DEVELOPMENT OF THE INFORMATION EXTRACTION SYSTEM FROM TEXTS IN RUSSIAN FOR SUBJECT DOMAIN CRIMINALISTICS

N. O. Krutikov, N. G. Podakov, V. A. Zhilyakova

Novosibirsk national research state university,
630090, Novosibirsk, Russian Federation

This article describes an approach to Russian language information extraction systems development presented for subject domain Criminalistics. At first, we should describe the task in details. The developed system should extract named entities from the text, such as people and organizations, and events. Also, attributes of the extracted entities, such as name, gender and date of birth for individuals and name and type of organization, time and place for the event should be filled. Relationships between named entities and events should be extracted, semantic role should be defined for each dependent entity (for example "subject" and "object" of event). Different semantic entities that describe single real object (person, organization or event) must be glued together by resolving coreference, their attributes should be united. For text analysis RCO FX Ru library is used in system. This library used rule-based approach and provides the following results: list of the extracted from the text semantic entities, their morphological and syntactic attributes and semantic graph of each proposal. To resolve the problem corpus of texts has been builded, domain ontology has been developed and a system of rules and patterns based on the RCO FX has been realized. After processing texts are translated to RDF structure and saved in RDF store. Also, the system has visualization module that allows the user to view text analysis results, search among the extracted information, and use a variety of filters that discard the most important information. The approach allows extract information from texts with level precision 70–80

Key words: information extraction, rule-based approach, CAPE, named entity, event extraction, relation extraction, ontology.

References

1. RCO Fact Extractor SDK. [Electron. resource]. http://www.rco.ru/?page_id=3554 (Accessed 11.05.2016).

2. Tomita-parser / Yandex. Website of the Tomita-parser technology. [Electron. resource]. https://tech.yandex.ru/tomita/ (Accessed 11.05.2016).

3. GitHub - yandex/tomita-parser / GitHub, Inc. Open source code of the Tomita-parser project. 2016. [Electron. resource]. https://github.com/yandex/tomita-parser/ (Accessed 11.05.2016).

4. About ABBYY Compreno technology / ABBYY. Describing ABBYY Intelligent Search SDK technology. 2016. [Electron. resource]. http://www.abbyy.ru/isearch/compreno/ (Accessed 11.05.2016).

5. PolyAnalyst – data analysis. Text analysis / Megaputer Intelligence, Inc. Website of the PolyAnalyst product 2015. [Electron. resource]. http://megaputer.ru/polyanalyst.php (Accessed 11.05.2016).

6. Apache Jena / The Apache Software Foundation. Website of the Apache Jena project. 2015. [Electron. resource]. https://jena.apache.org/ (Accessed 11.05.2016).

7. OpenRdf Sesame. [Electron. resource]. http://www.openrdf.org/ (Accessed 01.03.2016).

8. dotNetRdf – Semantic Web, RDF and SPARQL Library for C#/.NET / Rob Vesse. Website of the dotNetRdf project. 2015. [Electron. resource]. http://dotnetrdf.org/ (Accessed 11.05.2016).

9. Kormalev D. A. Obobshenie i specializatsiya pri postroenii pravil izvlecheniya informatsii // Conference. KII–2006. T. 2. M.: Phismatlit, 2006. P. 572–579.

10. Kurshev E. P., Kormalev D. A., Suleymanova E. A., Trofimov I. V. Issledovanie metodov izvlecheniya informatsii iz tekstov s ispolzovaniem avtomaticheskogo obucheniya i realizatsiya issledovatelskogo prototipa sistemi izvlecheniya informatsii // Matematicheskie metodi raspoznovaniya obrazov: 13 All-Russian Conference. Leningrad region, Zelenogorsk, 30th of September – 6th of October 2007: Reports collection. M.: MAKS Press, 2007. P. 602–605.

11. Ermakov A. E. Izvlechenie znaniy iz texta i ih obrabotka: sostoyanie i perspektivi // Information technologies. 2009. N 7.

12. Simakov K. V. Modeli i metodi izvlecheniya znaniy iz tekstov na estesstvennom yazike. Candidate of Technical Sciences dissertation: 05.13.17. M. 2008.

13. Andreev A. M., Berezkin D. V., Simakov K. V. Metod obucheniya modeli izvlecheniya znaniy iz estesstvenno yazikovih tekstov // Vestnik MGTU. Instrumentation. 2007. № 3. P. 75–94.

14. Tolpegin P. V. Novie metodi i algoritmi avtomaticheskogo razresheniya referentsii mestoimeniy tretiyego litsa russkoyazichnih tekstov. M.: KomKniga, 2006. P. 88.

15. The results of the competition Dialogue 2016 on the extraction of named entities [Electron. resource]. http://pullenti.ru/DownloadFile.aspx?file=FactRuEval.pdf // (Accessed 29.05.2016).

16. Brat rapid annotation tool [Electron. resource]. http://brat.nlplab.org/ (Accessed 29.03.2016).

17. Conference materials DIALOGUE 2014 [Electron. resource]. http://www.dialog-21.ru/dialogue2014/results // (Accessed 29.05.2016).

18. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian [Electron. resource]. http://www.dialog-21.ru/digests/dialog2014/materials/pdf/ToldovaSJu.pdf // (Accessed 29.05.2016).

РАЗРАБОТКА СИСТЕМЫ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ ИЗ ТЕКСТОВ НА РУССКОМ ЯЗЫКЕ В ОБЛАСТИ КРИМИНАЛИСТИКИ

Н. О. Крутиков, Н. Г. Подаков, В. А. Жилякова

Новосибирский национальный исследовательский государственный университет,
630090, Новосибирск, Россия

УДК 004.852

Представлен подход к созданию системы извлечения информации из текстов на русском языке, относящихся к предметной области „Криминалистика“. Для решения задачи был построен и размечен корпус, разработана онтология, на основе RCO FX реализована система правил и шаблонов, извлекающих необходимые сущности, события и связи между ними. Текст преобразовывается в RDF-структуру, к которой можно обращаться при помощи графического пользовательского интерфейса. Подход позволяет извлекать информацию из предметных текстов с точностью 70–80 % при полноте 30–35 %.

