Эксперименты по извлечению информации из аналитических текстов финансовых обзоров

Е.И. Большакова, Ю.А. Жеребцова

Национальный исследовательский университет «Высшая школа экономики»
eibolshakova@gmail.com, julia.zherebtsova@gmail.com

Аннотация

В статье кратко описывается система автоматической обработки текста на основе лексико-синтаксических шаблонов, которая разработана для извлечения событийной информации из неструктурированных русскоязычных текстов аналитических финансовых обзоров, размещенных в сети Интернет. Приводятся и обсуждаются результаты экспериментальной оценки реализованного метода извлечения информации.

Введение

На сегодняшний день значительная часть информации для человека доступна только в виде неструктурированных текстов на естественном языке. Постоянно растущий объем неструктурированных текстовых данных в сети Интернет, находящихся в свободном доступе, значительно затрудняет процесс поиска необходимой информации, а также отделение значимой информации от незначимой. Попытки справиться с этой проблемой дали дополнительный импульс развитию научной области под названием компьютерная лингвистика [5].

Одной из актуальных и сложных задач компьютерной лингвистики является задача извлечения информации [10] (Information Extraction, IE) – выявление в текстах на естественном языке объектов заданной предметной области и их связей, построение их формализованного представления, например, в виде записей реляционной базы данных.

В рамках задачи извлечения информации выделяют следующие направления [9]:

- распознавание именованных сущностей (имен персоналий, географических названий, названий организаций, дат и т.п.);
- выделение семантических отношений распознанных сущностей (например, отношение «работать в» для выявленных персоналий и названий организаций);
- извлечение информации о заданных событиях и их атрибутах (например, событие «кораблекрушение» с атрибутами «дата», «время», «место» и др.).

Каждое из указанных направлений являлось предметом исследования серии международных конференций MUC (Message Understanding Conferences) [11].

Большинство современных IE-систем по основному методу извлечения информации делится на два вида: системы, основанные на представленных в виде правил знаниях (Knowledge Based, или Rule Based), и системы, основанные на машинном обучении (Machine Learning). Системы первого типа имеют, как правило, высокую точность извлечения (Precision) и довольно низкий показатель полноты (Recall), в то время как системы, использующие методы машинного обучения, наоборот, извлекают информацию с меньшей точностью, но с большей полнотой.

На сегодняшний день наиболее используемым программным инструментом для разработки программных систем извлечения информации, основанных на правилах, является инструментальная система GATE [6]. Для построения систем автоматической обработки текстов на русском языке реализованы системы RCO Pattern Extractor [3] и LSPL [2], встроенные средства которых упрощают разработку.

В настоящей работе рассматривается задача автоматического извлечения событийной информации из русскоязычных текстов на базе методов, основанных на знаниях. Объектом обработки выступают тексты ежедневных финансовых обзоров, выпускаемых аналитическими департаментами инвестиционных компаний и публикуемых в сети Интернет. Каждый из обрабатываемых текстов содержит упоминание о выпуске некоторой компанией финансовой отчетности за определенный временной период, что представляет собой извлекаемое из текста событие.

В качестве инструмента извлечения был выбран язык LSPL и его программные средства [2]. Язык LSPL предназначен для формального описания конструкций русского языка с целью их представления в системах автоматической обработки текстов, основанных на частичном синтаксическом анализе. Ключевым в языке LSPL является понятие лексико-синтаксического шаблона, рассматриваемого как структурный образец языковой конструкции, который задает ее лингвистические свойства: лексический состав и поверхностно-синтаксические связи.

Труды XV Всероссийской объединенной конференции «Интернет и современное общество» (IMS-2012), Санкт-Петербург, Россия, 2012.

В данной статье описывается IE-система, построенная на основе лексико-синтаксических шаблонов, и результаты экспериментальной оценки реализуемого ею метода извлечения событийной информации из аналитических текстов финансовых обзоров. При построении системы была собрана коллекция текстов интернет-обзоров, в результате анализа которой построен набор LSPL-шаблонов. Этот набор лег в основу применяемого метода извлечения.

В статье последовательно рассматриваются лингвистические особенности обрабатываемых текстов коллекции, принципы построения LSPL-шаблонов, ключевые идеи метода извлечения, а также полученные в результате экспериментов оценки эффективности метода – точность, полнота и F-мера извлеченной событийной информации.

