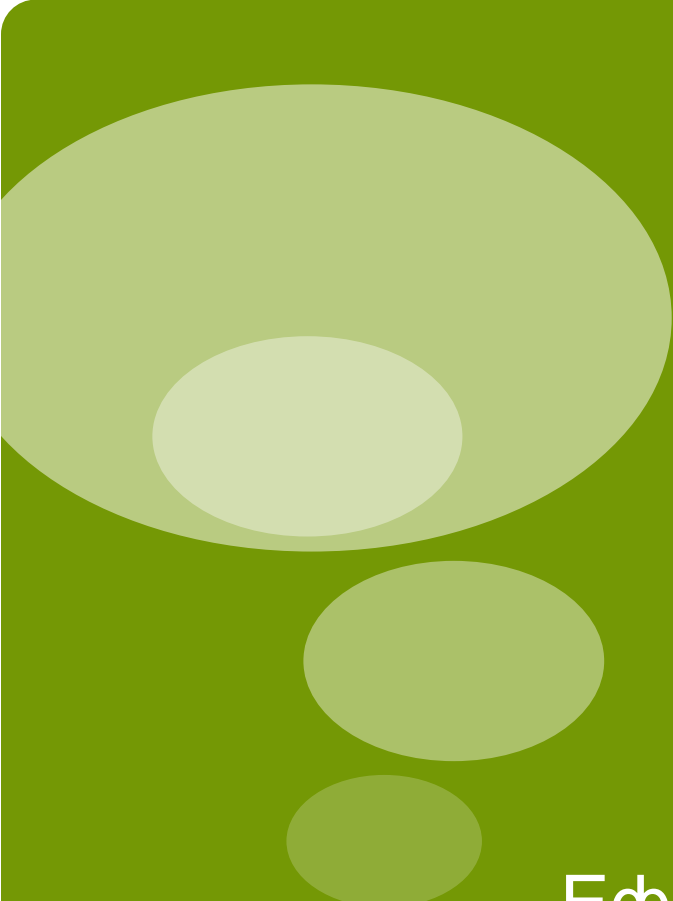




Извлечение информации

Ефремова Наталья Эрнестовна
Грацианова Татьяна Юрьевна

Содержание



- ❖ Постановка задачи
- ❖ Виды извлекаемой информации и особенности ее извлечения
- ❖ Подходы к извлечению информации
- ❖ Инструменты построения прикладных систем
- ❖ Оценка качества извлечения
- ❖ Результаты соревнований по извлечению информации

Извлечение информации (1)



- Сейчас (например, в Интернет) количество информации настолько огромно, что человек не в состоянии охватить ее за приемлемое время

Нужны программы извлечения и преобразования информации в форму, удобную для дальнейшей обработки

- Первые прикладные исследования – начало 1980-х, обработка новостных и военных текстов, выделения из них событий
- В настоящее время данные извлекаются также из художественных произведений, научно-технических статей, текстов сети Интернет и др.

Извлечение информации (2)



- Сейчас под извлечением информации понимают извлечение любых семантически значимых данных
- Ведутся исследования в области обработки мультимедийных документов: на основе содержимого аудио/видео файлов составляется их описание
- Приложения:
 - ✓ мониторинг новостных лент
 - ✓ составление дайджестов, рефератов, досье
 - ✓ сбор данных для анализа экономической, производственной и пр. деятельности

Постановка задачи



Извлечение информации (*Information Extraction, IE*): автоматическое извлечение данных из текста на ЕЯ

- обрабатывается отдельный текст или коллекция текстов, неструктурированные (без метаданных)
- извлекаются данные, релевантные определенной проблеме, вопросу, теме →

IE – разновидность информационного поиска

- извлеченные данные:
 - ✓ структурируются в виде таблиц, шаблонов
 - ✓ накапливаются в базах знаний
 - ✓ обрабатываются: сортируются, размечаются, визуализируются, сохраняются в базах данных

Виды извлекаемой информации



- ❑ **Значимые объекты для темы/области**
именованные сущности: *персоны, организации, герои произведений*
- ❑ **Отношения между объектами**
быть частью, быть владельцем
- ❑ **Атрибуты объектов:** для *персоны* – *место работы, должность, телефон, подразделение*
- ❑ **Факты/события:** *прошла встреча, выдан кредит, вступать в реакцию*

Виды извлекаемой информации: пример



Грейс Патриша Келли (12.11.1929 – 14.09.1982) – американская актриса, с 1956 года – супруга князя Монако Ренье III, 10-я княгиня Монако, мать ныне правящего князя Альбера II.

