Система извлечения и поиска структурированной информации из больших текстовых массивов СМИ. Архитектурные и лингвистические особенности.

Л. М. Гершензон
ЗАО «Интегрум-Техно», Москва
leva@integrum.ru

И. М. Ножов
ЗАО «Интегрум-Техно», Москва
nozhov@integrum.ru

Д. В. Панкратов
ЗАО «Интегрум-Техно», Москва
pankrat@integrum.ru

Введение

Возрастающие потоки текстовой информации ставят новые задачи перед информационно-поисковыми системами. Простая выдача даже абсолютно релевантных запросу текстов не удовлетворяет потребности в получении актуальной и точной фактической информации. Во многих случаях сама эта выдача будет очень большой и в ней будет содержаться много дублированной информации. Information Extraction, Text Mining – извлечение информации из текста – это направления, которые призваны помочь справиться с большими текстовыми потоками профессиональным аналитикам. Общие подходы, которые использовались при создании описываемой в докладе системы, близки общепринятым принципам мировой традиции создания технологий IE [1].

Постановка задачи

Информационное агентство «Интегрум-Техно» (www.integrum.ru) предоставляет он-лайн доступ к текстовым массивам через собственную поисковую систему «Артефакт». Одна из самых частотных поисковых задач – это выяснить, что из себя представляет конкретный человек, где работал и чем занимался, кто в определенной компании занимается маркетингом или возглавляет отдел продаж, кому принадлежит конкретный завод или газета. Решение таких задач традиционными средствами поиска не всегда может удовлетворить пользователя своей полнотой и точностью. Это тот случай, когда пользователя интересует в первую очередь именно информация, а не множество документов, в которых эта информация содержится. Технологии выделения на этапе индексирования из текста объектов более крупных, чем слово и установление содержательных отношений между ними (не просто взаимная встречаемость в одном предложении или документе) могли бы существенно облегчить и убыстрить процесс поиска информации. Интеграция технологий извлечения фактов из текста с полнотекстовым информационным поиском, позволяет получить для определенных задач точность и лаконичность вопросно-ответных систем, а при традиционном поиске документов по запросу - большую полноту и удобство работы.

В компании Интегрум-Техно разработана автоматически пополняемая база данных «Кто? Где? Когда?», содержащая информацию о людях и организациях, связанных отношением «занимать должность». Создание такой базы потребовало разработки общих технологий по выделению из текста и хранению данных определенного вида. Таким

образом, для настройки системы на другую предметную область теперь требуется написание грамматик и наполнение словарей с помощью уже созданных средств.

Система состоит из двух основных частей: модуль работы с документами и модуль работы с базой данных. Первый модуль занимается только извлечением данных из входного потока документов. Второй модуль загружает эти данные в реляционную базу данных, осуществляет поиск по ней и кластеризует результаты.

Интересующее нас отношение Человек-Организация выражается в тексте двумя способами:

1) Именная группа, состоящая из группы ФИО и согласованной с ней группы «должность+организация».

[известный предприниматель В. В. Пупкин], [директор по развитию компании «Лингвистические супертехнологии»], ...)

Выраженное таким образом отношение обычно соответствует тематическому компоненту актуального членения предложения.

2) Предикат, стоящий в вершине клаузы, с актантами, соответствующими группам ФИО, Должность и Организация.

[Известный предприниматель В. В. Пупкин], занимавший ранее должность [директора по развитию компании «Лингвистические супертехнологии»], назначен несколько дней назад [исполняющим обязанности директора ЗАО «Лингвистические суперпроцессоры»]

Выделение атомарных объектов

После традиционных графематического и морфологического модулей (www.aot.ru), следует этап выделения неразрывных атомарных объектов. В настоящее время в системе распознаются три типа объектов: ФИО, Дата, Число. Это простые и часто встречающееся в тексте объекты, имеющие понятную всем структуру. Благодаря этому, помимо использования этих объектов при извлечении фактической информации, их возможно лемматизировать, поместить в индекс ИПС «Артефакт» и осуществлять по ним поиск, пользуясь специальными операторами в языке запросов. Для каждого типа объектов существует нетерминал, по которому будет найдено любое обозначения любого объекта данного типа. Так, по запросу

(Александр Пушкин /фио) родился @дата /n

будут найдены все предложения, в которых есть упоминание Пушкина, форма слова «родиться» и обозначение какой-нибудь даты.

