомы могут содержать в своем составе другие идиомы. Так, название *Министерство по делам гражданской обороны, чрезвычайным ситуациям и ликвидации последствий стихийных бедствий* содержит такие выражения, как *гражданская оборона*, *чрезвычайная ситуация* и *стихийное бедствие*, каждое из которых рассматривается как идиома.

Обработка сложных идиоматических выражений требует наличия в базе идиом составных частей, являющихся идиомами. Поэтому построение базы идиом выполняется последовательно, от наименьшего количества словоформ в выражении к наибольшему.

Концепты действий, состояний и различных видов деятельности

Согласно Лингвистическому энциклопедическому словарю [Маслов], глаголы могут быть динамическими (предельными и непредельными) или статическими. Предельные динамические глаголы характеризуются завершённостью процесса, то есть обозначают действия субъекта, которые предполагают завершение. Непредельные динамические глаголы не предусматривают предела в протекании процесса и обозначают деятельность субъекта. Статические глаголы обозначают состояния (состоянием может быть, в частности, отношение субъекта к чему-либо). С концептуальной точки зрения, в онтологии AIIRE деятельность — частный случай состояния (субъект находится в состоянии осуществления какого-либо вида деятельности); далее действия, виды деятельности или состояния будут обозначаться ДС. Для предельных динамических глаголов в онтологии выделяются концепты, соответствующие процессу (несовершенный вид) и завершению процесса (совершенный вид). Внесение в онтологию концепта ПС покрывает все возможные варианты обозначения действия (ср. *суд приговорил*, *суд приговаривает*, *приговор суда*). Все три концепта должны сохранять валентности, в том числе связи с предлогами, что означает сохранение параллелизма трёх классификаций. Для этого ДС подразделяются на направленные (переходные глаголы) и ненаправленные, а также адресованные (управляющие дативом) и неадресованные. Для непредельных динамических глаголов указывается концепт, соответствующий завершению, который, в свою очередь, является действием (например, деятельность — *работать*, завершение — *доработать*, которому соответствует *дорабатывать*). Для статических глаголов в онтологии достаточно указания процесса, соответствующего состоянию.

При формировании параллельных классификаций в онтологии AIIRE созданы концепты, соответствующие различным семантическим классам значений глаголов и предлогов. В ходе вычисления СГ

для выражения с предложными группами с помощью правил наследования атрибутов производится необходимое связывание и снятие неоднозначности. Так, классы глаголов перемещения, передачи информации, наблюдения, соответствуют различным классам предлогов.

Синтаксические средства обозначения действий и деятельности ОГВ

Действия и деятельность ОГВ могут быть обозначены простыми двусоставными предложениями со сказуемым в действительном (*Минфин продаёт активы*) или страдательном залоге (*активы продаются Минфином*). На уровне синтаксической семантики приведённые примеры получают одинаковую интерпретацию в виде нормализованных СГ (рис. 1).

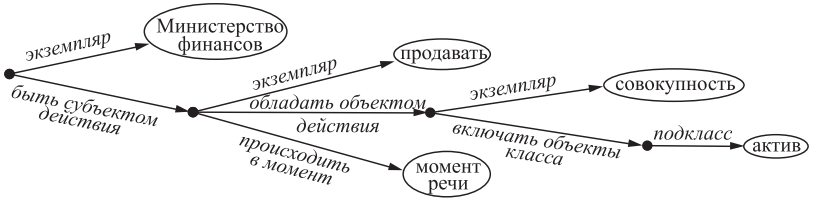

Рис. 1. Семантический граф (СГ)

Простые двусоставные предложения могут содержать в своей структуре распространители, обозначающие действия или деятельность ОГВ, например деепричастные обороты (ср. *продавая активы, Минфин постановил…*). В нормализованном СГ значения деепричастных оборотов отображаются так же, как и значения иных глагольных групп, при этом устанавливается связь субъекта основного ДС с ДС, обозначаемым деепричастным оборотом, а также отношение одновременности или предшествования между ними. Указанные распространители могут быть выражены группами ПС, и их значения встраиваются в СГ в соответствии с синтаксической позицией группы.

