Семантический анализ новостных сообщений по теме «Электронные услуги»: опыт применения методов онтологической семантики

А.В. Добров[1 2 3], А.Е. Доброва[3], Н.Л. Сомс [1 3], А.В. Чугунов[1]
[1] Санкт-Петербургский национальный исследовательский университет информационных технологий, механики и оптики (Университет ИТМО)
chugunov@egov-center.ru
[2] Санкт-Петербургский государственный университет
[3] Компания «АИРЕ»
{adobrov, adobrova, nsoms}@aiire.org

Аннотация

В статье представлен опыт создания семантических средств для контент-анализа коллекции текстов об электронных государственных услугах. Пилотное исследование осуществляется на массиве текстов, представляющих материалы коллекции новостных сообщений (ленты новостей и информационного бюллетеня) Центра технологий электронного правительства Университета ИТМО. Работа является частью серии исследований, ориентированных на изучение «повестки дня», формируемой средствами массовой информации и интернет-ресурсами по тематике, связанной с развитием электронного правительства. Лингвистический компонент системы был создан на основе компьютерного лингвопроцессора AIIRE и одноимённой онтологии.

Ключевые слова: электронные услуги, контент-анализ, новостные СМИ, повестка дня, автоматическая обработка текста, онтологическая семантика.

Представляемая работа продолжает серию исследований, ориентированных на изучение «повестки дня», формируемой в медиапространстве по темам, связанным развитием электронного правительства и онлайн-услуг, а также выявление специфики обсуждаемости процесса внедрения электронных государственных услуг в сети Интернет, в том числе в социальных медиа и блогосфере [1, 2]. В 2014 году автоматизированный контент-анализ новостных сообщений с использованием методов кластеризации был осуществлен как совместный исследовательский проект Центра технологий электронного

Сборник научных статей XVIII Объединенной конференции «Интернет и современное общество» IMS-2015, Санкт-Петербург, 23-25 июня 2015 г.

правительства Университета ИТМО и Лаборатории интернет-исследований НИУ ВШЭ – Санкт-Петербург [3].

Для исследования на основе методов онтологической семантики с использованием лингвопроцессора AIIRE была взята та же коллекция текстов, что и в исследовании 2014 года.

Коллекция текстов, которая является материалом исследования, составляет 3,5 тыс. новостных сообщений за три года. Массив информации сформирован на основе отбора новостных сообщений СМИ, который производился еженедельно в рамках текущей информационно-аналитической деятельности Центра технологий электронного правительства Университета ИТМО. Данная информация является основой для выпуска еженедельного бюллетеня, выпускаемого в электронном виде. Новостные сообщения также размещаются на сайте Центра технологий электронного правительства (ЦТЭП — http://egov.ifmo.ru/) со ссылками на первоисточники.

Важным фактором, который позволяет считать эти тексты репрезентативными, т.е. освещающими весь основной поток публикаций СМИ по указанной тематике, является (а) наличие перечня источников, освещающих тематику, и подлежащих обязательному просмотру; (б) регулярность и систематичность сбора информации экспертами ЦТЭП; (в) отсутствие в массиве републикаций, т.е. в подборку включаются только самые информативные сообщения СМИ, «перепечатки» игнорируются; (г) положительные отзывы подписчиков бюллетеня — представителей органов власти и экспертов, использующих данную информацию в своей деятельности.

Массив сообщений представляет собой набор файлов, каждый из которых включает информацию за определенный месяц. Всего для анализа было подготовлено 39 файлов (в каждом от 70 до 100 новостных сообщений).

Контент-анализ проводился на основе изучения содержания новостных сообщений из следующих групп информационных источников (СМИ и информационные ресурсы, регулярно публикующие новости по тематике электронного правительства и электронных услуг):

1. Сайты органов власти и официальные тематические порталы (Сайты Президента России, Правительства РФ, Минкомсвязи и Минэкономразвития, сайт «Административная реформа в Российской Федерации», Комиссии по модернизации и технологическому развитию экономики России и др.).