Ключевые слова: извлечение информации, правила, именованные сущности, события, отношения, онтологии.

Введение. Одна из самых важных задач криминалиста – это поиск информации о некотором событии по базе документов или из открытых источников, таких как Интернет. Необходимо узнать, что произошло, время, место события, установочные данные физических и юридических лиц, принимающих участие в данном событии; собрать информацию, которая касается физических и юридических лиц: что собой представляет конкретный человек, где работал и чем занимался, пол, возраст, когда и почему упоминался в других событиях базы. Решение таких задач традиционными методами поиска не всегда может удовлетворить пользователя полнотой и точностью, в то время как ручная обработка документов невозможна в связи с большими объемами базы документов.

1. Постановка задачи. Требуется разработать специализированную программную систему извлечения информации из текстов на русском языке в предметной области криминалистики TextPro. Заранее известно, что обрабатываемые тексты относятся к криминальной тематике, принадлежат к публицистическому стилю (криминальные сводки, новостные статьи).

Система должна предоставлять пользователю возможности пакетной обработки текстовых файлов, извлечения данных из текстов, а также визуализации полученных данных. Обработанные тексты представляются в системе в виде семантического графа, вершинами которого являются именованные сущности и события, а ребрами графа являются отношения между ними. Атрибуты сущностей и событий не являются отдельными вершинами графа.

Работа выполнена при финансовой поддержке Минобрнауки РФ (договор № 02.G25.31.0146).

Для комфортной работы пользователя с таким представлением нужно уметь строить досье текста и каждой вершины, входящей в него. Поэтому система должна иметь графический интерфейс, позволяющий просматривать досье и семантический граф текста. Должна быть возможность осуществлять поиск по графам обработанных текстов как при помощи запроса SPARQL, так и при помощи удобного графического пользовательского интерфейса.

2. Обзор методов решения задачи. На данный момент существуют три основных направления в задаче извлечения информации: статические методы, методы на основе правил и на основе машинного обучения [1–5].

Статические методы – наиболее простые и справляются с задачей извлечения информации, когда исходный текст представлен в структурированном виде. Например, задача извлечения полей в анкете. Часто для работы можно ограничиться данным методом. Однако он крайне узкоспециализирован и не позволяет обрабатывать тексты, которые изначально не имеют жестко заданной структуры.

Методы, основанные на правилах, более универсальны и позволяют обрабатывать тексты, не имеющие жесткой структуры. Описание текстовых ситуаций выполняется при помощи правил на специальном языке. Обычно правило состоит из двух частей: образец, которому должна соответствовать речевая ситуация, и действие, которое правило должно совершить с данной ситуацией. Такой подход подразумевает разработку правил лингвистами и специалистами в конкретной предметной области. Правила в результате достаточно общие, чтобы находить схожие речевые ситуации, но достаточно специализированные, чтобы не срабатывать на ложных.

В связи с трудоемкостью построения правил получило развитие *машинное обучение*. Одним из вариантов его применения является автоматизированное создание правил на основе размеченного текстового корпуса. Другой вариант – автоматическое наполнение словарей. Преимуществом такого подхода перед ручным созданием правил является отсутствие необходимости в большом количестве лингвистов, разбирающихся в определенной предметной области, а также быстрая адаптация к новым речевым ситуациям. Для этого достаточно разметки текстового корпуса. Данный подход позволяет сэкономить на разработке правил, но требует тщательного выбора критериев для разметки больших текстовых корпусов.

3. Обзор готовых решений. Процесс извлечения информации из текстов состоит из следующих этапов:

1) графематический анализ;

2) морфологический разбор;

3) синтаксический разбор;

4) семантический разбор.

Задача извлечения событий, именованных сущностей, а также разрешения кореференции возникает на уровне семантического разбора.

Существуют три наиболее распространенных инструмента для извлечения информации на русском языке. Каждый из них по-разному справляется с каждым их представленных выше четырех этапов.

3.1. *Яндекс Томита-парсер.* Исходный код проекта Томита-парсер [6] открыт [7], допускается бесплатно использовать продукт в коммерческих целях. Извлечение фактов происходит при помощи правил, записанных на языке расширенных контекстно-свободных

грамматик, и словарей ключевых слов. Парсер позволяет конструировать собственные правила и добавлять словари для нужного языка.

Томита-парсер включает в себя три стандартных лингвистических процессора: сегментатор (разбиение текста на предложения), токенизатор (разбиение на слова), и морфологический анализатор Mystem (присвоение словам морфологических характеристик). Основные компоненты Томита-парсера: газеттир, набор правил и множество описаний типов фактов, которые порождаются правилами в результате процедуры интерпретации.

Томита-парсер расширяем, однако высокий порог вхождения для написания собственных правил плохо сказывается на скорости, качестве разработки и поддержке.

3.2. *ABBYY Compreno.* В процессе полного семантико-синтаксического разбора ABBYY Compreno [8] определяет семантические значения слов в тексте и выделяет все связи между ними, создавая не зависящее от языка представление текста.

ABBYY Compreno преобразует тексты в синтактико-семантические деревья Compreno. Узлы деревьев соответствуют словам или коллокациям в предложении, а дугами обозначаются отношения между ними с точки зрения грамматики зависимостей. Деревья являются формальным представлением информации, заключенной в высказывании, поэтому на данном этапе в них уже не возникает проблем с решением такой неоднозначности слов, как омонимия.

Система преобразует эти деревья в сущности или факты. Таким образом, ABBYY Compreno создает универсальное и не зависящее от языка представление текста.

Compreno является продуктом, предоставляющим качественный текстовый анализ на корпусах общей тематики, обладает хорошими словарями и правилами, однако не позволяет расширять библиотеку правил для более точной специализации инструмента под конкретную тематику.

