

Приложение 1.

Дополнительная информация о технологиях и продуктах.

Базовые технологии

В настоящее время RCO владеет следующими технологиями компетенциями, используемыми в продуктах и решениях компании:

- Технологии в области компьютерной лингвистики и информационного поиска - Морфология (анализ/синтез, словарный/бессловарный), синтактико-семантический анализ (семантическая сеть текста, разрешение анафоры, выделение объектов, поиск фактов, тональность), поиск с учетом семантической разметке, нечеткий поиск, поиск сложных фактов в больших данных, поиск в базах знаний;
- Технологии «очистки данных» - нормализация ФИО и наименований организаций, нормализация российских адресов, идентификация нечетких дублей среди записей о физических и юридических лицах, проверка по «черным спискам», консолидация клиентских баз и другие применения;
- Компетенции по продуктам коммерческих вендоров – Oracle Database (администрирование, проектирование, разработка высоконагруженных баз размером в десятки терабайт, разработка расширений), Oracle WebLogic Server (администрирование, проектирование, разработка), MS SQL Server (администрирование, проектирование, разработка средненагруженных баз, разработка расширений), MS Sharepoint Portal Server (администрирование, разработка расширений), MS OLAP Services (проектирование, разработка);
- Компетенции по популярным продуктам с открытым исходным кодом – Redis, Nginx, Node.js, Django, Drools, PostgreSQL, RabbitMQ.

RCO ведет деятельность по следующим основным направлениям:

- Разработка технологий, создание продуктов и решений в области компьютерной лингвистики - синтаксический и семантический анализ, фактографический анализ, классификация текстов и др.;
- Компоненты русскоязычного полнотекстового поиска, семантического анализа для корпоративных решений от Oracle, Microsoft, IBM;
- Разработка заказных решений типа «полнотекстовые БД», «фактографические БД» и «системы очистки данных».

Примеры программных продуктов

RCO Fact Extractor

Комплексный инструментарий для разработки информационно-поисковых и аналитических систем, требующих лингвистического анализа текста на русском, английском, украинском, белорусском, казахском, киргизском, армянском, узбекском языках. Производит выделение различных классов сущностей, упомянутых в тексте (персоны, организации, география, предметы, действия, атрибуты и др.), и строит сеть отношений, связывающих эти сущности, а также предоставляет всю грамматическую информацию о составляющих анализируемого текста.

Используется для построения систем, решающих следующие типовые задачи:

- Автоматическое формирование досье;
- Конкурентная разведка (целевой сбор неструктурированной информации из разных источников в т.ч. интернет, с последующим структурированием и анализом);
- Расширение поисковых, навигационных и аналитических возможностей в информационных системах;
- Обработка строковых полей в БД, очистка данных.

RCO Zoom

Система RCO Zoom предназначена для осуществления поиска сущностей и фактов в документах, наряду с обычным контекстным поиском. В процессе индексирования документов, помимо построения классического полнотекстового индекса, в него добавляется еще и «семантическая разметка». Таким образом, семантическая поисковая система знает, не только какие слова встречаются в тексте, но и то, что именно эти слова означают.

Основные преимущества системы:

- Поддерживаемые языки – русский, английский, украинский, белорусский, казахский, киргизский, армянский;
- Расширенный язык запросов (шаблоны, расстояние, ...);
- Источники (хранилища документов) – файловая система, сайты, серверы электронной почты, базы данных;
- Форматы – TXT, HTML, PDF, все форматы Microsoft Office;
- Разграничение доступа на уровне сегмента базы, документа;
- Горизонтальная масштабируемость;
- Работа в кластере.

RCO Address Parser

Продукт, предназначенный для «очистки» больших массивов данных, содержащих адресную информацию. Не только нормализует записи, но и восстанавливает отсутствующую в источнике адресную информацию.

Устраняемые виды ошибок/опечаток:

- Использование схожих по написанию латинских букв вместо кириллических;
- Опечатки;
- Пропуск разделителей между адресными элементами;
- Пропуск ключевых слов (ул., г. и т.п.) в элементах адреса;
- Неполнота задания адреса;

- Использование старых названий городов и улиц;
- Автозамена часто встречающихся устойчивых сокращений.

Восстанавливаемая адресная информация

- Почтовый индекс;
- Код ФИАС;
- Пропущенные элементы адреса.

RCO СВД

Решение для ведения картотеки объектов, интеграции информации об объектах из различных источников, поиска и навигации в картотеке, анализа связей между объектами. Основные функции продукта:

- Создание и ведение карточек;
- Запрос информации об объекте во внешних источниках;
- Импорт списков, анкет;
- Автоматическое добавление информации в картотеку;
- Поиск карточек по реквизитам, прямым связям, связям через общий объект;
- Поиск первичных документов по контексту;
- Визуальный анализ связей между объектами, построение когнитивных карт.