2. Региональные ресурсы (Интернет-представительство глав регионов РФ «Клуб Регионов»; сайты «Электронных правительств» Астраханской и Самарской областей, региональные новостные СМИ, имеющие соответствующие разделы или тематические рубрики и др. ресурсы).

3. Экспертные структуры (ВЦИОМ, ГосМенеджмент: электронный журнал, Всероссийский научно-исследовательский институт проблем вычислительной техники и информатизации, Экспертный центр электронного государства, Фонд информационной демократии и др.).

4. СМИ и новостные порталы (CNews, ComNews, Российская газета, Комсомольская правда, Известия, Независимая газета, Коммерсантъ, Ведомости, ПРАЙМ-ТАСС, РБК daily, РИА «ФедералПресс», ТАСС-Телеком, PC-Week, ИА REGNUM Новости, Портал Право.ру и др.)

5. Ресурсы в социальных сетях и блоги (блог «Госуслуги» в Livejournal, группа «Электронное правительство» в Facebook и др.).

Для анализа коллекции текстов была разработана информационная система. Лингвистический компонент данной системы был создан на основе компьютерного лингвопроцессора AIIRE и одноимённой онтологии (принципы работы и структура AIIRE описаны в работах [4, 5, 6, 8]). При помощи лингвопроцессора каждый текст в коллекции подвергался комплексной лингвистической обработке (морфологический, синтаксический и семантический анализ), результатом которой являлись версии семантического анализа — семантические графы (структура и содержание этих графов описана в работах [4] и [7]). Семантический анализ позволил во многом решить проблемы неоднозначности языковых единиц и ложной корреляции при анализе текстов [4], тем самым обеспечивая возможность повышения точности этого анализа. Полнота анализа была многократно повышена по сравнению с обычными методами (см., например, [9, 10]) при помощи компьютерной онтологии.

Онтология позволила для каждого концепта в каждом семантическом графе найти все остальные концепты онтологии, связанные с этим концептом любыми транзитивными разновидностями отношения обладания и / или принадлежности. Среди концептов семантического графа, а также найденных связанных с ними концептов онтологии выбирались подклассы трёх основных классов, существенных, для контент-анализа в области электронных государственных услуг — ‘орган государственной власти’, ‘область государственного управления’ и ‘населённая территориальная единица Российской Федерации’.

В результате такой обработки тексты были привязаны непосредственно к концептам онтологии, каждый из которых входит в одну из трёх иерархий, что позволило представить коллекцию в наглядном интерактивном виде, позволяющем осуществлять произвольные выборки текстов по соответствующим трём иерархическим фильтрам и осуществлять выгрузку данных по произвольной выборке в формате электронных таблиц (веб-интерфейс системы доступен по адресу http://aiire.org/chuca).

Для того, чтобы построить родо-видовую иерархию органов государственной власти, была произведена классификация этих органов по различным основаниям. В частности, все органы власти были классифицированы по ветвям власти, к которым они относятся — соответственно, были введены классы ‘орган исполнительной власти’, ‘орган законодательной власти’, ‘орган судебной власти’ и т.д. При этом концепт ‘орган исполнительной власти’ был в явном виде отнесён к концепту ‘исполнительная власть’ специальным отношением «непосредственно входить в состав организации», а его непосредственными подклассами стали департаменты, мэрии, государственные службы, правоохранительные органы, федеральные органы исполнительной власти и правительство.

В связи с тем, что исследовавшийся неструктурированный массив текстов содержал в себе информацию в основном о государственных услугах конкретных ведомств, далее в онтологии AIIRE была разработана непротиворечивая система (родо-видовая иерархия и таксономия по отношению

организационной принадлежности) органов государственной власти. Абстрактный концепт ‘министерство’ вместе с его строгими эквивалентами ‘ведомство’ и ‘департамент’ был отнесен к концепту ‘правительство’ отношением ‘непосредственно входить в состав организации’. Это отношение с правительством, таким образом, наследуется всеми конкретными подклассами и экземплярами министерств.