3.3. *RCO Fact Extractor SDK.* RCO Fact Extractor SDK [9] представляет собой набор инструментов и библиотек, позволяющих разрабатывать информационно-поисковые и аналитические системы, требующие поддержки лингвистического анализа текстов на русском языке.

Ядром SDK является библиотека RCO FX Ru, которая осуществляет полный синтактико-семантический разбор русского текста. Библиотека строит семантико-синтаксический граф, выделяет различные классы сущностей, упомянутых в тексте (персоны, организации, география, предметы, действия, атрибуты и др.), и строит сеть отношений, связывающих эти сущности, частично исправляет грамматические ошибки и опечатки, а также предоставляет всю грамматическую информацию о составляющих текста.

Система использует для описания распознаваемых в тексте конструкций язык Cape, подобный Jape. Язык расширен встроенными средствами создания морфологических характеристик выделяемых языковых конструкций.

В состав лингвистического обеспечения пакета, помимо общих словарей и правил русского языка, входят правила выделения специальных объектов (дат, адресов, документов, телефонов, денежных сумм, марок автомобилей и пр.), правила для распознавания различных классов событий и фактов (сделок, экономических показателей, конфликтов, биографических фактов и пр.), высказываний прямой и косвенной речи.

Часть встроенных библиотек возможно менять и расширять, часть библиотек зашифрована. Это препятствует оценке их полноты. Есть графическое средство для разработки правил – RCO Fact Tuner. Оно нестабильно, но позволяет визуализировать структуру

шаблонов и упростить работу над их созданием. Кроме того, для работы с системой требуется платная подписка, благодаря которой большинство библиотек и словарей можно получить в открытом виде.

RCO Fact Extractor создавалась как система извлечения информации из текстов экономической и общей тематики. Для поставленной задачи система не позволяет извлекать информацию с достаточной полнотой по причине недостаточного количества специализированных правил и словарей криминалистической тематики. Не развита классификация событий по семантическому признаку, а существующая классификация не имеет иерархии. Практически не обрабатываются случаи кореференции в тексте, поэтому при построении синтаксико-семантического дерева возрастает количество сущностей, упомянутых в тексте, в то время, когда в тексте их количество было значительно меньше. Помимо этого, у RCO Fact Extractor возникают сложности при обработке сложносочиненных и сложноподчиненных предложений. Нередко при построении синтаксического графа предложения получается лес – набор несвязанных между собой деревьев.

3.4. *Вывод.* Ни одно из рассматриваемых средств не является решением поставленной задачи. Тем не менее, расширяемость RCO FX Ru, наличие инструментов для доработки, возможность построения синтактико-семантического графа, большое количество встроенных правил и словарей позволяет использовать его в качестве основы для разработки модуля, ответственного за синтаксическую и морфологическую обработку текста, извлечение и связывание синтактико-семантических сущностей и отображения их в граф знаний.

Для решения проблем RCO FX Ru необходимо разработать правила, связывающие сущности в предложении, разработать словари для классификации сущностей и событий по семантическому признаку, улучшить алгоритмы разрешения анафоры.

4. Проектирование. Для анализа текстов мы спроектировали следующую систему, схема которой изображена на рис. 1. Ядро принимает входные данные и при помощи библиотеки анализа текстов (RCO Fx Ru) обрабатывает их. Для обработки текстов библиотека анализа использует хранилище шаблонов, правил и словарей. После этого ядро, задействовав внешний классификатор, разрешает кореференцию, а полученный граф транслируется в хранилище онтологических структур.

4.1.*Корпус.*

4.1.1. *Сбор данных для разметки.* Разработка правил требует наличия корпуса, на основе которого правила создаются. Корпус должен представлять модель языка, поэтому, если в рабочих текстах конкретная речевая ситуация встречается с определенной вероятностью, то и в текстах документов корпуса подобная ситуация должна встречаться с той же вероятностью. Это означает, что далеко не каждый корпус подходит для задачи обучения системы.

Так как по выбранной тематике не было найдено ни одного корпуса в общем доступе, а существующие корпусы не удовлетворили требованиям системы, было решено создать и разметить свой. За основу были взяты статьи новостных сайтов lenta.ru, lib.ru, ria.ru, news.ru, chaskor.ru, так как они содержат большой объем криминальных сводок, существуют достаточно давно и имеют структуру, которая легко поддается извлечению данных. С каждого сайта было получено по 1000 последних новостей, после этого были удалены все тексты, не связанные с преступлениями.

При создании корпуса мы столкнулись со следующими проблемами: отсутствие инструментов коллективной работы для формирования корпусов на русском языке и неоднозначность разметки разными экспертами. Первая проблема решалась использованием brat