□ Объекты (сущности): ФИО, род занятий, даты

□ Отношения:

➤ **жена-муж:** Грейс Патриша Келли, Ренье III

➤ **родитель-ребенок:**

Грейс Патриша Келли, Альбер II

□ Факты и события:

*Альбер II – ныне правящий князь Монако
(Ренье III – отец Альбера II)*

Именованные сущности (*Named Entities, NE*)



- Изначально именованные сущности – это:
 - ❖ Имена персоналий: *И. Сечин, Ben White*
 - ❖ Географические названия: *р. Ока, гор. Москва*
 - ❖ Названия фирм/компаний/организаций: *ОАО «Я»*
 - ✓ Они имеют имя и референт – объект внешнего мира с данным именем

- Сейчас в зависимости от задачи еще выделяют:
 - ❖ Даты и временные отрезки: *02.03.1913, 2 р.т.*
 - ❖ Адреса: *3-ая улица Строителей д. 25, кв.12*
 - ❖ Марки товаров: *Nokia, Apple, Land Rover*
 - ❖ Ссылки на литературу: *[2], [Иванов, 1995]*
 - ❖ Гены, белки, хим. вещества: *$H_2N-CH(R)-COOH$*

Решение задачи извлечения *NE*



1. Нахождение наименований сущности в тексте
 2. Определение категории сущности
 3. Разрешение кореференции
 4. Связывание сущности с референтом (если сущность является именем собственным)
- ☐ Данная задача изучена лучше всего
 - ☐ Порядок пунктов 2-4 может быть другим

Нахождение наименований сущности



- Опора на соответствующие словарные ресурсы (имен, частей имен, географических названий, денежных единиц и т.п.)
- Учет особенностей именования:
 - ✓ регистр букв (первая или все большие)
 - ✓ определенные последовательности букв:
 - ов, -дзе – окончания фамилий
 - банк, -инвест – окончания названий компаний
 - ✓ внутренняя структура: +1(23)45-67
- Учет контекста: **ПАО** «Аэрофлот», ул. Полянка

Определение категории сущности



- ◆ Учет локального контекста – соседних слов

Скала «Три Сестры» расположена к востоку от Уральских гор

→ «Три сестры» – географический объект (на основе слова **скала**)

А.П. Чехов в 1900 году приступил к работе над пьесой «Три сестры»

→ «Три сестры» –художественное произведение (на основе слова **пьеса**)

- ◆ Учет глобального контекста – тематики и структуры текста, проверка употреблений по тексту/корпусу
Michigan State – часть названия университета
New York State – только лишь название штата

Разрешение кореференции



*Сбербанком поддержана акция «Красная звезда», которую по всей стране проводит благотворительный фонд «Память поколений». В крупных отделениях **банка** каждый клиент, совершивший любую операцию, получает значок в форме звезды*

Используются стандартные методы:

- ☐ выявления
- ☐ нормализации
- ☐ отождествления
- ☐ приведение к каноническому виду (ПАО «Сбербанк **России**»)

Связывание сущности с референтом



Лена — это не только женское имя, но и название:

- ✓ реки (не единственной)
 - ✓ населённого пункта (не единственного)
 - ✓ автомобильной дороги
 - ✓ железнодорожной станции
- Для связывания сущности с референтом используются внешние источники знаний о существующих в мире персонах и объектах
 - При использовании Википедии каждая ее страница рассматривается как отдельный референт

Связывание сущности с референтом: пример



Глава МИД РФ рассчитывает, что встреча президентов России и США 7-8 июля внесет ясность в перспективы отношений двух стран. Об этом Лавров заявил на международном форуме «Примаковские чтения». Я надеюсь, что возобладает прагматизм, — заключил министр.

