Компьютерный анализ текстов

Автоматическое извлечение финансово-экономической информации из текстов на русском языке*

М.И. Ананьева, Д.А. Девяткин, М.А. Каменская, М.В. Кобозева, И.В. Смирнов

Аннотация. В статье рассматриваются проблемы создания методов и программных средств автоматического извлечения из текстов информации о финансово-экономических событиях и фактах, связанных с заданной географической областью (на примере Арктической зоны), с целью поддержки принятия решений на основе анализа информационного пространства. Предложен метод извлечения информации об инвестировании средств из текстов на русском языке, который позволяет выявлять сам факт вложения средств, сумму инвестирования, организацию-инвестора и географическую локацию (регион), в которой расположен объект финансирования. Представлен экспериментальный корпус из материалов СМИ, статей в профильных изданиях, посвященных Арктической зоне. Работоспособность предложенного метода была подтверждена экспериментально на представленном корпусе.

Ключевые слова: *информационно-поисковая система, извлечение финансово-экономической информации, поддержка принятия решений.*

Введение

Для поддержки принятия решений в качестве источника структурированной информации часто используются статистические базы, такие как FAOSTAT, Росстат и др. [1]. Однако, являясь хорошим источником ретроспективной информации, они, как правило, не содержат данных о недавних событиях, кроме того, полнота и степень детализации данных в таких базах тоже весьма ограничены. Все это приводит к невозможности или сложности оперативного принятия корректных решений. Решением этой проблемы могло бы стать извлечение информации из текстов сообщений СМИ и социальных сетей, статей в профильных журналах и других источников. В связи с этим становится актуальной задача автоматического извлечения информации о событиях и фактах из текстов на естественном языке.

В настоящей работе решается задача извлечения из текстов информации о финансово-экономических событиях и фактах, связанных с Арктической зоной. Под финансово-экономической информацией мы понимаем все события, связанные с вложением средств в инфраструктуру и социально-экономическое благоустройство региона (экологические программы, образование и здравоохранение коренных народов, строительство новых военных баз, разработка месторождений), а также покупку акций компаний, ведущих свою деятельность в Арктике и др. Таким образом, из текстов необходимо извлекать события и их участников, а также атрибуты событий: время, место, затраченные суммы. Нам интересны не только сущности указанных типов, но и отношения между ними. Кроме того, мы планируем устанавливать причинно-следственные связи между описываемыми в текстах событиями. Методы, позволяющие выявлять тексты на русском языке, относящиеся к определенной географической области, и извлекать из них всю перечисленную совокупность информации, отсутствуют на текущий момент, что делает настоящую работу значимой.

Несмотря на важность Арктического региона, существует ограниченное количество источников структурированной информации о нем, которые могли бы быть использованы при принятии

* Работа выполнена при финансовой поддержке РФФИ грант №15-29-06053 офи-м.

решений, что повышает практическую ценность настоящей работы. В качестве источников анализируемой текстовой информации выступили сайты информационных агентств, профильных журналов и другие интернет-ресурсы. Так как финансово-экономическая информация встречается лишь в незначительной части сообщений, был сформирован сбалансированный тестовый набор данных, состоящий из 3 тыс. текстов. Экспериментальная оценка представленного метода, выполненная на этом тестовом наборе, показала его работоспособность.

1. Связанные работы

В последнее время среди российских исследователей растет интерес к проблеме извлечения информации из текстов, о чем говорит, например, проведение соревнований по выявлению именованных сущностей и фактов на конференции по компьютерной лингвистике Диалог-2016*. В работе [2] представлен метод, в котором используется глубокий синтаксический и семантический анализ. С помощью этого метода можно извлекать такие типы информации, как лицо, местоположение, организация, время, дата, затраченные суммы.