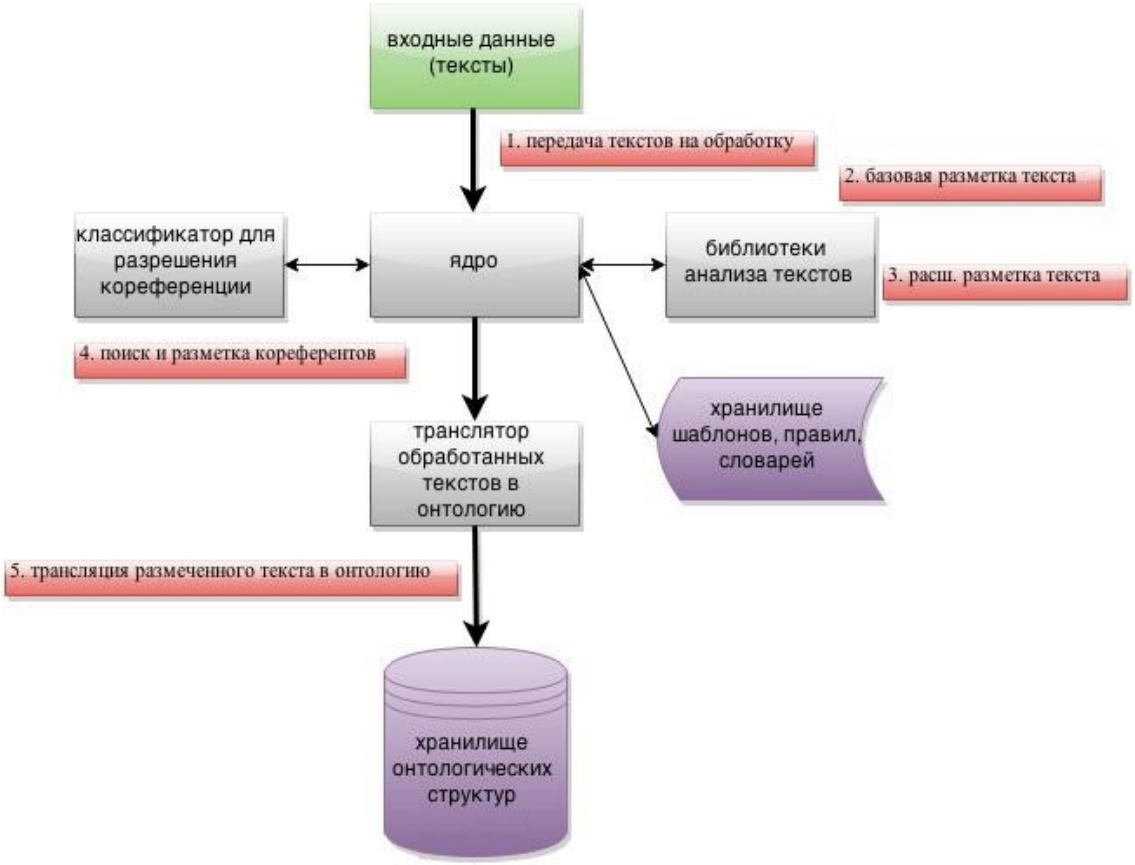

Рис. 1. Архитектура разработанной системы

[10] и обработкой его результатов, а позже – разработкой собственного средства разметки, дополняемого функционалом по мере необходимости, вторая – перекрестной классификацией текстов. Таким образом, был разработан собственный корпус текстов для решаемой задачи объемом 4566 текстов.

4.2. *Разметка и разработка онтологии по предметной области „Криминалистика“.* Для создания онтологии мы воспользовались эмпирическим подходом: собрали информацию о типах сущностей, событиях и их атрибутах на ограниченном корпусе, проанализировали результат, разработали онтологию, в которую укладывается информация о типах, применили онтологию на весь текстовый корпус.

Несколько человек разметили 200 случайных текстов вручную независимо друг от друга. Цель разметки – создание списка именованных сущностей и их предположительных классов, фактов (событий) и их классов, а также атрибутов. Примерами таких классов являются: физическое лицо, юридическое лицо, мошенничество, грабеж; примерами атрибутов: ФИО физического лица, полное наименование юридического лица, время и место совершения преступления. После обработки результатов была создана онтология (рис. 2, 3), содержащая окончательный список классов, исходя из количества встреченных сущностей и событий такого класса в выборке, а также набор отношений между сущностями.

Далее каждый текст из корпуса вручную размечается уже согласно полученной онтологии. В текстах выделяются сущности и события, указываются их типы, события свя-

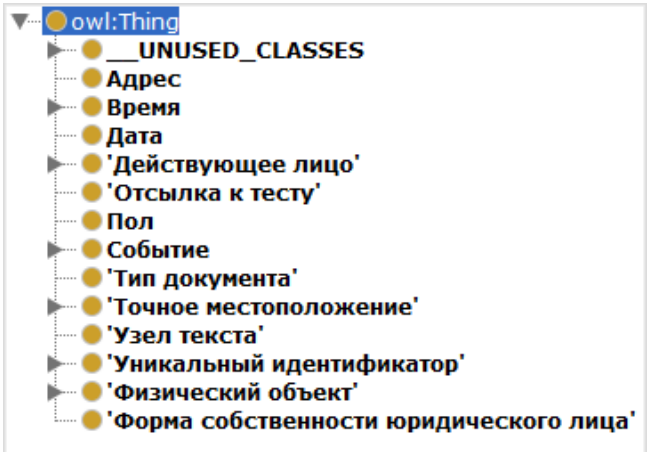

Рис. 2. Верхний уровень онтологии

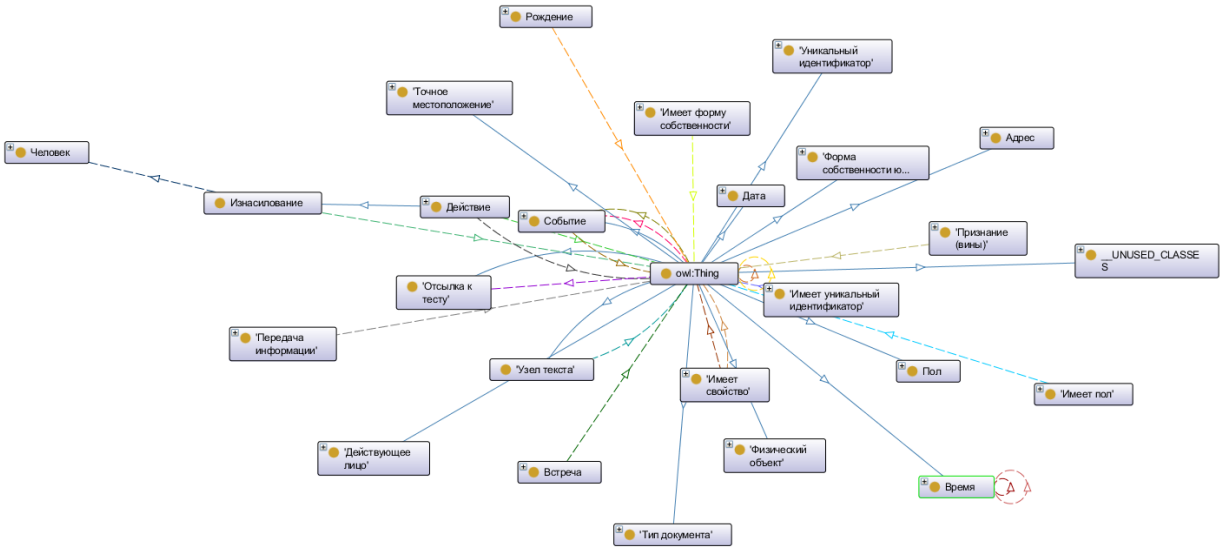

Рис. 3. Граф верхнего уровня онтологии

зываются с сущностями, указываются атрибуты. Разметка проводилась в программном средстве brat, так как оно позволяет указывать список типов, а также имеет удобную и наглядную структуру выходного файла.