Лингвистические особенности текстов

Поскольку обрабатываемые тексты относятся к узкой предметной области аналитичеких обзоров финансовой отчетности, а событие выпуска отчетности представляется ограниченным множеством языковых конструкций, это дает возможность довольно полно описать эти конструкции набором LSPL-шаблонов.

Коллекция из 38 текстов аналитических обзоров, на основе которой составлялся набор LSPL-шаблонов, бралась из ежедневных финансовых обзоров Банка Москвы за период 2010 – 2011 гг. Приведем в качестве примера фрагмент одного из текстов коллекции [13]:

Вчера Автоваз подвел финансовые итоги за 3-й квартал 2010 года. Выручка компании выросла на 57 %, а себестоимость – на 40 %, в результате чего маржа по валовой прибыли составила приличные 12.2 %...

Комментарий. Своими позитивными финансовыми результатами Автоваз в значительной степени обязан государственной программе утилизации. Поскольку большая её часть пришлась на II полугодие 2010, его результаты выглядят более сильными, чем результаты первого… Вышедшая отчетность в целом ожидаема нами и, на наш взгляд, не окажет существенного влияния на котировки Автоваза…

В качестве извлекаемых атрибутов рассматриваемого события выпуска отчетности были выбраны следующие четыре (в приведенном фрагменте обзора они подчеркнуты):

Название компании, опубликовавшей отчетность – целиком или его аббревиатура (в приведенном фрагменте – *Автоваз*).

Период, за который представлены финансовые результаты (в приведенном фрагменте – *3-й квартал 2010 года*). Обычно они записываются в следующем виде:

<№_квартала> *квартал*
<№_полугодия> *полугодие*
<год> *год*

Изменение выручки компании (в приведенном фрагменте – *выросла на 57%*). В зависимости от характера изменения записывается в одной из форм:

\+ <кол-во_процентов> *%*
\- <кол-во_процентов> *%*
увеличилась в <кол-во_раз> *раз(а)*
уменьшилась в <кол-во_раз> *раз(а)*

Качество отчетности относительно ожиданий финансового аналитика (в приведенном фрагменте – *отчетность в целом ожидаема нами*). Чаще всего записывается в виде одной из трех конструкций:

в рамках ожиданий
лучше ожиданий
хуже ожиданий

Каждый из рассмотренных атрибутов события представляет один из четырех видов извлекаемой информации: именованную сущность, дату (как отдельный вид именованной сущности), количественный показатель и оценочную информацию (мнение). Выбор разнотипных атрибутов позволяет нам дополнительно оценить выразительную мощность языка LSPL.

Первым из извлекаемых атрибутов является название компании, выпустившей финансовую отчетность. Это имя собственное, как правило, употребляющееся без кавычек, иногда содержащее цифры и знак дефис («-») (например, Лукойл, Газпром нефть или ОГК-5). Типичными конструкциями, описывающими контекст употребления названия компании, являются простые предложения типа

Газпром представил финансовую отчетность;
ОГК-5 опубликовал отчетность;
вчера Лукойл обнародовал финансовые результаты

Основной сложностью извлечения названия компании является то, что оно может состоять из нескольких слов, количество которых четко не определено.

Финансовая отчетность, выпущенная по международным стандартам (МСФО) [1], отражает деятельность компании за определенный период времени, поэтому в качестве второго извлекаемого атрибута был взят отчетный период. Чаще всего компании отчитываются каждый квартал или полугодие, в связи с этим типичными языковыми конструкциями, описывающими отчетный период, являются фразы вида:

– *результаты за 1-й квартал 2011 г.;*
– *отчетность за 2-е полугодие;*
– *финансовые итоги за 2010 год.*

Отметим, что в текстах коллекции, как правило, редко встречается конструкция 4-й квартал 2010 г., вместо этого употребляется просто 2010 г., что есть одно и то же. Сложность извлечения отчетного периода заключается в большом количестве чередований чисел, слов и дополнительных знаков, таких как точка («.») и дефис («-»). Также зачастую в тексте упоминаются другие периоды и даты, например, даты выпуска предыдущей отчетности или периода прогнозов. Указанная особенность составляет

сложность извлечения информации о текущем отчетном периоде.