В ряде работ представлено решение частных задач, близких настоящему исследованию: извлечение числовых характеристик (км., см., мм. и др.) из текста [3], триплетов «Субъект – Предикат – Объект» [4], информации о людях и организациях, связанных отношением «занимать должность» [5]. В вопросе анализа текстовой информации, наибольший интерес для нас представляет область политики и юриспруденции. Для анализа политических событий в мировом сообществе в сфере компьютерной лингвистики разработан ряд программных методов и систем по извлечению информации. Отечественная система ИСИДА-Т [6, 7] принимает факты как события и состояния, участниками которых выступают лица, организации, роли лиц, геополитические единицы. Факты описывают отставки/назначения и структурные отношения между сущностями. Авторы используют понятие текстовой ситуации, под которым понимают событие, описанное в одном предложении при помощи предикатного слова (уволить, назначить, отставка) и трех именных групп, называющих трех участников ситуации (2 лица и должность).

В [8] представлена система извлечения сущностей и фактов из новостных статей. Данная система дает возможность делать сложные запросы в новостных статьях. Авторы предлагают механизм извлечения именованных сущностей, которые относятся к различным классам, таким как «Правительство», «Глава», «Президент», и кластеризации их на три типа – лица, организации, геолокации. Особое внимание уделяется анализу географических сущностей.

В работе зарубежных авторов [9] представлена вероятностная модель для извлечения событий между главными политическими деятелями из новостных корпусов. В работе используется комбинация лингвистических данных и политической информации (отношения между двумя действующими лицами в каждый момент времени). В основе семантического анализа данной системы – словарь глаголов, связанных с теми или иными политическими изменениями.

Наряду с анализом политических событий задача извлечения событий актуальна для экономической и финансовой сферы. Так, специальная система SPEED (Semantics-based Pipeline for Economic Event Detection) [10] извлекает финансовые события (такие как слияние корпораций, приобретение компаний и т.д.) из новостных статей на основе семантического анализа.

Существует ограниченное количество доступных лингвистических ресурсов, корпусов на русском языке, пригодных для обучения методов извлечения информации о событиях, поэтому наиболее перспективным видится создание такого метода в соответствии с парадигмой открытого извлечения информации (open information extraction). В статье [11] предложен метод формирования крупных многоязычных корпусов на основе Википедии. Представлен также подход к автоматическому построению новых связей и обучающих примеров на основе существующей сети из категорий и статей Википедии, а также связей между ними. В [12] представлен метод извлечения именованных сущностей, для обучения которого не требуются размеченные наборы данных, аннотированные параллельные корпуса или другие, зависящие от языка ресурсы. Этот метод состоит в обучении векторных представлений для слов (word embeddings), которые кодируют признаки слов в каждом языке. С использованием этих векторов, структуры связей между статьями из Википедии и атрибутами объектов из Freebase можно построить набор данных для обучения метода извлечения именованных сущностей из текстов на любом распространенном языке.

В работе [13] представлен подход к выявлению именованных сущностей, имеющих отношение к финансам и экономике, и связей между ними. Этот подход основан на использовании результатов семантического анализа текстов с использова-

* http://www.dialog-21.ru/dialogue2016/results/

нием заранее сформированной онтологии. Однако извлечение данных о финансировании, в том числе сумм финансирования, не предусмотрено.

2. Метод извлечения финансово-экономической информации из текстов

Метод извлечения финансово-экономической информации из текстов на русском языке, предложенный авторами, позволяет выявлять информацию о самом факте вложения средств, сумме инвестирования, организации-инвесторе и географической локации, в которой расположен объект финансирования. Извлечение сущностей, связанных с описанием финансово-экономической ситуации из текстов, проводится в несколько этапов (рис. 1). На первом этапе с помощью библиотеки АОТ [14] выполняется токенизация и морфологический анализ текстов, для всех словоупотреблений выявляются их нормальные формы и части речи. Для нормализации числительных применяется библиотека Freeling [15]. В этой библиотеке реализован метод, позволяющий выявлять нормальные формы числительных, в том числе составных, непосредственно в виде чисел. Нормальные формы существительных, извлеченные из текстов, сопоставляются также с базой наименований географических объектов в Арктической зоне, сформированной на основе GeoNames*. Далее проверяется соответствие текстов ряду лексико-морфологических шаблонов, указывающих на возможное упоминание процессов финансирования (табл. 1).