Эксперты могут иметь различные точки зрения на разметку некоторых речевых ситуаций, а также могут ошибаться, поэтому после разметки была проведена перекрестная валидация.

В разработанной онтологии все события представляются узлами, т. е. классами онтологии. Индивиды – это объекты классов; каждое конкретное событие и сущность являются индивидами. Участники событий связываются с самими событиями при помощи отношений объектов (object property, ребра в терминах графовых БД). Также имеются отношения данных (data property, свойства узлов в терминах графовых БД), они позволяют хранить данные примитивных типов, при помощи этих отношений хранятся атрибуты индивидов.

5. **Дополнение ресурсов библиотеки RCO FX Ru.**

5.1. *Типизация выделяемых сущностей при помощи словарей.* Библиотека RCO FX Ru выделяет такие основные типы сущностей, как Event (событие), Person (человек), Organization (организация), Date (дата или время), Geoplace (место). Но для заполнения построенной онтологии требуется выделять большее количество семантических типов.

Процесс создания словаря состоит из нескольких этапов:

1) определяется название типа, который необходимо выделять;

2) посредством поиска в словарях синонимов находятся слова или словосочетания, которые представляют собой семантические сущности с искомым типом;

3) проверяется качество выделения этих слов и словосочетаний, а также слов и словосочетаний, полученных путем изменения морфологических атрибутов исходных слов;

4) по результатам проверки словари дополняются элементами, полученными путем изменения морфологических атрибутов. Также в случае некорректного определения морфологические атрибуты для некоторых элементов задаются вручную в словарях.

Для решения этой задачи было создано 16 словарей, содержащих в общей сложности 564 цепочки лексем и выделяющих 16 новых типов семантических сущностей, среди которых события криминальной направленности (убийства, кражи, похищения и т. д.), люди (словари для выделения людей, которые являются жертвами преступлений, людей, которые являются сотрудникам правоохранительных органов, преступников) и организации (словари для правоохранительных организаций и преступных группировок).

5.2. *Улучшение качества разбора текстов при помощи правил.* Основной инструмент улучшения качества разбора текстов – правила на формальном языке Jape. Каждое правило состоит из двух частей. Первая часть правила представляет собой описание цепочки лексем, которую необходимо найти в тексте. Для каждой лексемы можно задавать ограничения на значения ее морфологических и синтаксических атрибутов. Также можно использовать простые логические операторы. Правая часть представляет собой описание сущности, которая будет выделена, с заданием ее атрибутов.

Для улучшения качества разбора текстов были разработаны правила, выделяющие языковые конструкции, которые наиболее часто встречались в текстах и некорректно выделялись библиотекой RCO FX Ru (наиболее частые случаи – языковые конструкции, содержащие специальные символы).

Были созданы наборы новых правил, выделяющие более 20 различных форм описания дат, наиболее часто встречающихся в собранном корпусе текстов, и определяющие для них 8 новых атрибутов. Впоследствии эти атрибуты используются для семантического поиска по датам.

5.3. *Выделение связей между сущностями.* Для выделения связей между сущностями из текстов использовался такой инструмент библиотеки RCO FX Ru, как семантические шаблоны. Семантический шаблон представляет собой набор ограничений на синтактико-семантический граф разбора предложения. В ограничениях могут использоваться простые логические операторы, а также, помимо обычного сравнения значений атрибутов, могут использоваться сравнение с учетом морфологии и сравнение с регулярным выражением.

Были разработаны несколько групп шаблонов:

1) шаблоны для поиска выделенных типизированных сущностей (людей, организаций, событий, дат);

2) шаблоны для поиска выделенных связей между сущностями и их атрибутами;

3) шаблоны для выделения связей, которые не были найдены библиотекой RCO Fx Ru;

6. Разрешение местоименной анафоры 3 лица. Если не обработать случаи кореференции в тексте, то получаем множество „двойников“ одной и той же сущности. Поэтому информация, относящаяся к одному объекту, будет распределена по нескольким. Для решения этой проблемы мы обратились к результатам участников конференции „Диалог“ [11].

В 2014 г. в рамках конференции „Диалог“ в секции „Соревнование по оценке систем автоматического разрешения анафоры и кореферентности“ принимали участие такие команды, как ABBYY, RCO, MAIL.RU и т. д. Лучший результат решения местоименной анафоры для 3 лица: точность 80 %, полнота 65 % [12].

Целью разработки модуля разрешения местоименной анафоры 3 лица было получение схожих результатов по точности, по полноте не ниже 20 %. Для обучения использовали корпус „Диалога“, потому что в данный момент существует мало размеченных текстов на предмет анафорической кореференции.

Корпус представляет собой файл разметки и набор текстовых файлов на русском языке, содержащих в себе анафоры. Файл разметки – файл в формате XML, содержащий путь к размечаемому текстовому файлу, пары анафор–антецедент, находящиеся в тексте файла, их местоположение в виде пары length-offset.

Корпус обрабатывается RCO. Происходит морфологический анализ каждой сущности в тексте. После этого ищутся соответствия извлеченных сущностей RCO и сущностей в терминах разметки при помощи пересечения на основании пары length-offset.

Далее происходит поиск анафоров по тексту. Если сущность является местоимением, считаем, что эта сущность – анафор, требующий распознания. Сопоставляем с ней истинную пару, полученную из размеченного корпуса

В данный момент модуль разрешения анафора умеет работать только с личными местоимениями третьего лица. Поэтому из списка рассматриваемых объектов убираем те, которые построены на основе других местоимений, так как они требуют других методов разрешения анафора.