RCO КАОТ

Платформа для построения автоматизированных систем, работающих с неструктурированной информацией. Платформа позволяет осуществлять разнообразную обработку текстов на естественном языке:

- Извлечение сущностей и фактов (Персоны, Организации, Геопонятия, Должности, Собственность, Высказывания, ...);
- Кластеризация (Новостные сюжеты, Выборка);
- Рубрицирование (Темы, Регионы);
- Построение досье на объект, досье по тематике (запросу).

Области применения по источникам документов

Технологии RCO прошли успешную апробацию на обработке текстов самых различных стилей: СМИ, нормативно-правовые документы, научно-технические отчеты, досье, сводки, социальные сети Интернета, web-сайты, записи баз данных. Вот типовые задачи, которые мы эффективно решаем с помощью компьютерной обработки текста:

- **для текстов СМИ:** выявление упоминаний персон и организаций, извлечение фактов заданного типа и их участников (*биографические данные, связи, владение собственностью, что он говорит и что о нем говорят*), мониторинг упоминаний о событиях заданного типа (*кадровые перестановки, купля-продажа, договора, судебные разбирательства*), новостная кластеризация, рубрицирование;
- **для научно-технических отчетов и статей:** построение тезаурусов понятий и связей предметной области, выявления направлений проводимых исследований и достигнутых результатов, специалистов в соответствующих областях, распознавание ссылок на публикации и построение индексов цитирования, выявление плагиата и повторений научных исследований;

- **для социальных сетей Интернета**: анализ мнений, предпочтений и интересов;
- **для нормативно-правовых документов**: идентификация в тексте ссылок на документы, поиск похожих судебных решений, вымарывание персональных данных, автоматическое рубрицирование;
- **для решений арбитражных судов**: выявление атрибутов решения, выявление участников процесса и их ролей в процессе, а также прочих упомянутых лиц и объектов, выявление атрибутов участников процесса и прочих лиц, выявление связей между участниками процесса и прочими лицами;
- **для досье, биографий, технических описаний, "карточек" проектов и прочих документов специального вида**: извлечение фактографических данных, стандартизация и ввод в базу данных. Собственная система ведения досье с возможностью автоматической «пробивки» объекта по различным источникам;
- **для запросов к поисковым машинам**: разбор запроса на естественном языке и трансляция в релевантные запросы на языке поисковика, расширение слов запроса всеми грамматическими формами;
- **для записей в базе данных**: стандартизация записей ФИО и наименований организаций, извлечение реквизитов из несоответствующих им полей, идентификация записей о физических и юридических лицах с опорной базой, формирование единого реестра физических и юридических лиц из различных источников, нормализация российских почтовых адресов, разбор строковых полей типа «назначение платежа».

Области применения по задачам

Технологии RCO позволяют решать с высоким уровнем качества следующие задачи, без которых трудно себе представить современную систему, предназначенную для работы с текстом на естественном языке:

- **содержательный портрет текста**: Построение информационного портрета документа, который характеризовал бы в компактной форме основное содержание текста - описанные в нем предметы, лица, ситуации и т.п. Позволяет находить похожие документы, производить автоматическую категоризацию и кластеризацию документов; автоматически строить глоссарии, частотные словари терминов;
- **упоминания персон и организаций**: Распознавание и разбор наименований объектов с выделением всех элементов наименования (ФИО, ОПФ, форма хозяйственной деятельности, название, география и т.д.), отождествление различных вариантов наименования одного и того же объекта в тексте, в том числе косвенных обозначений, не содержащих в себе имени собственного;
- **упоминания особых объектов**: Распознавание объектов, отличающихся специального вида написанием – почтовые адреса, идентификационные и паспортные данные, марки товаров и модели устройств и т.п. Используется язык, который позволяет оперировать как формальными особенностями написания текста, так и всеми грамматическими атрибутами слов. Образцы сложных конструкций могут строиться иерархически, включая образцы более простых. Возможно как бесконтекстное, так и контекстно-зависимое распознавание;
- **связи между объектами в тексте**: Выявление связи между описанными в тексте событиями, именованными и неименованными сущностями. Сеть связей,

построенная по коллекции документов, помогает при поиске заранее неизвестной информации, служит основой для решения различных аналитических задач;