На основе этой информации для каждого текста формируется вектор признаков, по которым на втором этапе производится фильтрация – отбрасываются сообщения, не содержащие информации о финансово-экономической деятельности, а также не относящиеся к Арктической зоне. Проблема фильтрация текстов в этом исследовании рассматривается как задача бинарной классификации, и для ее решения применяются методы машинного обучения с учителем (SVM, логистическая регрессия). С помощью метода поиска похожих документов [16] отфильтровываются также нечеткие дубликаты текстов.

Затем, с помощью семантико-синтаксического анализатора [17] выполняется полный лингвистический анализ отобранных текстов. Далее производится извлечение именованных сущностей (персоны, организации, географические локации, названия валют), для чего применяется комбинированный подход: метод с частичным обучением с учителем, реализованный в библиотеке Polyglot [12], дополняется методом, использующим лексико-синтаксические шаблоны. Благодаря такому подходу удается повысить полноту извлечения информации из текстов на русском языке. На последнем этапе производится сохранение всей извлеченной метаинформации в реляционную базу данных (БД), суммы финансирования при этом конвертируются в рубли по курсу ЦБ РФ на дату публикации сообщения.

Основным отличием представленного подхода от аналогов является:

- использование методов открытого извлечения информации для выявления именованных сущностей, что позволяет сократить трудозатраты, связанные с разметкой обучающего корпуса;
- использование методов выявления нечетких дубликатов текстов, а также извлечение данных о суммах финансирования. Эти величины могут являться важной информацией при поддержке принятия решений.

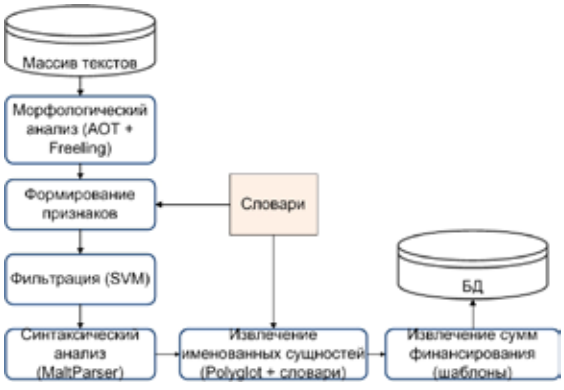

Рис. 1. Процесс извлечения финансово-экономической информации из текстов

3. Эксперименты

Для проведения экспериментов по извлечению финансово-экономической информации авторским коллективом был создан корпус текстов. На данный момент корпус включает 66 000 текстов следующих категорий:

– российские научные журналы (такие как «Проблемы Арктики и Антарктики», «Арктика и Север» и другие);
– российские военные журналы («Красная звезда», журнал Минобороны «Ориентир» и другие);
– российские СМИ (АрктикИнфо, Арктика Сегодня и другие).

Все ресурсы обходились автоматическим краулером, который настраивался под каждый тип ресурсов и загружал тексты в хранилище.

* http://www.geonames.org

Табл. 1

Примеры шаблонов упоминаний процессов финансирования

Шаблон	Пример текстового фрагмента
НФ(«выделить») + * + [ЧР(Числ) + ЧР(Сущ)&КСК(количественное)?] + НФ(«рубль»)	На разработку ресурсов полярного региона в ближайшее время будет выделено почти 100 млрд долларов.
НФ(«привлечь») + * + [ЧР(Числ) + ЧР(Сущ)&КСК(количественное)?] + НФ(«доллар») + * + НФ(«инвестиция»)	Мурманская область в 2011 году привлекла 75 млн долларов иностранных инвестиций.
НФ(«объем»)||НФ(«величина»)|| НФ(«порядок») + * +ЧР(Глаг) + [ЧР(Числ) + ЧР(Сущ)&КСК(количественное)?] + НФ(«рубль»)	За январь-ноябрь 2013 года объем инвестиций в основной капитал составил 41 947,7 млн рублей.
Обозначения: НФ – нормальная форма; ЧР – часть речи; КСК – категориально-семантический класс; & – логическое и; || – логическое или; * – любой текстовый фрагмент; [...] – фрагмент повторяется 1 или более раз; ? – фрагмент необязателен.	