Третьим извлекаемым атрибутом является динамика выручки, которой будем понимать ее изменение относительно предыдущего периода выпуска отчетности или по сравнению с аналогичным периодом прошлого года, полугодия или квартала. Наиболее распространенными конструкциями, описывающими динамику выручки, являются фразы вида:

выручка компании увеличилась в 2010 г. по сравнению с 2009 г. на 58.4 %;

выручка сократилась на 6 %;

выручка НОВАТЭКа превзошла аналогичный показатель прошлого года (+38.4 %).

Основная трудность извлечения динамики выручки заключается в удаленности в рассматриваемом тексте названия показателя от величины его изменения. Например, во фразе:

выручка Газпрома в 4-м квартале 2010 года практически совпала с нашим прогнозом и превзошла аналогичный показатель 3-го квартала (+15.3%)

слово *выручка* и ее изменение *+15.3%* употребляются далеко друг от друга, разделяясь довольно разнообразными конструкциями.

Последним извлекаемым атрибутом является мнение аналитика о представленной финансовой отчетности относительно его ожиданий, либо относительно совокупного консенсус-прогноза рынка, учитывающего мнение нескольких аналитиков из других информационных агентств.

Основными языковыми конструкциями, описывающими качество отчетности, являются такие предложения, как:

финансовые показатели Газпрома оказались лучше наших ожиданий;

отчетность Уралкалия оказалась лишь немногим хуже прогнозов;

НОВАТЭК представил финансовые результаты, которые оказались на уровне ожиданий.

По результатам проведенного анализа текстов коллекции был вручную определен набор типичных языковых конструкций для каждого извлекаемого атрибута события.

Метод извлечения

На основе выявленных типичных языковых конструкций для каждого извлекаемого атрибута на языке LSPL был составлен первоначальный набор шаблонов, который впоследствии корректировался и дополнялся. Рассмотрим методику составления набора шаблонов на примере одного из атрибутов – названия компании, представившей финансовую отчетность.

Первоначально для извлечения названия компании были составлены шаблоны вида

NAME = N1<c=nom> {NAME1}<0,2>

COMPANY = NAME1 V1<опубликовать, t=past>

COMPANY = NAME2 V2<обнародовать, t=past>

Данный набор шаблонов учитывает случаи, когда название компании состоит из нескольких слов, при этом первый шаблон NAME описывает название компании – последовательность из одного, двух или трех существительных в именительном падеже, а шаблон COMPANY выражает контекст употребления названия компании. При помощи COMPANY распознают конструкции, где после названия компании идет один глагол в прошедшем времени, выражающий факт публикации отчетности, например, «*Газпром нефть опубликовал*».

Однако после тестирования этих первоначальных шаблонов с помощью программной системы LSPL на рассматриваемой коллекции текстов, не были выделены довольно регулярно встречающиеся фразы, такие как *Газпром нефть вчера вечером раскрыл финансовые результаты и Лукойл подвел итоги финансового года* и т.п. Набор шаблонов был дополнен с учетом вновь выявленных языковых конструкций (упоминание времени публикации отчетности – *вчера вечером*) и контекстных синонимов (*подвел итоги, раскрыл финансовые результаты* и т.д.), в частности:

COMPANY =NAME1 {Av1}<0,2> V1<опубликовать, t=past>

COMPANY =NAME3 {Av3}<0,2>V3<подвести, t=past>

Затем расширенный набор шаблонов был протестирован на прежней коллекции текстов и дополнен снова. Подобная итеративная процедура расширения множества шаблонов позволила составить набор из восьми расширенных шаблонов, наиболее полно описывающий конструкции употребления имен отчитавшихся компаний в коллекции из 38 текстов.

Аналогичная процедура использовалась при составлении LSPL-шаблонов для остальных атрибутов события. В результате для описания отчетного периода было составлено 27 шаблонов, для изменения выручки – 36, а для качества отчетности – 32.

Извлечение информации на базе шаблонов в разработанной нами IE-системе происходит в несколько этапов:

- выделение из входных текстов необходимых языковых конструкций, соответствующих составленным шаблонам, с помощью программных средств LSPL;
- извлечение из выделенных конструкций требуемой информации об атрибутах рассматриваемого события;
- формирование из извлеченных атрибутов записи (строки) и занесение в базу данных.

На первом этапе программной системой LSPL в каждом из входных текстов на базе составленных шаблонов поочередно выделяются фразы, содержащие название компании, затем фразы, упоминающие отчетный период, далее – изменение выручки и качество отчетности относительно ожиданий.