ФИО – самый сложный из этих трех объектов. При выделении цепочек типа ФИО используется хранящаяся в морфологическом словаре информация об именах и отчествах. Наличие фамилии в словаре необязательно – различные эвристики позволяют предсказать, что данное слово в этом контексте является фамилией. При этом проверяются все согласования между именем и отчеством и именем и фамилией. Важным этапом здесь является отождествление формально различных ФИО в одном документе: исходя из предположения, что в одном документе (статье) вряд ли могут встретиться два ФИО с одними и теми же именем и фамилией, обозначающими разных людей, мы сливаем неполные ФИО с более полными. Например, упоминания «А. В. Петров», «Александр Петров» и «Александр Владимирович Петров» будут считаться относящимися к одному человеку.

В качестве чисел выделяются цепочки, записанные не только цифрами, но и словами. Так, запросу «*2575100!ч*» релевантны такие фрагменты:

мобилизовано 2.575 млн. руб. платежей;

объем торговой сессии составил $2575 тыс.;

США предоставит Грузии еще 2 млн 575 тыс. долл.

В качестве дат распознаются как абсолютные даты («в мае-июне 2001 года»), так и относительные («две недели назад»). Относительные даты вычисляются исходя из

абсолютной даты документа (в системе каждому документу приписана дата его публикации).

Из таких простых многословных объектов и одиночных слов на следующем этапе строятся неразрывные группы, выражающие отношение Человек-Организация. На специально для этого разработанном языке структура таких групп описывается правилами бесконтекстной грамматики.

Язык описания грамматики

Язык описания использует множество терминалов, соответствующих названиям частей речи, которые приписываются словарным единицам на этапе морфологического анализа, а также позволяет задавать терминалы, содержащие синтаксические единицы, построенные другими грамматиками, или содержащие группы слов/словосочетаний непосредственно хранящихся в словаре ключевых слов. Разработанный язык позволяет хранить информацию о согласовании между синтаксическими единицами в правой части правила и дополнительные грамматические ограничения для каждого терминала/нетерминала в пределах правила. Форма описания дает возможность определять синтаксическую вершину группы для каждого правила грамматики и эксплицитно приписывать граммемы построенной группе. Заданные правила грамматики преобразуются в LR(0)-таблицу, которая используется синтаксическим анализатором для обработки входной цепочки терминалов, приписанных каждой словоформе исходного предложения. Анализатор, основанный на алгоритме Томиты [2], интегрирован в систему из разработки немецкого синтаксиса Берлинской АН (www.bbaw.de). Выбор алгоритма Томиты обусловлен возможностью, в отличие от стандартных LR-парсеров, эффективно работать с морфологической и синтаксической омонимией внутри предложения: построение вариантов синтаксического представления исходной цепочки терминалов и последующий выбор дерева с наибольшим покрытием.

Выделение неразрывных цепочек ФИО-Должность-Организация

Основной принцип, используемый при выделении цепочек, состоит в следующем: в предложении выделяются ключевые слова (или словосочетания), указывающие на то, что в данном месте может встретиться группа должности или компании, затем вокруг этих слов с помощью грамматик строятся определенные именные группы, в которых вершинами являются найденные слова. (Например, «управляющий директор», «зампредправления» - ключевые слова для группы должности; «холдинг», «отдел» - ключевые слова для группы компании.) Эти же ключевые слова используются при распределении построенных групп по предопределенным ролям: ФИО, должность, имя компании, описание компании, география.

Во второй секции был заслушан доклад заместителя гендиректора по науке и информационным технологиям научно-производственной фирмы «Медиалинг» (Н. Новгород) кандидата технических наук Александра Ярмакова.

Должность	**Описание организации**	**Имя организации**	**География**	**Звание**	**ФИО**
Заместитель гендиректора по науке и информационным технологиям	Научно-производственная фирма	Медиалинг	Нижний Новгород	Кандидат технических наук	Александр Ярмаков

Вспомогательный словарь ключевых слов обладает возможностью объявлять цепочку слов, распознанную некоторой грамматикой, «ключевым словом» для других грамматик, имеющим ту же информацию, что и обычный член предложения. Например, цепочки

слов, распознанные грамматикой географии или временных групп, могут стать терминалами грамматики для цепочек Человек-Организация.