Так как финансово-экономическая информация встречалась лишь в незначительной части сообщений библиотеки, на ее основе был сформирован сбалансированный тестовый набор данных, состоящий из 3 тыс. текстов. Результаты анализа этого набора оценивались несколькими экспертами по следующим критериям: качество выделения фрагментов текста, содержащих финансово-экономическую информацию, качество выявления локаций, качество выявления организаций. В табл. 2 приведены примеры информации, извлеченной из новостных сообщений, посвященных финансированию различных проектов в Арктической зоне.

На основе оценок экспертов вычислялись показатели точности, полноты и F_1-меры, используемые обычно для оценки качества работы методов, основанных на машинном обучении с учителем [18]. Полученные в результате оценки качества извлечения информации на тестовом наборе представлены в табл. 3.

Результаты показывают, что факт финансирования и суммы финансирования выделяются достаточно надежно. Относительно невысокая пол-

Табл. 2

Примеры извлекаемой информации

Фрагмент новости	Извлекаемая информация		
	Организации	Локации	Суммы
Ямал направит на поддержку северного оленеводства 0,5 млрд рублей 05 мая 2012 Версия для печати В бюджете ЯНАО на 2012 год и плановый период – 2013 и 2014 годы учтена субсидия на поддержку северного оленеводства в размере 527, 8 млн рублей	ЯНАО	Ямал	527800000 руб.
РФ выделит средства на инфраструктуру плавучей АЭС 13 мая 2015 Версия для печати Правительство РФ планирует до 2020 года выделить из госбюджета 5 млрд рублей на сооружение на Чукотке объектов береговой и гидротехнической инфраструктуры для первой российской плавучей атомной теплоэлектростанции «Академик Ломоносов»	РФ	Чукотка	5000000000 руб.
«Роснефть» вложит 7,3 млрд рублей в работы на шельфе моря Лаптевых 17 мая 2016 Версия для печати Дочернее предприятие «Роснефти» «РН-Шельф-Арктика» в 2016-2017 годах выполнит проектно-изыскательские и строительные работы на поисково-оценочных скважинах на шельфе моря Лаптевых и Охотского моря	Роснефть	море Лаптевых, Охотское море.	7300000000 руб.

Табл. 3

Результаты экспериментов по извлечению финансово-экономической информации

Тип извлекаемой информации	P	R	F_1
Факт финансирования	0,87	0,93	0,90
Сумма финансирования	0,96	0,75	0,84
Организация	0,75	0,76	0,76
Географическая локация	0,79	0,90	0,84

нота выявления сумм финансирования связана с ограниченностью набора шаблонов, используемых для нормализации числительных. В ходе дальнейших экспериментов этот набор будет расширен. Для повышения точности выявления организаций и локаций предполагается ввести дополнительный этап фильтрации извлеченных именованных сущностей с помощью методов, основанных на машинном обучении с учителем.

Заключение

В работе представлен аналитический обзор методов и систем извлечения информации из текстов на естественных языках для поддержки принятия решений. Предложен метод извлечения из текстов на русском языке информации об инвестировании средств, который позволяет выявлять информацию о самом факте вложения средств, сумме инвестирования, организации-инвесторе и географической локации (регионе), в которой расположен объект финансирования. Представлен также экспериментальный корпус из материалов СМИ, статей в профильных изданиях, посвященных Арктической зоне. Работоспособность предложенного метода была подтверждена экспериментально на представленном корпусе.