Таким образом, для рассмотренного в предыдущем разделе фрагмента текста на первом этапе из-

влечения информации будут последовательно выделены четыре фразы:

Автоваз подвел финансовые итоги;

подвел финансовые итоги за 3-й квартал 2010 года;

Выручка компании выросла на 57%;

отчетность в целом ожидаема нами.

На втором этапе происходит обработка выделенных на первом этапе языковых конструкций и извлечение необходимой информации. Из приведенных выше конструкций для каждого из атрибутов будет извлечено:

Автоваз;

3-й квартал 2010 года;

выросла на 57%;

отчетность ожидаема.

На последнем этапе осуществляется построение строки базы данных и занесение ее в соответствующую таблицу – см. Табл. 1.

Таблица 1. Пример заполнения атрибутов события выпуска финансовой отчетности

№ события	Компания	Отчетный период	Изменение выручки	Качество отчетности
1	Автоваз	3 квартал 2010 год	+57%	в рамках ожиданий

Экспериментальная оценка

Оценка эффективности разработанной IE-системы проводилась на новой коллекции из 35 финансовых обзоров, где каждый текст содержал ровно одно событие. Качество извлечения информации оценивалось с помощью классических метрик точности, полноты и F-меры [7, 8] – как для каждого из атрибутов события в отдельности, так и для всей системы в целом.

Числовые характеристики эффективности извлечения каждого из атрибутов события по отдельности, а также всего события целиком (последний столбец), приведены в Табл. 2. F-мера во всех случаях вычислялась по стандартной формуле

$$F = \frac{2*R*P}{(R+P)}$$

где R – полнота, а P – точность.

Таблица 2. Точность, полнота и F-мера извлечения атрибутов события и событий в целом (%)

	Атрибуты события				Событие в целом
	Название компании	Отчетный период	Изменение выручки	Кач-во отчетности	
Точность	96	94	92	93	93
Полнота	77	86	63	78	78
F-мера	86	90	75	85	85

Показатель точности извлечения каждого атрибута превышает 90%, что является ожидаемо высоким результатом для системы извлечения информации, основанной на знаниях. Показатель полноты в среднем также довольно высок, он превышает 75%. Существенно ниже, чем все остальные, оказался показатель полноты для изменения выручки. Данный атрибут выражается в тексте довольно широким набором сложных языковых конструкций, что потребовало составить для него LSPL-шаблонов больше, чем для любого другого атрибута, и в то же время даже такой набор покрывает далеко не все возможные случаи, что говорит о необходимости расширения его в будущем.

В целом F-мера извлечения для каждого из атрибутов не опускается ниже 75%. Если рассмотреть задачу извлечения названия компании и отчетного периода как отдельную задачу распознавания именованных сущностей, то F-мера извлеченной информации является в среднем близкой к 90%, что немногим хуже лучших систем, основанных на знаниях, представленных на MUC-7 (F-мера = 93%).

Для общей оценки эффективности применяемого метода извлечения событийной информации использовался следующий способ, примененный на конференциях MUC-5 и MUC-6 [12].

Для каждой строки в заполненной базе данных проверяется, все ли атрибуты заполнены верно. Если это так, то строка считается *корректно заполненной*. При этом если какого-то для атрибута значение не найдено, например, изменение выручки, и в тексте про изменение выручки ничего не говорится, то в этом случае ошибка не фиксируется, т.к. атрибут найти нельзя. Таким образом, строка является корректно заполненной, если корректно заполнены все атрибуты, которые представляется возможным заполнить.

Заполненный атрибут считается *частично корректным*, если извлеклась не вся информация, к нему относящаяся, или же извлечена лишняя. Например, в качестве значения отчетного периода найден только «4 квартал», в то время как в тексте есть «4 квартал 2010 года». Строка, имеющая кроме корректных еще и частично корректные атрибуты, является *частично корректной*.

Таким образом, точность P, полнота R и F-мера извлеченной информации рассчитывались по следующим формулам [7]:

$$P = \frac{correct + 0.5 * partial}{actual}$$

$$R = \frac{correct + 0.5 * patrial}{possible}$$

$$F = \frac{2*P*R}{(P+R)}$$

где *correct* – количество корректных строк базы;

partial – количество частично корректных строк;

actual – количество заполненных строк, имеющих пропуски только тех значений атрибутов, которые отсутствуют в тексте;

possible – количество строк, которые можно извлечь из текстов.