Субъект, равно как и объект действия или деятельности может быть выражен при помощи ИГ в РП (ср. *приговор суда*, *роспуск Госдумы*). Выбор субъектной или объектной интерпретации основыва-

ется исключительно на семантических валентностях, что не всегда позволяет разрешить неоднозначность: для выражения *приговор суда* возможна трактовка, в которой суд не приговаривает кого-либо, а сам является объектом приговора. Такая неоднозначность разрешается при помощи контекста или путём регистрации отдельных идиом. Контекстуальное разрешение указанной неоднозначности возможно в случаях, когда ПС образованы от переходного глагола и могут присоединять к себе ИГ в творительном падеже, обозначающую субъект ДС, и ИГ в РП в роли объекта (ср. *продажа активов Минфином*).

Действия и деятельность ОГВ могут быть выражены не только самими ИГ, но и их распространителями, в частности причастными группами в препозиции (*продающий активы Минфин*) и обособленными причастными оборотами в постпозиции (*Минфин, продающий активы*). Сходную с причастными группами функцию выполняют определительные придаточные предложения (*Минфин, который продаёт активы*). Действия или деятельность ОГВ могут также обозначаться группами страдательных причастий с обозначением субъекта, формально выраженным ИГ в творительном падеже (*проданные Минфином активы*; *активы, проданные Минфином*). Нормализованный СГ для ИГ, распространённой группой страдательного причастия, должен совпадать с СГ ИГ, распространённой группой действительного причастия, с учётом перестановки ИГ (то есть *проданные Минфином активы* после нормализации СГ получает такую же интерпретацию, как и *продавший активы Минфин*).

Выводы

В ходе выполнения работы были выделены различные синтаксические конструкции системы наименований ОГВ, а также синтаксические конструкции, обозначающие действия и деятельность ОГВ. Были разработаны принципы построения отношений между концептами, соответствующими названиям ОГВ и их видам деятельности. Онтология AIIRE была дополнена в соответствии с разработанными принципами. В ходе исследования были разработаны параллельные классификации ДС в соответствии с различными классами глаголов и ПС, представленных в имеющейся текстовой коллекции, обеспечивающие вычисление семантических валентностей

исследуемых единиц. Были рассмотрены и решены некоторые проблемы анализа идиоматических единиц в названиях ОГВ. Таким образом, была обеспечена корректная обработка ЛП AIIRE русскоязычных языковых единиц, обозначающих различные виды деятельности и действия органов государственной власти Российской Федерации.

Литература

Апресян Ю. Д. Лексическая семантика (синонимические средства языка). М., 1974.

Богданов В. В. Моделирование семантики предложения // Прикладное языкознание: Учебник. СПб, 1996.

Вакарюк Л. А. Структурно-семантический анализ имен существительных со значением процесса, не мотивированных глаголами (на материале русского языка): автореф. дис. … канд. филол. наук; 10.02.01. Черновцы, 1985.

Герд А. С. Основы научно-технической лексикографии. Л., 1986.

Золотова Г. А. Коммуникативные аспекты русского синтаксиса: моногр. М., 1982.

Казаков В. П. Синтаксис имен действия. СПб., 1994.

Маслов Ю. С. Глагол // Лингвистический энциклопедический словарь / гл. ред. В. Н. Ярцева. 2-е изд., доп. М., 2002.

Нгуен Т. Т. В. Терминология государственного управления в современном русском языке: автореф. дис. … канд. филол. наук. М., 2001.

Цейтин Г. С. Программирование на ассоциативных сетях // ЭВМ в проектировании и производстве. Л., 1985. Вып. 2.

Чудинов А. П. Российская политическая метафора в начале XXI века // Политическая лингвистика. Екатеринбург, 2008. Вып. 1(24). С. 86–93.

Юдина Т. Н., Богомолова А. В. УИС РОССИЯ: онтология предметной области «государственное управление» // Интернет и современное общество: сб. науч. статей. Материалы XIV Всероссийской объединенной конференции «Интернет и современное общество». СПб., 2011.

Dobrov A. V. Semantic and Ontological Relations in AIIRE Natural Language Processor // Computational Models for Business and Engineering Domains. Rzeszow, Sofia, 2014.

Serial Combination of Rules and Statistics: A Case Study in Czech Taggin / J. Hajic [et al.] // Proceedings of the 39th Annual Meeting of the Association for Computational Linguistics (ACL 2001). Tolouse, 2001.

Sowa J. F. Conceptual Graphs: Draft Proposed American National Standard // International Conference on Conceptual Structures ICCS-99. Lecture Notes in Artificial Intelligence 1640. Berlin; New-York, 1999. P. 1–65.