Непосредственными подклассами концепта ‘министерство' стали такие промежуточные подклассы, как ‘министерство конкретной компетенции’ и ‘министерство конкретной страны’ — эти подклассы потребовались в связи с наличием двух разных оснований для классификации министерств. Подклассы ‘министерства конкретной страны’ были выделены по географическому принципу — это такие классы, как 'министерство Америки’, ‘министерство Австралии’, ‘министерство Евразии’ и ‘министерство Океании’. Для этих классов известно географическое положение страны (материк), но ещё не известна та область государственного управления, которая относится к компетенции министерства. Данные подклассы имеют дальнейшую классификацию до самых нижних классов и до экземпляров, для которых известны и конкретная административно-территориальная единица, и область государственного управления, например: ‘Министерство Кавказа’ — ‘Министерство Нагорно-Карабахской Республики’ — ‘Министерство здравоохранения Нагорно-Карабахской Республики’. Помимо всего прочего, концепт ‘министерство конкретной страны’ отношением ‘непосредственно входить в состав организации’ отнесён к ‘правительству конкретной страны’, что означает, что у концепта ‘правительство’ производится параллельная классификация по странам и регионам.

Классификации в онтологии AIIRE вообще часто воспроизводят структуру других классификаций этой онтологии в соответствии с правилами логического вывода и могут перемножаться при наличии нескольких оснований (это явление подробно описано в работе [6]). Классификация органов государственной власти оказалась параллельна одновременно иерархии населённых территориальных единиц (регионов) и областей государственного управления. В результате возникли промежуточные классы, например — ‘министерство России конкретной компетенции’. Этот класс имеет два надкласса — ‘министерство России’ и ‘министерство конкретной страны и определённой компетенции’. В подклассах можно увидеть семнадцать конкретных министерств, выделенных и обработанных с учётом требований онтологии. Так, концепт ‘Министерство юстиции РФ’ — это экземпляр одновременно нескольких классов: ‘министерство юстиции Азии’, ‘министерство юстиции Европы’, ‘министерство России конкретной компетенции’ и ‘правоохранительный орган России’. У этих классов данный экземпляр наследует связи с различными другими концептами, что позволяет лингвопроцессору строить дополнительные гипотезы о связи текста ними.

Отношением ‘осуществлять деятельность в области’ концепт 'Министерство юстиции РФ’ привязан к концепту ‘юстиция (судебная деятельность)’, отношением ‘непосредственно входить в состав организации’ — с концептом ‘органы правосудия’, обратным же отношением (‘непосредственно включать в свой состав организацию’) связан с концептами ‘Федеральная служба

исполнения наказаний РФ’ и ‘Федеральная служба судебных приставов’. Кроме того, ‘Министерство юстиции РФ’ является «типичным представителем» ‘министерства юстиции’ — это позволяет лингвопроцессору понять, что в русскоязычном тексте, вопреки формальной логике, словосочетание министерство юстиции, скорее всего, означает ‘Министерство юстиции РФ’, а не ‘Министерство юстиции любой страны’.

Следует уточнить правила интерпретации таких отношений, как ‘непосредственно входить в состав организации’ и обратного отношения ‘включать в свой состав организацию’. Данные отношения предполагают, что, например, что Министерство Юстиции РФ включено в органы правосудия, а обратное отношение говорит о том, что Министерству Юстиции РФ подведомственны Федеральная служба исполнения наказаний и Служба судебных приставов. Это отношение является частным случаем транзитивного отношения принадлежности, что позволяет инструментарию контент-анализа относить информацию к сфере компетенции Министерства Юстиции даже в том случае, когда в тексте упомянута только та или иная подведомственная ему служба.