Распознавание предикативного выражения ФИО-Должность-Организация

Для анализа второго способа выражения этого отношения – с помощью предикативной вершины и актантов – требуется использование еще одного общелингвистического процессора, фрагментации. Модуль фрагментационного анализа разбивает сложное предложение на простые, выделяет причастные и деепричастные обороты, подчиненные предложения и т.д. [3] Вершина каждой клаузы ищется в словаре ключевых слов (например, «занимать», «назначить»). Таким словам приписаны словарные статьи, описывающие грамматические признаки, лексический состав и взаимное расположение актантов для ситуаций назначения, отставки, нахождения в должности и т.д. При проверке синтаксических связей между вершиной и актантами используется информация о связях, полученная на этапе фрагментации: информация о найденном подлежащем, об определяемом слове причастного оборота и т.д. Структура актанта описывается некоторой грамматикой, так что актанты обычно являются многословными и про них известно больше, чем просто часть речи и граммемы. Такое детальное описание актантов позволяет, не строя полного синтаксического дерева клаузы, распознавать довольно сложные построения.

Интерпретация

Цель интерпретации состоит в сопоставлении множества строящихся нетерминалов (фразовых категорий) грамматики множеству выделяемых объектов для определенного типа отношения (например, Человек-Должность), т.е. распределение подцепочек слов по заранее заданным полям таблицы. Интерпретация осуществляется на процедуре Reduce (свертка правой части правила). Основная трудность, решаемая на этапе интерпретации, - анализ сочинения.

Пример анализа сочиненных членов двух групп, связанных отношением Человек-Должность-Организация:

Вагит Алекперов, Михаил Ходорковский, Евгений Швидлер и Герман Хан, президенты НК ЛУКОЙЛ, НК ОАО ЮКОС и НК "Сибнефть" и директор Тюменской нефтяной компании, подпишут меморандум о взаимодействии по строительству нефтепровода в Мурманский порт.

Должность	**Описание организации**	**Имя организации**	**ФИО**
Президент	Нк	Лукойл	Алекперов Вагит
Президент	Нк ОАО	Юкос	Ходорковский Михаил
Президент	Нк	Сибнефть	Швидлер Евгений
Директор		Тюменская нефтяная компания	Хан Герман

На этом примере видно, что интерпретация сочинения реализуется через последовательный анализ двух связей:

1. «один ко многим»: группы Должности с вершиной во множественном числе [*президенты*] и группы сочиненных Организаций [*НК ЛУКОЙЛ, НК ОАО ЮКОС и НК "Сибнефть"*];

2. «многие ко многим»: группы сочиненных ФИО [*Вагит Алекперов, Михаил Ходорковский, Евгений Швидлер и Герман Хан*] и группы сочиненных Должностей [*президенты*[…] *и директор* […]].

Нормализация

Работу нормализации для построенного с помощью актантной структуры предиката 'назначить' отношения Человек-Должность-Организация можно продемонстрировать примером:

Сергей Лавров несколько дней назад был назначен новым министром иностранных дел Российской Федерации.

Должность	Имя организации	География	ФИО
новый министр	Министерство иностранных дел России	Россия	Лавров Сергей

Нормализация выделенных объектов внутри отношения реализует два принципа:

1) морфологическая нормализация: когда выбирается синтаксическая вершина именной группы и выставляется в форму именительного падежа, а вслед за этим и все ее премодификаторы, синтаксически непосредственно связанные с вершиной; например, вершина группы [*новым министром*] 'министром' -> 'министр' -> 'новый министр';
2) нормализация через тезаурус и географический словарь: восстановление с помощью тезауруса части названия организации, вложенной в название занимаемой должности, 'министр' -> 'министерство'; лемматизация географического названия через словарь географии, 'Российской Федерации' -> 'Россия'.

Кореференция. Восстановление эллипсиса

Внутри выделенных цепочек для отношения Человек-Должность-Организация на месте ФИО может стоять местоимение, для которого осуществляется поиск антецедента слева по тексту документа в пределах 3-4 предложений. Если в ходе анализа остается несколько кандидатов на роль антецедента, то выбирается ближе стоящий к анафоре в тексте. В общем виде алгоритм приближается к модели установления кореференциальных связей, описанной в работах S.Lappin и H.Leass [4]. Похожий механизм восстанавливает эллипсис имени организации в исходной цепочке. Специальная грамматика занимается обнаружением эллипсиса имени организации, используя результаты работы предшествующих грамматик, осуществляющих выделение организаций и их дескрипторов, а также их отождествление в пределах текста документа.

Пример:

НП "Уралалмаз" и ООО "Кама-кристалл" создали совместное предприятие – ЗАО "Уралалмаз". Оно было зарегистрировано 20 августа. Как пояснил генеральный директор ЗАО Геннадий Галкин, цель создания новой структуры – проведение геолого-разведочных работ.