Далее ищем гипотетические антецеденты. Для этого рассматриваем фрейм, содержащий в себе три предложения после анафора и одно до него. Во фрейме ищутся все сущности, главным словом которых являются существительные и прилагательные. Эти сущности считаем гипотетическими антецедентами. Только один из них является истинным, остальные – ложные.

Следующим этапом является определение признаков. В список признаков взяли все синтаксические, морфологические и семантические признаки, влияние которых на разрешение местоименного анафора было доказано в трудах Толпегина [13] и которые возможно извлечь средствами RCO. Кроме того, был создан ряд своих признаков, основанных на метриках – количестве слов, букв или специализированных сущностей, находящихся между анафором и антецедентом:

1) номер антецедента как слова в предложении;
2) расстояние в одушевленных сущностях;
3) расстояние в предложениях;
4) расстояние в существительных и местоимениях;
5) расстояние в глаголах;
6) расстояние в причастиях;
7) расстояние в местоименных прилагательных;
8) расстояние в подлежащих.

Таблица

Точность распознавания

Корпус	Число объектов	Точность распознавания, %
Диалог	86	70,9
Корпус lenta.ru	173	63,1
Корпус lib.ru	146	69,2
Корпус ria.ru	168	66,7
Совокупная точность распознавания		66,8

Для классификатора было решено использовать алгоритм RandomForest. В реализации библиотеки sklearn на Python необходима таблица булевых или числовых значений без пропусков. Каждый столбец – признак, строка – объект. Объектом является пара антецедент–анафор и все значения признаков для нее. Значения каждого признака-столбца либо полностью булевы, либо полностью численные.

При обучении получили точность на исходном корпусе $\approx$ 60 %, полнота $\approx$ 32–33 %. Однако вариант использования классификатора только на основании метрик не является верным. Следует сравнивать результирующие вероятности для различных антецедентов одного анафора между собой. После такой постобработки точность возросла до 67 %, полнота до 34 %.

7. Визуализация извлеченных сущностей в виде графа. В результате обработки текстов пользователю предлагается визуальное представление извлеченных данных в виде графа.

RCO FX Ru без дополнительных правил и словарей предоставит нам такую схему (рис. 4) разбора предложения:

15.01.2015 года около 18:00 час. из квартиры дома 73 по ул. Октябрьская гражданка К. похитила деньги в сумме 4 000 рублей.

Для сравнения, на рис. 5 то же самое предложение, обработанное системой семантического анализа Pullenti. В 2016 году в рамках Dialogue Evaluation проводилось соревнование по извлечению информации из новостных текстов на русском языке. Система Pullenti участвовала в соревновании на трех дорожках, на двух заняла первое место, на одной – второе [14].

Граф, построенный при помощи TextPro, показан на рис. 6.

Вершинами графа являются сущности и события, а ребрами – связи. При одинарном клике на вершину открывается досье сущности или события в той же вкладке сбоку от графа, а при двойном — открывается новая вкладка с досье.

Заметно, что и RCO Fx Ru без дополнительных библиотек, и граф Pullenti верно выделяют все сущности и события, однако они не находят связи между ними. А система TextPro и верно находит все сущности и события, и верно связывает их в синтаксико-семантический граф.

Заключение. Таким образом, был разработан инструмент, решающий задачу извлечения информации из текстов на русском языке в области криминалистики и предоставляющий пользователю следующую функциональность:

1) возможность обрабатывать блоки текстов;

2) сохранение результатов анализа в онтологическую базу данных;