- **распознавание ситуаций в тексте**: Позволяет найти в тексте описания ситуаций нужного типа, выделить всех участников ситуации в соответствии с их ролями, классифицировать описания ситуаций, сгруппировав их по заданным критериям;
- **отношение к объекту в тексте**: Анализ текста на предмет выражения в нем положительного или отрицательного отношения к объекту. Позволяет выявить как явную характеристику объекта с использованием тонально-окрашенной лексики, так и неявную характеристику объекта, связанную с упоминанием в тексте таких ситуаций, при восприятии которых возникает эмоциональная реакция;
- **анализ предметной области**: Методика автоматизированного построения онтологии, которая позволяет выявлять следующие составляющие онтологии: термины, названия артефактов, атрибуты и характеристики объектов, ситуации;
- **разбор частично-структурированного текста**: Комплексная обработка частично-структурированных документов. Автоматическая идентификация типов входных документов и блоков текста. Извлечение из блоков требуемых сущностей и связей как на основании формальных признаков, так и на основании лингвистического анализа текста на естественном языке;
- **категоризация текстов**: Упорядочение информационного массива, когда документы, близкие по определенным содержательным критериям, объединяются в группы, называемые категориями, рубриками, тематическими подборками. Разработана методика формирования описаний категорий – профилей;
- **кластеризация новостей**: Связывание сообщений, описывающих одни и те же события, в кластеры – сюжеты, и ведение сюжетной линии во времени. Построение обзорных рефератов, категоризация сюжетов, поиск похожих сюжетов. Значительно повышает эффективность анализа информационного потока;
- **поиск документов**: С учетом словоформ (в том числе и для слов, отсутствующих в словаре), с учетом опечаток, поиск на естественном языке, поиск похожих фрагментов, поиск с учетом семантической разметки (индексируются произвольные объекты, концепты, возможно использование в запросе одновременно ограничений на концепты и контекст: <компания> РЯДОМ «банкротство» - вернет список компаний, рядом с которыми было слово «банкротство»). Собственная поисковая машина с возможностью контент-анализа в реальном времени, с поддержкой многостраничных документов;
- **сопутствующие решения**: Распознавание языка и кодовой страницы документа; Извлечение требуемых блоков текста, очистка от элементов оформления и навигации; Обнаружение информационных дублей; Реферирование текста, в том числе по контекстному запросу; Подсветка найденных в документе слов; и многое другое.

Анализатор текста компании RCO

По данным различных источников, более 90% используемой корпоративной информации представлено в виде текста на естественном языке. Для эффективного использования неструктурированных данных в корпоративных информационных системах необходимо использовать инструментарий, позволяющий решать следующие основные задачи: классификация и поиск текстов, извлечение из текстов информации и представление ее в структурированном виде. Анализатор текста на естественном языке – это программный компонент, интегрируемый в корпоративную информационную (информационно-аналитическую) систему и осуществляющий обработку неструктурированных данных для обеспечения ее функций.

Флагманским продуктом компании является анализатор текста RCO Fact Extractor SDK. Продукт представляет собой комплексный инструментарий для разработки информационно-поисковых и аналитических систем, требующих лингвистического анализа текста на русском, английском, украинском, казахском и армянском языках. Продукт производит выделение различных классов сущностей, упомянутых в тексте, и строит сеть отношений, связывающих эти сущности, а также предоставляет всю грамматическую информацию о составляющих анализируемого текста.

Продукт, как правило, используется для построения систем, решающих следующие типовые задачи:

- Мониторинг СМИ и социальных сетей;
- Разбор обращений клиентов и граждан;
- Расширение поисковых, навигационных и аналитических возможностей в существующих информационных системах;
- Обработка строковых полей в БД, очистка данных.

В зависимости от языка, в RCO Fact Extractor SDK доступны следующие функции:

	Русский	Английский	Украинский	Белорусский	Казахский	Киргизский	Армянский
Морфологический анализ	+	+	+	+	+	+	+
Простые объекты (выявление слов и словосочетаний)	+	+	+	+	+	+	+
Выявление сложных объектов (даты, денежные суммы, номера автомобилей, адреса и т.д.)	+	+	+	+	+	+	+
Именованные объекты (персоны, организации, география)	+	+	+	+	+	+	+
Синтаксический анализ	+	+	+	+	+	+	
Анафорические связи между объектами	+	+	+	+			
Поиск событий и фактов (с выявлением участников и ролей)	+	+	+	+	+	+	
Выявление тональности упоминаний объектов	+						

Пресса о компании

MskIT, 27.12.2007

Текстовый анализ обрел самостоятельность. RCO пустилась в свободное плавание

Дарья Тренина

<http://www.mskit.ru/news/n44422/>

...

Как сообщили в компании **RCO**, к моменту ее выхода из состава «Гарант-Парк-Интернет» уже фактически сложились независимые технологические процессы и сформировалась собственная клиентская база, поэтому было принято решение о начале самостоятельной работы. «Существование в рамках одной компании нескольких подразделений с существенно различающимися видами деятельности не оптимально, как с точки зрения продвижения на рынке, узнаваемости бренда, так и по прозрачности и привлекательности для потенциальных инвесторов», - прокомментировал генеральный директор **ООО «ЭР СИ О» (RCO) Владимир Плешко**.