В ходе дальнейших исследований планируется разработать метод выявления целей финансирования из текстов на русском языке, а также построения связей, между выявленными суммами, организациями, локациями и целями. Основной проблемой здесь видится построение связей между сущностями, находящимися в разных предложениях текста. Будут также созданы экспериментальные программные средства информационной поддержки принятия решений, использующие методы фасетного поиска с учетом извлеченной финансово-экономической информации.

Литература

1. *Maes J. et al.* Mapping ecosystem services for policy support and decision making in the European Union // Ecosystem Services. 2012. Vol. 1. №. 1. pp. 31-39.
2. *Starostin A.S., Smurnov I.M., Stepanova M.E.* A production system for information extraction based on complete syntactic-semantic analysis // Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". 2014. URL: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/StarostinAS.full.pdf
3. *Харабет Я.К.* Автоматическое выделение количественных конструкций в русскоязычных научно-популярных текстах // Сборник трудов XVIII Всероссийской объединененной конференции IMS-2015. С. 100-102.
4. *Хайрова Н., Шаронова Н., Гаутам А.П.С.* Логико-лингвистическая модель генерации фактов из текстовых потоков информационной корпоративной системы // Information Theories and Applications. 2015. № 2. Т. 22. С. 142-152.
5. *Гершензон Л.М., Ножов И.М., Панкратов Д.В.* Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности // Сборник “Компьютерная лингвистика и интеллектуальные технологии”. 2005. URL: http://www.dialog-21.ru/Archive/2005/Gershenzon%20Nozhov%20Pankratov/Gershenzon_Nozhov_Pankratov.pdf
6. *Кормалев Д.А., Куршев Е.П., Сулейманова Е.А., Трофимов И.В.* Извлечение информации из текста в системе ИСИДА-Т // Труды 11-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL’2009. Петрозаводск, Россия. 2009. URL: http://resources.krc.karelia.ru/math/doc/rcdl2009/247_253_Section07-2.pdf
7. *Власова Н.А.* Извлечение информации о ситуациях отставок-назначений в новостных текстах. Опыт разметки коллекции. Результаты тестирования // Труды 15-й Всероссийской научной конференции «Электронные библиотеки: перспективные методы и технологии, электронные коллекции» - RCDL-2013. Ярославль, Россия. 2013. URL: http://ceur-ws.org/Vol-1108/paper6.pdf
8. *Zharikov A., Kristalovsky K., Pivovarov V.* Information Retrieval System for News Articles in Russian // Proceedings of the Fifth Russian Young Scientists Conference in Information Retrieval. St. Petersburg. 2011. P. 5-14. URL: http://elar.urfu.ru/bitstream/10995/3707/3/RuSSIR_2011_01.pdf

9. *O'Connor B., Stewart B., Smith N. A.* Learning to extract international relations from political context. 2013.
10. *Hogenboom A., Hogenboom F., Frasincar F., Schouten K., O. van der Meer.* Semantics-based information extraction for detecting economic events. URL: http://link.springer.com/article/10.1007/s11042-012-1122-0/fulltext.html
11. *Nastase V., Strube M.* Transforming Wikipedia into a large scale multilingual concept network // Artificial Intelligence. 2013. Vol. 194. P. 62-85.
12. *Al-Rfou R., Kulkarni V., Perozzi B., Skiena S.* Polyglot-NER: Massive multilingual named entity recognition //Proceedings of the 2015 SIAM International Conference on Data Mining. – Society for Industrial and Applied Mathematics, 2015. P. 586-594.
13. *Дмитриев А.С., Соловьев И.С., Заболеева-Зотова А.В.* Извлечение взаимосвязей между объектами и терминами в текстах на экономическую тематику //Известия Волгоградского государственного технического университета. 2015. №. 13. С. 55-60.
14. *Сокирко А.В.* Морфологические модули на сайте www. aot.ru // Компьютерная лингвистика и интеллектуальные технологии: Труды международной конференции «Диалог'2004». 2004.
15. *Padró L., Stanilovsky E.* Freeling 3.0: Towards wider multilinguality // LREC2012. 2012.
16. *Суворов Р.Е., Соченков И.В.* Определение связанности научно-технических документов на основе характеристики тематической значимости //Искусственный интеллект и принятие решений. 2013. №. 1. С. 33-40.
17. *Смирнов И.В., Шелманов А.О., Кузнецова Е.С., Храмоин И.В.* Семантико-синтаксический анализ естественных языков //Искусственный интеллект и принятие решений. 2013. №. 1. С. 43.
18. Flach P. Machine learning: the art and science of algorithms that make sense of data // Cambridge University Press. 2012.