Результат анализа:

Должность	Описание организации	Имя организации	ФИО
генеральный директор	ЗАО	"Уралалмаз"	Геннадий Галкин

Модуль поиска. Отождествление фактов

База фактов имеет пользовательский интерфейс, который позволяет проводить два основных вида поисков: по человеку и по организации. Задавая фамилию и имя или инициалы, пользователь получает все подходящие ФИО, для каждого из которых выдается множество организаций, связанных с ФИО отношением «занимать должность». При поиске по названию организации, выдается список людей, занимающих в ней какую-либо должность.

Одной из основных задач, ставившихся перед системой и не решаемой при полнотекстовом поиске, является определение тождественности информации, выраженной формально различными способами (например, с использованием разного лексического состава) в различных документах. Определение тождества объектов и фактов происходит на двух этапах. При парсировании документа отождествляются варианты обозначений одного ФИО и краткие и полные названия одной организации (окказиональные сокращения). Однако основная работа происходит в реляционной базе, где хранятся извлеченные объекты. Основной принцип определения тождества двух связей звучит так: две связи Человек-Организация считаются тождественными, если обозначения людей могут являться вариантами одного ФИО и обозначения организаций могут относиться к одной организации. Так, будут «слиты» факты, полученные из двух фрагментов:

Глава МЭРТа Г.Греф

Греф Герман Оскарович – глава Министерства экономического развития и торговли России

Для этого достаточно, чтобы существовал хотя бы один документ, из которого можно извлечь информацию, о том что МЭРТ – сокращение от Министерства экономического развития и торговли, например, в таком контексте:

В Министерстве экономического развития и торговли России (МЭРТ) 25 марта в рамках процесса доработки Лесного кодекса ...

Достижения. Проблемы. Дальнейшее развитие

Мы считаем, что положительный результат достигнут, в частности, из-за выбора решения применять синтаксический анализатор локально, вокруг «опорных» слов, а не выделять цепочки, содержащие искомые факты, на результатах синтаксического анализа всего документа. Однако именно невозможность в настоящей версии построения всех синтаксических вариантов разбора предложения и выбора из них наиболее правдоподобного является одной из причин ошибок – выбор неправильного синтаксического варианта.

[Адвоката председателя правления ЮКОСа] [Михаила Ходорковского]] или

[Адвоката [председателя правления ЮКОСа Михаила Ходорковского]]

Одним из частотных случаев ошибок являются сочиненые конструкции, в которых не всегда можно определить, сочиняются ли должности, относящиеся к одному человеку, или речь идет о нескольких людях.

В частности, членами клуба является семья [Юрия Лужкова, посла Японии,] префекта Центрального округа Москвы Александра Музыкантского, etc.

Еще одной частой проблемой, специфичной для данной задачи, является определение правой границы многословного названия организации

Андрей Орлов и Николай Петров назначены заместителями председателя [Государственного комитета РФ по поддержке и развитию малого предпринимательства] решением председателя правительства.

Основным результатом нашей работы по созданию полностью автоматизированной системы извлечения фактической информации из текстового массива является реально работающий вопросно-ответный модуль, базирующийся на большом (около 500 Гб), ежедневно пополняющемся текстовом массиве. Тот факт, что оказалось возможным создание вопросно-ответной системы, оперирующей ограниченным набором информативных объектов и отношений, дает импульс к дальнейшему развитию этого подхода как экстенсивно, увеличивая количество извлекаемых объектов, отношений и событий, так и интенсивно, совершенствуя собственно лингвистические модули: контекстное разрешение морфологической, лексической и синтаксической неоднозначности, расширение области действия механизмов восстановления элиппсиса, референтных связей.

Список литературы

1. Ralph Grishman. Information extraction: Techniques and challenges.In *Information Extraction (International Summer School SCIE-97)*. Springer-Verlag, 1997.

2. Elisabeth Scott, Andrian Johnstone, and Shamsa Sadaf, Hussain: Tomita-Style Generalised LR Parsers

3. Л. М. Гершензон, Д. В. Панкратов. Фрагментационный нализ русского предложения в системе Artefact. // Компьютерная лингвистика и интеллектуальные технологии: труды Международного семинара Диалог'2002. В двух томах. Т.2. "Прикладные проблемы". – Москва, Наука, 2002.

4. S.Lappin, H.J. Leass. An algorithm for pronominal anaphora. Computational Linguistics, 20(4), pp. 535-561, 1994