...

Digest.CNews, 28.01.2008

Издание CNews запустило новую услугу для своих читателей - новостной агрегатор

Сергей Ершов

http://www.cnews.ru/news/top/digest_cnews_polnyj_monitoring_itpressy

...

«Развивая лингвистические и математические методы на протяжении почти десяти лет и обеспечивая такими программными модулями не одну организацию, нам, в то же время, по понятным причинам редко удавалось видеть конечные прикладные системы и результат их работы, — рассказывает генеральный директор компании **RCO Владимир Плешко**. — Это наш первый публичный интернет-проект, и мы надеемся, что в его следующую версию будут включены и иные наши разработки, которые позволят продемонстрировать новые возможности новостных порталов».

...

РБК, 02.02.2015

Горе от ума: как заработать на искусственном интеллекте

Андрей Бабицкий

<http://www.rbc.ru/business/02/02/2015/54ceaf759a79472263c89c57>

В деле извлечения фактов у ABVYU есть конкуренты. Один из них — знаменитая программа Watson, созданная IBM (два года назад она победила живого чемпиона в «Свою игру»). Конкуренция с гигантом не слишком пугает Яна : рынок очень большой.

...

Есть конкуренты поменьше. В компании **RCO**, которая называет своими клиентами Центробанк, «Газпром» и ФСБ, работает 30 человек, но ее руководитель Владимир Плешко не боится ABVYU. «Они росли в оранжерейных условиях», — уверен он. 10



лет назад **Плешко** увидел телеинтервью Давида Яна, который рассказывал о своей революционной системе анализа текста. «Я подумал: нам конец. Прошел год, другой, третий, мы выпустили собственную систему [извлечения фактов. – Прим. РБК], а у них так ничего и не появилось. То, на что они потратили 15 лет, мы сделали за пять усилий нескольких людей», – говорит **Плешко**.

...

Лингвисты **RCO** не пытаются создать универсальный искусственный интеллект, для каждого клиента делается своя система, которая удовлетворительно работает еще и потому, что анализирует специализированные тексты.

ВЕДОМОСТИ, 14.07.2015

Rambler покупает разработчика RCO

<http://www.vedomosti.ru/technology/news/2015/07/14/600485-rambler-pokupaet-razrabotchika-rco>

Rambler & Co стала владельцем 51% компании **RCO**, [сообщили](#) "Коммерсанту" в группе компаний, уточнив, что была выкуплена доля финансового инвестора.

...

RCO занимается разработкой софта интеллектуальной обработки текстов на русском и других языках.

...

Основные заказчики - крупные корпорации и госструктуры. **RCO** выступает для них и как поставщик ПО, и как разработчик информационно-поисковых и аналитических систем.

D-Russia, 30.03.2018

Реестр отечественного ПО пополнился впервые за четыре месяца

<http://d-russia.ru/reestr-otechestvennogo-po-popolnilsya-vpervye-za-chetyre-mesyatsa.html>

В Единый реестр российских программ для электронных вычислительных машин и баз данных 29 марта был внесен 171 новый продукт. Предыдущее пополнение реестра, [напомним](#), состоялось 11 декабря 2017 года.

...

Полный перечень добавленных в реестр программных продуктов

...

4336	RCO Zoom	Библиотеки подпрограмм (SDK), Лингвистическое программное обеспечение, Системы сбора, хранения, обработки, анализа, моделирования и визуализации массивов данных, Поисковые системы
------	--------------------------	---

...

RUSBASE, 07.02.2020

Мягкая киберугроза: как технологии борются с фейками в сети

Илья Калагин

<https://rb.ru/opinion/myagkaya-kiberugroza/>

Фейковые новости сегодня обсуждают на разных площадках и уровнях — от комментариев в Instagram до кабинетов министров. Насколько серьезно влияние фальшивок на развитие событий — вопрос дискуссионный, но запрос на их выявление и нейтрализацию вполне реальный.

...

В России подобные задачи могут решать организации, специализирующиеся на обработке естественного языка и/или анализе соцмедиа: наша компания, а также «Яндекс», ABBYY, **RCO**, Brand Analytics и «Крибрум».

...

RUSBASE, 04.08.2021

Компания RCO создаст сервис для мониторинга за биткоин-транзакциями

[Карина Пардаева](#)

<https://rb.ru/news/bitcoin-monitoring-service/>

Росфинмониторинг выбрал подрядчика для создания модуля для мониторинга криптовалютных транзакций. Им стала подконтрольная «Сберу» компания **RCO**. Она выполнит работы за 14,7 млн рублей.