В общей сложности в рамках данного пилотного исследования было проработано 570 концептов. Системой было выполнено 1170 привязок текстов к органам государственной власти, 32737 привязок к областям государственного управления и 3537 привязок к административно-территориальным единицам. Все выполненные привязки отображаются в интерфейсе системы.

Исследование находится на первом этапе реализации, и его развитие требует введения новых концептов, а также уточнения имеющихся в онтологии.

В завершение следует подчеркнуть, что данная работа имеет пилотный характер и является составной частью более широкого исследования.

Авторы планируют продолжение данной работы, в том числе заинтересованы в сотрудничестве с другими исследовательскими коллективами, готовыми использовать данную методологию и тем самым развивать созданный инструментарий.

Литература

[1] Бершадская Л.А., Чугунов А.В. Услуги электронного правительства: исследование дискуссий в социальных сетях // Межотраслевая информационная служба. 2014. № 1 (166). С. 10-17. URL: http://elibrary.ru/item.asp?id=21278501

[2] Бершадская Л.А., Биккулов А.С., Болгова Е.В., Чугунов А.В., Якушев А.В. Социальные сети и социометрические исследования: теоретические основания и практика использования автоматизированного инструментария изучения виртуальных сообществ // Информационные ресурсы России. 2012. № 4. С. 19-24. URL: http://elibrary.ru/item.asp?id=17910592

[3] Видясова Л.А., Кольцов С.Н., Чугунов А.В. Формирование «повестки дня» в сфере электронного правительства: результаты контент-анализа новостных сообщений // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-

2014), Санкт-Петербург, 19 – 20 ноября 2013 г. – СПб.: Университет ИТМО, 2014. С. 124 – 128.
[4] Добров А.В. Автоматическая рубрикация новостных сообщений средствами синтаксической семантики -- дисс. ... канд. филол. наук, 10.02.21, дата защиты - 26.03.2014.
[5] Сомс Н.Л., Добров А.В., Доброва А.Е. Использование средств лингвистической обработки текстов в системе мониторинга информационных ресурсов по пользовательским предпочтениям // Технологии информационного общества в науке, образовании и культуре: сборник научных статей. Труды XVII Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2014), Санкт-Петербург, 19 – 20 ноября 2014 г. СПб: НИУ ИТМО, 2014.
[6] Dobrov A.V. Semantic and Ontological Relations in AIIRE Natural Language Processor // Computational Models for Business and Engineering Domains. Rzeszow-Sofia: ITHEA, 2014. P. 147-157
[7] Добров А.В. К вопросу об универсальном представлении концептуальных структур в системах индексирования и автоматической рубрикации текстов // Материалы XLI международной филологической конференции — секция прикладной и математической лингвистики 26-31 марта 2012 г., 2012.
[8] Добров А.В. Автоматическая рубрикация текстов средствами комплексного лингвистического анализа // Структурная и прикладная лингвистика. 2012. № 9. С. 135-147.
[9] Мангейм Дж. Б., Рич Р. К. Политология. Методы исследования: Пер. с англ. / Предисл. А.К. Соколова = Empirical Political Analysis: Research Methods in Political Science. — М.: Весь Мир, 1997.
[10] Почепцов Г. Г. Теория коммуникации. М.: Рефл-бук, 2001.

Semantic Analysis of News Items on `Electronic Services' Subject Domain: Experience of Applying Methods of Ontological Semantics

A. Dobrov[1 2 3], A. Dobrova[3], N. Soms[1 3], A. Chugunov[1]
[1]ITMO University, [2]St. Petersburg State University, [3]AIIRE company

The article describes the experience of creating semantic tools for content analysis of texts collection on electronic public services. A pilot study is carried out on the material of texts representing the collection of news items (news feeds and the newsletter) of the ITMO University Center for Electronic Government. This work is part of a series of studies focused on the "Agenda" through the media and Internet resources on topics related to e-government development. The linguistic component of the system was based on AIIRE natural language processor, and computer ontology of the same name.

Keywords: electronic services, content analysis, news media, news agenda, natural language processing, ontological semantics.