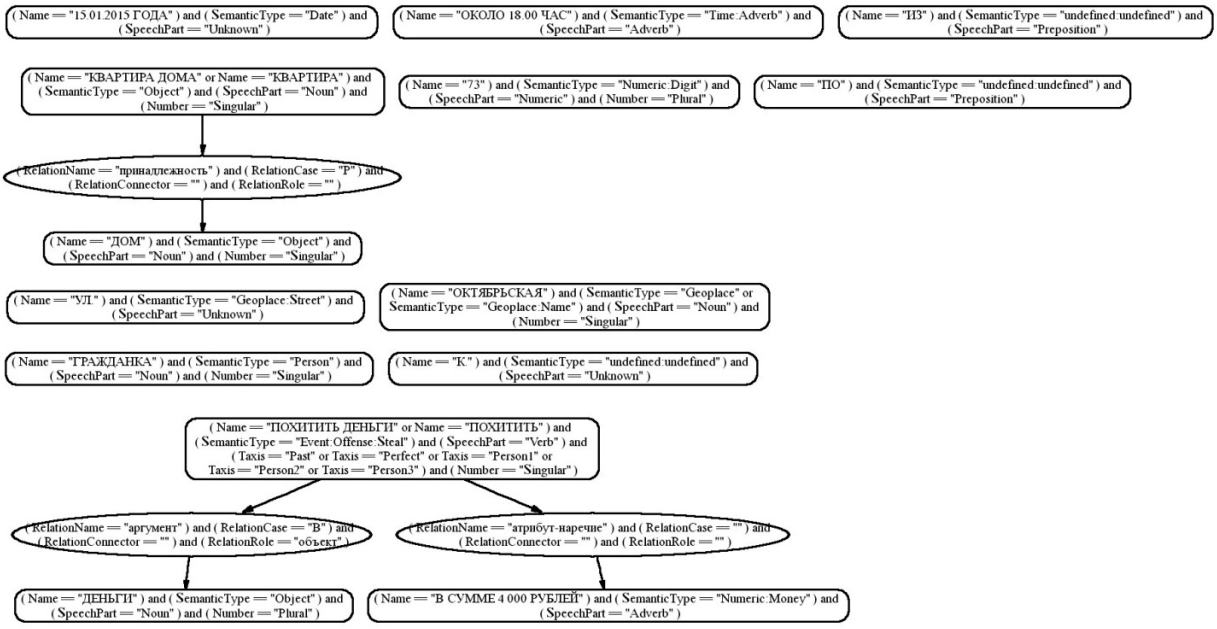

Рис. 4. Граф RCO FX Ru без дополнительных модулей

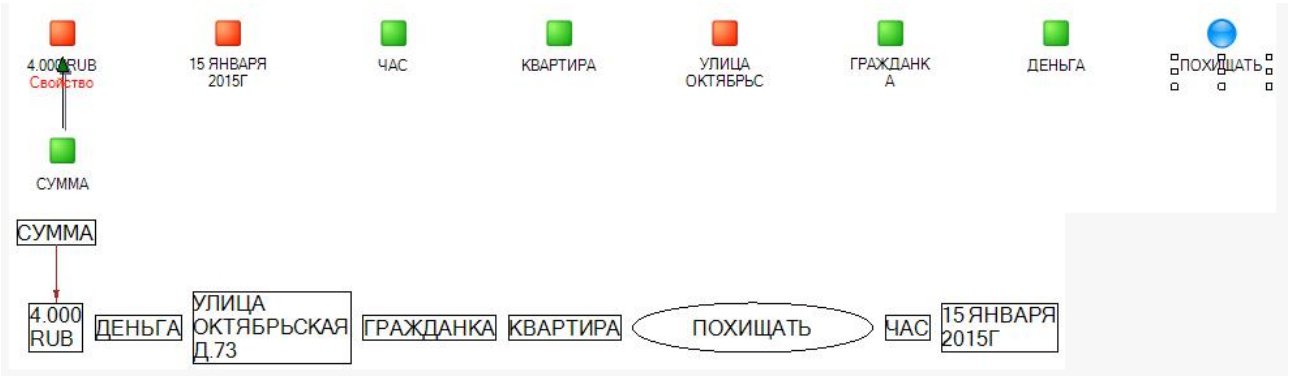

Рис. 5. Пример графа разбора текста Pullenti

3) возможность просматривать в графической оболочке досье текста (весь семантический граф разбора текста, все сущности, содержащиеся в тексте);

4) возможность просматривать в графической оболочке досье каждой сущности (все атрибуты сущности и набор сущностей, связанных с ней);

5) просмотр семантического графа текста с возможностью фильтрации результатов анализа по важности.

Разработанные 350 правил различной степени детализации позволили создать более глубокую, чем у RCO Fact Extractor, иерархическую систему типов объектов, событий и отношений, которые не покрывались стандартными шаблонами. Это открывает возможности поиска не только по ключевым словам, но и по смыслам.

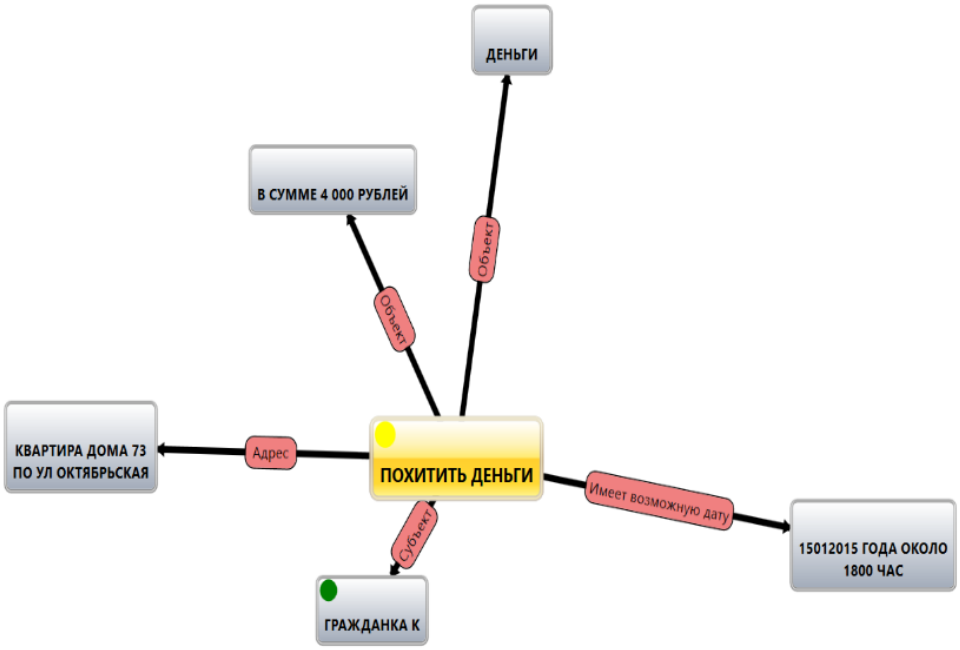

Рис. 6. Пример графа разбора текста TextPro

На тестовых текстах процент событий и сущностей, выделенных RCO Fact Extractor, которые представляют интерес для криминалистики, составляет 15–20 %. Трансляция в онтологию позволяет повысить точность до 25–30 %. При этом число напрасно выброшенных событий составляет менее 10 % от общего числа интересной нам информации.

В онтологии все события и сущности имеют иерархию. Если среди них рассмотреть только те сущности, которые имеют отношение к криминальной области, то точность возрастает до 70–80 %. При этом полнота падает до 30–35 %.

Среди найденных сущностей 5–10 % нужно объединить, так как они реферируют. RCO Fact Extractor справляется с этой задачей с точностью 90 %, полнотой 3 %, а разработанная система в частном случае, для местоимений 3 лица, с точностью 80 %, полнотой 20 %.

Список литературы

1. Кормалев Д. А. Обобщение и специализация при построении правил извлечения информации // Конф. КИИ–2006. Т. 2. М.: Физматлит, 2006. С. 572–579.

2. Куршев Е. П., Кормалев Д. А., Сулейманова Е. А., Трофимов И. В. Исследование методов извлечения информации из текстов с использованием автоматического обучения и реализация исследовательского прототипа системы извлечения информации // Математические методы распознавания образов: 13-я Всерос. конф. Ленинградская обл., г. Зеленогорск, 30 сентября – 6 октября 2007 г. Сборник докладов. М.: МАКС Пресс, 2007. С. 602–605.