Ананьева Маргарита Игоревна. Младший научный сотрудник ФИЦ ИУ РАН. Окончила в 2013г. Московский государственный лингвистический университет. Количество печатных работ: 9. Область научных интересов: компьютерная лингвистика, дискурсивный анализ текстов, корпусная лингвистика. E-mail: ananyeva@isa.ru

Девяткин Дмитрий Алексеевич. Младший научный сотрудник ФИЦ ИУ РАН. Окончил в 2011г. Рыбинскую государственную авиационную технологическую академию. Количество печатных работ: 21. Область научных интересов: машинное обучение, классификация и кластеризация текстов, методы обработки больших данных, методы анализа патентной и наукометрической информации. E-mail: devyatkin@isa.ru

Каменская Маргарита Александровна. Инженер-исследователь ФИЦ ИУ РАН. Окончила в 2014г. РУДН. Количество печатных работ: 5. Область научных интересов: компьютерная лингвистика, информационно-аналитические системы, интеллектуальный анализ информации. E-mail: mak@isa.ru

Кобозева Мария Вадимовна. Младший научный сотрудник ФИЦ ИУ РАН. Окончила в 2014г. МГУ им. М.В. Ломоносова, в 2016 г. – магистратуру Российского государственного гуманитарного университета. Количество печатных работ: 5. Область научных интересов: компьютерная лингвистика, автоматическая обработка естественного языка, дискурсивная структура текста. E-mail: kobozeva@isa.ru

Смирнов Иван Валентинович. Доцент ФИЦ ИУ РАН. Окончил в 2003 г. РУДН. Кандидат физико-математических наук. Количество печатных работ: 67. Область научных интересов: обработка естественного языка, интеллектуальный анализ информации. E-mail: ivs@isa.ru

Extraction of financial and economic information from texts in Russian

M.I. Ananyeva, D.A. Devyatkin, M.A. Kamenskaya, M.V. Kobozeva, I.V. Smirnov

Abstract. In this article we consider some problems that arise when developing methods and system for automatic extraction of economic events like investment of capital (e.g. in ecological projects), financial provision (e.g. of regions), purchase (e.g. of equipment), etc. In our research we focus on a particular geographical area – the Arctic Region. The aim of the project is to develop a pilot decision support system that analysis Internet media. In this article we propose a method for extraction of economic events, spent sums, investors, and location of an object to be financed. We created an experimental dataset in Russian which includes materials from electronic media and journals dedicated to the Arctic. The quality of the proposed method was confirmed experimentally on this dataset.

Keywords: *information extraction, detection of economic events, decision support.*