Сложности извлечения *NE*



- Нельзя зафиксировать в словаре, например, все названия компаний
- Особенности именования не являются строгими и однозначными:
Роза не любит жару
- В качестве извлекаемых объектов могут выступать обычные слова и словосочетания текста
заместитель директора, научный сотрудник
- В зависимости от контекста сущность может относиться к разным видам (категориям):
В России прошли ... – географический объект
Россия отказалась от ... – страна
- Из контекста может быть не понятен референт:
Нам задали читать Толстого

Атрибуты объектов



- После извлечения именованных сущностей/объектов можно устанавливать связи между ними
- Для извлечения атрибутов определяют основную категорию сущности и категории ее атрибутов

*Продам 2-комнатную квартиру по ул. Молодежная.
Площадь 45,1 кв.м. Цена договорная.*

- Для объекта **квартира** могут быть извлечены следующие **атрибуты** (и их значения):
 - ✓ количество комнат: **2**
 - ✓ адрес: **ул. Молодежная**
 - ✓ общая площадь (кв.м.): **45,1**
 - ✓ цена: **договорная**
- Множество атрибутов может быть шире: наличие лифта, газа, балкона, метраж комнат, тип дома и т. д.

Отношения сущностей



- ❑ Виды отношений:
 - ✓ общие (*часть-целое, причина-следствие*)
 - ✓ зависящие от тематики, предметной области (*компания-владелец, вступить в реакцию*)
- ❑ Обычно рассматриваются отношения только между двумя объектами
- ❑ При извлечении учитываются типичные конструкции выражения отношений
- ❑ Отношения могут быть непостоянными, например *сыграть главную роль* постоянно для *Караченцева* и фильма *«Ловушка для одинокого мужчины»*, но временно для спектакля *«Юнона и Авось»*
 - нужно отслеживать изменения, учитывать временные рамки

Отношения сущностей: примеры контекстов



быть режиссёром связывает *имя режиссёра* (ИМЯ)
и *название фильма* (НАЗВАНИЕ):

- фильм НАЗВАНИЕ режиссёра ИМЯ
*Фильм «Хохлатый убис» режиссёра Ляна Цяо
получил Гран-при 39-го ММКФ*
- фильм НАЗВАНИЕ ИМЯ
*Специальный приз жюри получил фильм «Мешок
без дна» Рустама Хамдамова*
- фильм режиссёра ИМЯ НАЗВАНИЕ
*Ранее приз зрительских симпатий ММКФ
получил фильм режиссёра Владимира Котта
«Карп отморозенный»*

Факты и события (Events)



«Who did what to whom, when, where, through what methods (instruments), and why»

- ❑ Объекты, их атрибуты и отношения извлекаются из текстов. Из них формируются факты и события
- ❑ События описываются набором параметров и их значений (именованные сущности)
- ❑ Такой набор образует так называемый **семантический фрейм события**
- ❑ Сущности в событии связаны определенным набором отношений
- ❑ Для извлечения фактов и событий используется информация о типичных конструкциях их выражения

Семантический фрейм события: пример



Вчера, 1 апреля 2007 года, представители корпорации Пепелац Интернэшнл посетили офис компании Гравицап Продакшнз.

- ☐ Событие – деловой визит
- ☐ Параметры:
 - ✓ визитер
 - ✓ принимающая сторона
 - ✓ дата
- ☐ Фрейм

<i><u>Визитер</u></i>	Принимающая сторона	<i>Дата</i>
<i>Пепелац Интернэшнл</i>	<i>Гравицап Продакшнз</i>	<i>1 апреля 2007 года</i>

Особенности и сложности извлечения событий и фактов



1. В предложении важно найти слово, которое выражает суть события
 2. Потом ищутся участники, устанавливаются их роли
- Событие в тексте может выражаться по-разному
 - Минобороны РФ ответило британскому министру обороны*
 - В Минобороны РФ ответили на обвинения британского министра*
 - В Минобороны РФ назвали «маловразумительными» заявления британского министра о Сирии*
 - Важно помнить о словах, меняющих суть (*почти, не*)
 - Необходимо слияние частичных описаний, полученных из разных предложений

Извлечения событий и фактов: пример



Мировые СМИ обсуждают предстоящую встречу Дональда Трампа с Владимиром Путиным. Уже официально подтверждено, что она пройдет в кулуарах саммита «двадцатки» 7-8 июля.