3. Ермаков А. Е. Извлечение знаний из текста и их обработка: состояние и перспективы // Информационные технологии. 2009. № 7.

4. Симаков К. В. Модели и методы извлечения знаний из текстов на естественном языке: автореф. дис. канд. техн. наук: 05.13.17. М. 2008.

5. Андреев А. М., Березкин Д. В., Симаков К. В. Метод обучения модели извлечения знаний из естественно-языковых текстов // Вестник МГТУ. Приборостроение. 2007. № 3. С. 75–94.

6. Томита-парсер / Сайт технологии Томита-парсер. [Электронный ресурс]. https://tech.yandex.ru/tomita/ (дата обращения: 11.05.2016).

7. GitHub - yandex/tomita-parser / GitHub, Inc. Открытый исходный код проекта Томита-парсер. 2016. [Электронный ресурс]. `https://github.com/yandex/tomita-parser/` (дата обращения 11.05.2016).

8. О технологии ABBYY Compreno / ABBYY. Описание технологии ABBYY Intelligent Search SDK. 2016. [Электронный ресурс]. `http://www.abbyy.ru/isearch/compreno/` (дата обращения: 11.05.2016).

9. RCO Fact Extractor SDK / ООО „ЭР СИ О“. Сайт продукта „RCO Fact Extractor SDK“. 2016. [Электронный ресурс]. `http://www.rco.ru/?page\_id=3554` (дата обращения: 11.05.2016).

10. Brat rapid annotation tool. [Электронный ресурс]. `http://brat.nlplab.org/` (дата обращения: 29.03.2016).

11. Материалы конференции DIALOGUE 2014. [Электронный ресурс]. `http://www.dialog-21.ru/dialogue2014/results` // (дата обращения: 29.05.2016).

12. RU-EVAL-2014: Evaluating Anaphora and Coreference Resolution for Russian [Электронный ресурс]. `http://www.dialog21.ru/digests/dialog2014/materials/pdf/ToldovaSJu.pdf` // (дата обращения: 29.05.2016).

13. Толпегин П. В. Новые методы и алгоритмы автоматического разрешения референции местоимений третьего лица русскоязычных текстов. М.: КомКнига, 2006.

14. Результаты соревнований Диалог–2016 по выделению именованных сущностей [Электронный ресурс]. `http://pullenti.ru/DownloadFile.aspx?file=FactRuEval.pdf` // (дата обращения: 29.05.2016).

15. PolyAnalyst – Анализ данных. Анализ текста. Единый инструментарий / Megaputer Intelligence, Inc. Сайт продукта PolyAnalyst 2015. [Электронный ресурс]. `http://megaputer.ru/polyanalyst.php` (дата обращения: 11.05.2016).

16. Apache Jena / The Apache Software Foundation. Сайт проекта Apache Jena. 2015. [Электронный ресурс]. `https://jena.apache.org/` (дата обращения: 11.05.2016).

17. OpenRdf Sesame. [Электронный ресурс]. `http://www.openrdf.org/` (дата обращения: 01.03.2016).

18. dotNetRdf – Semantic Web, RDF and SPARQL Library for C#/.NET / Rob Vesse. Сайт проекта dotNetRdf. 2015. [Электронный ресурс]. `http://dotnetrdf.org/` (дата обращения: 11.05.2016).

Крутиков Никита Олегович окончил бакалавриат факультета информационных технологий Новосибирского государственного университета в 2014 году. Темой диплома была „Защищенная база данных. Реализация сохраняющего порядок шифрования. Вероятностный вариант“. Работа была выполнена во время работы в научно-исследовательской лаборатории Parallels NSU. В 2016 г. защитил магистерскую диссертацию по специальности 09.04.01 („Информатика и вычислительная техника“), тема работы „Разработка модуля извлечения информации из текстов новостных сводок на русском языке“. Работа была выполнена в рамках совместной работы компаний „Исследовательские системы“ и „Сигнатек“. С 2016 года работает в АО „Сбертех“.

Krutikov Nikita Olegovich received his bachelor degree in faculty of information

technology from Novosibirsk State University (2014). Subject of diploma was „The protected database. Realization of the order–keep encryption. A probabilistic approach“. This scientific work was written at the time of working in the laboratory Parallels NSU. Krutikov Nikita received his master degree in Computer Science at 2016. Subject of master dissertation was „Development of the module of extraction of information from texts of news reports in Russian.“ This scientific work was written within collaboration of the company „research systems“ and „signatec“. Krutikov Nikita works in company „Sberbank-Technology“ from 2016.

Никита Подаков получил степень магистра по специальности 09.04.01 (Информатика и вычислительная техника) Новосибирского государственного университета в 2016 году. С 2014 года работал в области программирования. С 2014 по 2015 год работал в компании „Исследовательские системы“, где принимал участие в работе над командным проектом „Разработка аналитического программного комплекса потоковой обработки данных в телекоммуникационных сетях с целью обеспечения информационной безопасности“, выполненном при поддержке Минобрнауки РФ.

Nikita Podakov received his M. S. degree in Computer Science from the National Research University of Novosibirsk (2016). From 2014 he had a job in software industry. From 2014 to 2015 he work in „Issledovatelskie systemi“ organization and participated in the work on a team project „Development of the information extraction system from texts in Russian for subject domain Criminalistics“ that was carried out with financial support from the Ministry of Education.

Жилякова Валерия Андреевна окончила гуманитарный факультет Новосибирского государственного университета в 2015 году. С 2014 по 2015 годы работала в компании „Экспасофт“ лингвистом и участвовала в разработке системы извлечения информации из текстов новостных сводок на русском языке.

Zhilyakova Valeria Andreevna received her M. S. degree in humanitarian faculty from Novosibirsk State University (2015). From 2014 to 2015 she worked in the company „Expasoft“ as the linguist and she participated in development of the information extraction system from texts of news reports in Russian.

Дата поступления — 02.06.2016