References

1. *Maes J. et al.* 2012 Mapping ecosystem services for policy support and decision making in the European Union. Ecosystem Services. Vol. 1. №. 1. pp. 31-39.
2. *Starostin A.S., Smurov I.M., Stepanova M.E.* 2014. A production system for information extraction based on complete syntactic-semantic analysis. Computational Linguistics and Intellectual Technologies: Papers from the Annual International Conference "Dialogue". Available at: http://www.dialog-21.ru/digests/dialog2014/materials/pdf/StarostinAS.full.pdf
3. *Kharabet Ya.K.* 2015. Avtomaticheskoye vydeleniye kolichestvennykh konstruktsiy v russkoyazychnykh nauchno-populyarnykh tekstakh [Automatic allocation of quantitative constructions in Russian-language popular science texts]. Sbornik trudov VIII Vserossiyskoy ob"yedinennoy konferentsii IMS-2015 [Proceedings of the VIII All-Russian Joint Conference IMS-2015]. pp.100-102.
4. *Khayrova N., Sharonova N., Gautam A.P.S.* 2015. Logiko-lingvisticheskaya model' generatsii faktov iz tekstovykh potokov informatsionnoy korporativnoy sistemy [Logico-linguistic model for fact generation from texts of the corporate information system]. Information Theories and Applications. № 2. Vol. 22. pp. 142-152.
5. *Gershenzon L.M., Nozhov I.M., Pankratov D.V.* 2005. Sistema izvlecheniya i poiska strukturirovannoy informatsii iz bol'shikh tekstovykh massivov SMI. Arkhitekturnyye i lingvisticheskiye osobennosti [A system for search and extraction of structured information from large-scale media collections. Architectural and linguistic features]. Sbornik "Komp'yuternaya lingvistika i intellektual'niye tekhnologii" [The journal "Computer Linguistics and Intellectual Technologies"]. Available at: http://www.dialog-21.ru/Archive/2005/Gershenzon%20Nozhov%20Pankratov/Gershenzon_Nozhov_Pankratov.pdf
6. *Kormalev D.A., Kurshev E.P., Suleymanova E.A., Trofimov I.V.* 2009. Izvlecheniye informatsii iz teksta v sisteme ISIDA-T [Information extraction from the text by the ISIDA-T system]. Trudy 11-y Vserossiyskoy nauchnoy konferentsii «Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kollektsii» RCDL'2009 [Proceedings of the 11th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" RCDL'2009]. Available at: http://resources.krc.karelia.ru/math/doc/rcdl2009/247_253_Section07-2.pdf
7. *Vlasova N.A.* 2013. Izvlecheniye informatsii o situatsiyakh otstavok-naznacheniy v novostnykh tekstakh. Opyt razmetki kollektsii. Rezul'taty testirovaniya [Extraction of the resignations-appointments evets from news texts. Experience in marking the collection. Test results]. Trudy 13-y Vserossiyskoy nauchnoy konferentsii «Elektronnyye biblioteki: perspektivnyye metody i tekhnologii, elektronnyye kollektsii» RCDL'2013 [Proceedings of the 13th All-Russian Scientific Conference "Digital Libraries: Advanced Methods and Technologies, Digital Collections" RCDL'2013]. Available at: http://ceur-ws.org/Vol-1108/paper6.pdf
8. *Zharikov A., Kristalovsky K., Pivovarov V.* 2011. Information Retrieval System for News Articles in Russian // Proceedings of the Fifth Russian Young Scientists Conference in Information Retrieval. St. Petersburg. pp. 5-14. Available at: http://elar.urfu.ru/bitstream/10995/3707/3/RuSSIR_2011_01.pdf
9. *O'Connor B., Stewart B., Smith N.A.* 2013. Learning to extract international relations from political context. URL: https://brenocon.com/oconnor+stewart+smith.irevents.acl2013.pdf
10. *Hogenboom A., Hogenboom F., Frasincar F., Schouten K., O. van der Meer*. Semantics-based information extraction for detecting economic events. Available at: http://link.springer.com/article/10.1007/s11042-012-1122-0/fulltext.html
11. *Nastase V., Strube M.* 2013. Transforming Wikipedia into a large scale multilingual concept network // Artificial Intelligence. Vol. 194. pp. 62-85.