Особенности задачи извлечения информации



- Постоянно возникают новые приложения задачи со своей спецификой обрабатываемых текстов и распознаваемых данных
- Объекты разнородны, правила их именования и извлечения различны
- Современные прикладные системы извлечения информации, как правило, ориентированы на обработку текстов в **узких предметных областях**

Подходы к извлечению



□ **Машинное обучение**

- ✓ опора на статистические (вероятностные) методы
- ✓ необходим размеченный вручную обучающий корпус

□ **Инженерный подход** (применение лингвистических правил и шаблонов)

- ✓ правила и шаблоны пишут эксперты
- ✓ для записи нужны специальные языки и поддерживающие их программные средства

□ **Комбинирование** машинного обучения и инженерного подхода

Машинное обучение с учителем



- По корпусу размеченных данных (обучающей выборке) строится модель (машинный классификатор)
- Затем модель применяется к новым, неразмеченным текстам
- Применяется в задаче извлечения *NE*
- Используемые методы (различаются способом учета признаков):
 - ✓ наивный байесовский классификатор
 - ✓ деревья решений
 - ✓ метод опорных векторов
 - ✓ логистическая регрессия
 - ✓ HMM и CRF
 - ✓ нейронные сети

Обучение с учителем.

Разметка данных



- Для извлекаемых данных (сущности, атрибуты) определяются:
 - ❖ лингвистические и структурные признаки
 - ❖ ближайший контекст
- Для текстов может быть проведен графематический и морфологический анализ, реже – синтаксический
- Для разметки категорий именованных сущностей используются специальные схемы

*[Владислав]B-PERS [Сурков]I-PERS
[встретится]O [с]O [президентом]O
[Абхазии]B-LOC [Раулем]B-PERS
[Хаджимба]I-PERS*

Обучение с учителем.

Признаки



- Данные после разметки преобразуются в наборы признаков для каждого токена. Рассматриваются:
 - ✓ собственно токен, его длина, его соседи
 - ✓ вид токена (слово, знак препинания и пр.)
 - ✓ является ли началом/концом предложения
 - ✓ тег токена, полученный при разметке
 - ✓ входит ли токен в определенный словарь и т.п.
- Для токенов-слов дополнительно учитывается:
 - ✓ способ написания токена
 - ✓ лемма, часть речи, значения морфопризнаков
 - ✓ состав слова (корни, суффиксы и окончания, типичные для ФИО, названий организаций и т.д.)

Обучение с учителем.

Извлечение *NE*



- Классическая задача классификации токенов на несколько классов
 - ❖ Если используется обучение без учителя – кластеризация по схожим контекстам употребления
- Для обучения и использования классификаторов применяются разные стратегии:
 - классификатор одновременно распознает сущности разных категорий
 - для каждой категории работают отдельные классификаторы, результаты их работы объединяются
- При учете локального контекста токена логичнее использовать HMM и CRF:
 - ✓ категории – скрытые состояния
 - ✓ токены – наблюдаемые

Частичное обучение с учителем



- Используется при распознавании отношений и фактов
- Метод distant supervision:
 - ✓ берется много примеров сущностей, связанных определенным отношением/фактом. Источник, например, Википедия
 - ✓ автоматически готовится обучающая выборка: пусть все предложения с сущностями, связанными отношением, являются положительными примерами
 - ✓ применяется метод обучения с учителем
- Признаки учитывают контекст вокруг сущностей:
 - ❖ леммы слов, стоящих между сущностями, части речи
 - ❖ слова и их часть речи слева и справа
 - ❖ синтаксический путь между сущностями и его длина
 - ❖ категории именованных сущностей

Инженерный подход



Опора на **лингвистические шаблоны** – формальное описание языковых конструкций, их лексического состава и грамматических свойств