Результаты вычисления точности, полноты и F-меры извлечения события в целом приведены в Таблице 2.

Согласно примененному способу подсчета эффективности извлечения информации, разработанная система имеет очень высокий показатель точности (93%) и довольно высокий показатель полноты (56%). Для сравнения, IE-системы, представленные на MUC-5 в рамках решения задачи извлечения экономических событий из новостных статей the Financial Times, имели показатели точности и полноты извлечения около 60% и 43% соответственно [8]. Максимальная величина показателя F-меры всех систем извлечения событийной информации, представленных на конференциях MUC, не превышала 60% [4]. Основной причиной столь высокой точность является тот факт, что каждый текст коллекции содержит информацию ровно об одной событии.

Рост показателя F-меры извлечения информации разработанной системой (+10%) во многом обусловлен той же самой особенностью текстов коллекции (каждый текст – одно событие), а также тем, что количество атрибутов извлекаемого события примерно в два раза меньше среднего количества атрибутов, извлекаемых IE-системами на конференциях MUC.

Заключение

На основе языка и программных средств LSPL построена программная IE-система, основанная на правилах (лексико-синтаксических шаблонах), и получена экспериментальная оценка эффективности метода извлечения.

Эксперименты показали, что при обработке текстов, содержащих информацию о событии выпуска финансовой отчетности, извлекается информация о 56 % всех событий, причем 93 % извлеченных данных корректны.

В целом полученные результаты позволяют утверждать, что разработанная система имеет приемлемое качество извлечения информации, сравнимое с аналогичными показателями современных IE-систем, а язык шаблонов LSPL и его программные средства являются мощным и гибким инструментом для построения систем извлечения информации из текстов на русском языке в узкой предметной области.

Литература

[1] Берзон Н.И. Фондовый рынок. М.: Вита-Пресс, 2009.

[2] Большакова Е.И., Носков А.А. Программные средства анализа текста на основе лексико-синтаксических шаблонов языка LSPL. М.: Изд. отдел факультета ВМиК МГУ имени М.В.Ломоносова, МАКС Пресс, 2010.

[3] Ермаков А.Е., Плешко В. В., Митюнин В. А. RCO Pattern Extractor: компонент выделения особых объектов в тексте // Информатизация и информационная безопасность правоохранительных органов: XI Международная научная конференция. - Москва, 2003.

[4] Лукашевич Н. В. Тезаурусы в задачах информационного поиска. – М.: МГУ, 2011.

[5] Bolshakov I.A., Gelbukh A., Computational Linguistics: Models, Resources, Applications. IPN–UNAM–FCE, 2008.

[6] Bontcheva K., Maynard D., Tablan V., Cunningham H. GATE: A Unicode-based infrastructure supporting multilingual information extraction // Proc. of Workshop on Information Extraction for Slavonic and Other Central and Eastern European Languages (IESL'03)., Borovets, 2003.

[7] Chinchor N., MUC-5 Evaluation Metrics // Fifth Messages Understanding Conference (MUC-5). Morgan Kaufman. 1993.

[8] Constantino M.: Financial Information Extraction using pre-defined and user-definable Templates in the LOLITA system. PhD Thesis at University of Durham. 1997.

[9] Grishman R. Information Extraction // The Oxford Handbook of Computational Linguistics / Mitkov R. (ed.). Oxford University Press. 2003. P. 545-559.

[10] Grishman R. Information Extraction: Techniques and Challenges. NY: Computer Science Department, 1997.

[11] Grishman R., Sundheim B. Message Understanding Conference - 6: A Brief History // Proc. of COLING. NY., 1996.

[12] Grishman R., Sterling J., Description of the PROTEUS System as used for MUC-5 // Fifth Messages Understanding Conference (MUC-5). Morgan Kaufman, 1993.

[13] Аналитический департамент Банка Москвы, обзор российского рынка акций за 09.01.2011 URL: http://bm-am.ru/ru/analitika/analitika_nedel/pdf/2011/file_2141.pdf.

Experiments on information extraction from analytical texts of financial dailies

E. Bolshakova, Yu. Zherebtsova

The paper describes in short a natural language processing system based on lexico-syntactic patterns and developed for automatic extraction of event information from unstructured Russian analytical texts of financial dailies published in Internet. The results of experimental evaluation of the applied extraction method is presented and discussed.