12. *Al-Rfou R., Kulkarni V., Perozzi B., Skiena S.* 2015. Polyglot-NER: Massive multilingual named entity recognition //Proceedings of the 2015 SIAM International Conference on Data Mining. Society for Industrial and Applied Mathematics. pp. 586-594.
13. *Dmitriev A.S., Soloviev I.S., Zablenova-Zotova A.V.* 2015. Izvlecheniye vzaimosvyazey mezhdu ob'yektami i terminami v tekstakh na ekonomicheskuyu tematiku [Extraction of interrelations between objects and terms in economic texts]. Izvestiya Volgogradskogo gosudarstvennogo tekhnicheskogo universiteta [Bulletin of Volgograd State Technical University]. No. 13. pp. 55-60.
14. *Sokirko A.V.* 2004. Morfologicheskiye moduli na sayte www.aot.ru [Morphological modules on the site www.aot.ru]. Komp'yuternaya lingvistika i intellektual'nyye tekhnologii: Trudy mezhdunarodnoy konferentsii «Dialog'2004» [Computer linguistics and intellectual technologies: Proceedings of the international conference "Dialogue'2004"]. Available at: http://www.dialog-21.ru/media/2569/sokirko.pdf
15. *Padró L., Stanilovsky E.* 2012. Freeling 3.0: Towards wider multilinguality. LREC2012. Available at: http://www.lrec-conf.org/proceedings/lrec2012/pdf/430_Paper.pdf
16. *Suvorov RE, Sochenkov I.V.* 2013. Opredeleniye svyazannosti nauchno-tekhnicheskikh dokumentov na osnove kharakteristiki tematicheskoy znachimosti [Measuring similarity of scientific and technical documents using thematic importance characteristic] // Iskusstvennyy intellekt i prinyatiye resheniy [Artificial intelligence and decision-making]. Moskva: ISA RAN [Moscow: ISA RAS]. No. 1. pp. 33-40.
17. *Smirnov I.V., Shelmanov A.O. Kuznetcova E.S., Khramoin I.V.* 2014. Semantiko-sintaksicheskiy analiz yestestvennykh yazykov Chast' II. Metod semantiko-sintaksicheskogo analiza tekstov [Semantic-syntactic analysis of natural languages. Part II. Method for semantic-syntactic analysis of texts]. №. 1. pp. 11-24.
18. *Flach P.* 2012. Machine learning: the art and science of algorithms that make sense of data // Cambridge University Press. 395 p.

Ananyeva M.I. Junior research fellow in the Institute for Systems Analysis of the Federal Research Center "Computer Science and Control" (117312, prospekt 60-letiya Oktyabrya 9, Moscow). Graduated from Moscow State Linguistic University in 2013. Author of 9 papers. Scientific interests: computational linguistics, discourse analysis. E-mail: ananyeva@isa.ru

Devyatkin D.A. Researcher in the Institute for Systems Analysis of the Federal Research Center "Computer Science and Control" (117312, prospekt 60-letiya Oktyabrya 9, Moscow). Graduated from Rybinsk State Aviation Technology Academy after Pavel Solovyov in 2011. Authored 21 scientific papers. Scientific interests: machine learning, full-text clustering, data mining, scientometrics. E-mail: devyatkin@isa.ru

Kamenskaya M.A. Research engineer in the Institute for Systems Analysis of the Federal Research Center "Computer Science and Control" of Russian Academy of Sciences (117312, prospekt 60-letiya Oktyabrya 9, Moscow). Graduated from Peoples' Friendship University of Russia in 2014. Author of 5 scientific papers. Scientific interests: computational linguistics, information analysis system, data mining. E-mail: mak@isa.ru

Kobozeva M.V. Junior research fellow in the Institute for Systems Analysis of the Federal Research Center "Computer Science and Control" of Russian Academy of Sciences (117312, prospekt 60-letiya Oktyabrya 9, Moscow). Graduated from Moscow State University in 2014 and from Russian State University for the Humanities in 2016. Author of 5 scientific papers. Research interests: computational linguistics, natural language processing, discourse analysis. E-mail: kobozeva@isa.ru

Smirnov I.V. PhD, associate professor, head of the laboratory in the Institute for Systems Analysis of the Federal Research Center "Computer Science and Control" (117312, prospekt 60-letiya Oktyabrya 9, Moscow). Graduated from Peoples' Friendship University of Russia in 2006. Authored 67 scientific papers. Scientific interests: natural language processing, data and text mining. E-mail: ivs@isa.ru