- Виды шаблонов: лексические, лексико-синтаксические
- Элементы шаблонов:
 - ❖ словоформы, лексемы (с указанием характеристик)
 - ❖ грамматические образцы: именные и др. группы
- Для описания шаблонов созданы специальные языки (*JAPE*, *LSPL*, грамматика Томита-парсера) и поддерживающие их программные средства
- При использовании шаблонов достаточно частичного синтаксического анализа предложений

Лингвистические шаблоны: пример



- ✓ **Объекты и атрибуты**

$(^([0-9]+[.]+[0-9]+)|(0)\$)$ – 0, 1.2, 12.345

А N – прелестный кварк, солнечное сплетение

- ✓ **Отношения между объектами** (*Ng* – именная группа)

Ng1 «является частью» Ng2

Процессор является частью компьютера

- ✓ **Факты и события:** *Fact=(Loc, Ship)*

Loc «пошел ко дну» Ship

Ship «затонул» Loc

В Белом море пошел ко дну корабль ВМС Индии

Корабль ВМС Индии затонул в Белом море

Современные тенденции



- Учет предметной области, темы, стиля автора, ...
- Использование большого количества словарных ресурсов, в частности, больших внешних ресурсов знаний (Википедия, DBPedia, WordNet, графы знаний и др.);
- Машинное обучение:
 - поиск естественно размеченных данных
 - применение мета-алгоритмов обучения
 - использование сложной разметки
 - учет большого числа признаков разного вида
- Инженерный подход:
 - инструменты построения систем извлечения
 - автоматизация построения правил и шаблонов

Идея автоматического построения шаблонов



Используется итеративный подход (bootstrapping), относящийся к методам частичного обучения с учителем:

- ❖ Имеется множество пар (например, объект-отношение)
- ❖ В текстах находятся упоминания этих пар
- ❖ Анализируется контекст – слова слева и справа, их роль в предложении
- ❖ Формируются наиболее вероятные шаблоны
- ❖ Полученные шаблоны проверяются на других текстах и парах

Инструменты построения систем извлечения



- Разработка прикладных *IE*-систем является сложным и трудоемким процессом
- Существенная помощь – использование инструментальных систем
- Инструментальные системы, поддерживающие инженерный подход (*GATE*, Томита-парсер, *LSPL*):
 - ✓ имеют встроенный формальный язык для задания лингвистических правил и шаблонов
 - ✓ настройка систем на конкретную задачу *IE* – с помощью шаблонов и правил
- ◆ Инструментальные системы, поддерживающие машинное обучение позволяют использовать уже построенные программные модели, а также обучать новые (*OpenNLP*, *Stanford*)

Меры эффективности извлечения информации (1)



Результаты работы системы можно разделить

эксперт \ система	правильные (positive – P)	неправильные (negative – N)
правильные (P)	True P = TP	False P = FP
неправильные (N)	FN	TN

- Точность (*Precision*) – отношение найденных правильных к общему количеству найденных

$$P = TP / (TP + FP)$$

- Полнота (*Recall*) – отношение найденных правильных к общему количеству правильных

$$R = TP / (TP + FN)$$

Меры эффективности извлечения информации (2)



- F-мера – соотношение между P и R

$$F = \frac{(\beta^2 + 1) PR}{\beta^2 R + P},$$

где β – коэффициент относительной важности, обычно $\beta=1$

- Ошибка (*Error*) – отношение неправильно принятых решений к общему числу решений

$$E = (FP + FN) / (TP + FP + TN + FN)$$

- Аккуратность (*Accuracy*) – отношение правильно принятых решений к общему числу решений

$$A = (TP + TN) / (TP + FP + TN + FN)$$

Соревнования систем извлечения информации



- IE – одно из первых направлений, где стали проводиться открытые тестирования автоматических систем на одних и тех же задачах и данных
- *MUC (Message Understanding Conference)*, проводилась с 1987 по 1998 годы
 - 1995 (*MUC-6*) – служебные перемещения: назначения и отставки
 - 1998 (*MUC-7*) – запуски космических кораблей и ракет
- РОМИП (российский семинар по оценке методов информационного поиска)
 - 2004 – поиск событий, связанных с персоной
 - 2005 – выделение именованных сущностей и фактов заданных типов

MUC-6 и MUC-7. Оценки



Были выработаны специальные способы оценки качества извлечения событий и фактов

- ✓ если все ли атрибуты заполнены верно, факт **корректно извлечён**
- ✓ если извлечена не вся информация или же извлечена лишняя, параметр считается частично корректным
- ✓ извлечённый факт, имеющий частично корректные параметры, является **частично корректным**

MUC-6 и MUC-7.

Формула оценки



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

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

- *correct* – число корректно извлечённых фактов
- *partial* – количество частично корректных фактов
- *actual* – число всех выявленных фактов
- *possible* – число фактов, которые можно извлечь из текстов (размеченных экспертом)

MUC-6 и MUC-7. Результаты



- ◆ Примеры атрибутов в MUC-7: *запущенный аппарат, дата запуска, место запуска, тип задания (военный, гражданский)*
- ◆ Примеры результатов:
 - Извлечение именованных сущностей
 - Машинное обучение (HMM)
MUC-6: F=93% MUC-7: F=90,4%
 - Извлечение на основе правил:
MUC-6: F=96,4% MUC-7: F=93,7%
 - Извлечение событий и фактов
 - На основе правил: P=90%, R=20%

РОМИП 2004



- Дана персона (*Аристарх Евгеньевич Ливанов, российский актер театра и кино*). Необходимо найти все события связанные с ним. Всего 5052 персоны
- Система должна вернуть фрагмент текста и отнести его к заранее определенному типу событий
- Полностью релевантный ответ – есть описание и время события, ссылка на персону
- Частично релевантный ответ – нет части информации
- Планировалось 3 участника, ответ сдал 1 (RCO)
- Оценивалась только точность, значения колеблются от 0 до 100%. В среднем:

$$\begin{aligned} P_{\text{полностью релевантных}} &= 24\% \\ P_{\text{частично релевантных}} &= 69\% \end{aligned}$$

РОМИП 2005



- ❑ Новостная коллекции
- ❑ Две подзадачи выделения:
 - ✓ именованных сущностей (организация, персона, географический объект, другое)
 - ✓ фактов заданных типов (где работает человек, кто владеет организацией)
- ❑ 2 участника, у каждого по 2 прогона
- ❑ Оценивалась точность – от 95 до 99% и ошибка – 1,9% до 4,5%
- ❑ Лучшие значения – для фактов

FactRuEval 2016



- ❑ Новостная коллекции
- ❑ Три подзадачи выделения:
 - ✓ именованных сущностей (организация, персона, географический объект)
 - ✓ сущностей и их атрибутов (нужно заполнить поля, не должно быть дубликатов)
 - ✓ фактов из текстов (найм, сделка, владение, встреча)
- ❑ 13 участников, большинство использовали инженерный подход с элементами статистики
- ❑ Автоматическая (практически) система оценки результатов, в открытом доступе

FactRuEval 2016.

Правила оценки



Система оценки сравнивала тестовую разметку с разметкой текстов из Золотого стандарта и устанавливала между ними соответствие, причем:

- Соответствие устанавливалось только между объектами одного типа
- Каждый объект из тестов мог иметь только один соответствующий ему объект из Золотого стандарта
- Каждый объект в Золотом стандарте мог иметь для сущностей – только один соответствующий ему объект в тестах, для фактов – несколько
- Для каждой задачи – свои формулы

Лучшие значения – для 1 и 2 дорожки ($F1 = 93\%$), худшие – для фактов ($F1=66\%$)

Выводы



- Извлечение информации из текстов – развитое направление компьютерной лингвистики
- Существует множество методов и соответствующих инструментальных средств для построения различных прикладных систем
- Актуальность задач направления сохраняется
- Современными тенденциями развития данного направления являются:
 - ✓ учёт при извлечении нелокальных зависимостей текстовых единиц
 - ✓ проведение более глубокого синтаксического анализа и использование синтаксических признаков при машинном обучении
 - ✓ сдвиг фокуса от структуризации извлечённой информации к ее визуализации



Спасибо за внимание!