

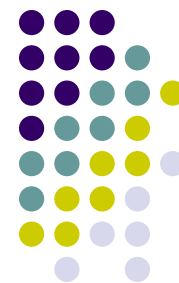


ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ ИЗ ТЕКСТОВ: ПОРТРЕТ НАПРАВЛЕНИЯ

Большакова Елена Игоревна

МГУ имени М.В.Ломоносова, фак-т ВМК

СОДЕРЖАНИЕ



1. Особенности задачи
Information Extraction (IE)
2. Виды извлекаемой информации и особенности ее извлечения
3. Подходы к извлечению информации
4. Инструментальные системы для
5. Оценка качества извлечения, соревнования методов и систем

АКТУАЛЬНОСТЬ ЗАДАЧИ



- Рост объема текстовой информации, особенно в сети Интернет: человек не в состоянии охватить ее за приемлемое время
- Нужны программы извлечения и преобразования информации в форму, удобную для дальнейшей обработки
- Возможные приложения:
 - ✓ мониторинг новостных лент
 - ✓ составление дайджестов, рефератов, досье
 - ✓ сбор данных для анализа экономической, производственной и др. деятельности

СПЕЦИФИКА ЗАДАЧИ



Information Extraction: автоматическое извлечение релевантных данных из текстов на ЕЯ

- Обрабатывается отдельный текст или коллекция текстов, неструктурированные (без метаданных)
- Извлекаются данные, релевантные определенной проблеме, вопросу, теме
- Важно: извлеченные данные
 - ✓ структурируются в виде таблиц, шаблонов
 - ✓ обрабатываются: сортируются, размечаются, отбираются, сохраняются в базах данных
 - ✓ накапливаются в базах знаний

ИЗВЛЕЧЕНИЕ ИНФОРМАЦИИ: ПРИМЕРЫ



- Извлечение информации о событиях, фактах (деловых визитах):

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

- Извлечение объектов, их атрибутов и отношений:

Грейс Патриша Келли – американская актриса, мать ныне правящего князя Альбера II.

- Извлечение новых терминов:

Такие слабовзаимодействующие массивные частицы называют вимпами.

ВИДЫ ИЗВЛЕКАЕМОЙ ИНФОРМАЦИИ



- *Именованные сущности (Named Entities, NE)* – значимые объекты: *персоны, названия фирм, белков, марки товаров, геогр. названия* и т.п.
- Атрибуты объектов: для *персоны* – *должность, место работы, телефон, подразделение*
- Отношения между объектами:
быть частью, быть владельцем
- Факты/события: *прошла встреча, выдан кредит*

А также:

- Термины ПО и их связи, ключевые слова текста
- Отзывы и мнения о товарах, услугах, кино и пр.

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 (имен.сущности): **ФИО**, род занятий +даты
- Отношения:
 - супружество: *Грейс Патриша Келли, Ренье III*
 - быть матерью: *Грейс Патриша Келли, Альбер II*
- Факты и события:
 - замужество (1956, *Грейс П. Келли, Ренье III*)
 - правящий князь (*Монако, Альбер II*)

Можно ли извлечь : *Ренье III – отец Альбера II ???*

ИМЕНОВАННЫЕ СУЩНОСТИ



Изначально именованные сущности – это:

- Имена персоналий: *И. Сечин, Ben White*
- Географические названия: *р. Ока, гор. Москва*
- Названия компаний/организаций: *РЖД, ОАО «Уют»*

Сейчас также выделяют:

- Даты и временные отрезки: *02.03.1913, 2 р.т.*
- Номера телефонов: *+7(123)456-78-90*
- Адреса: *3-ая улица Строителей д. 25, кв.12*
- Марки товаров: *Nokia, Apple, Land Rover*
- Обозначения денежных единиц: *руб., \$, GBP*
- Ссылки на литературу: *[2], [Иванов, 1995]*
- Гены, белки, хим. вещества: *$H_2N-CH(R)-COOH$*

СЛОЖНОСТИ ИЗВЛЕЧЕНИЯ *НЕ*



- Большое число разных сущностей/объектов, постоянно появляются новые
- Множество различных способов именования одной и той же сущности: *ВВП, В.В.Путин*
William H. Gates, Bill Gates, владелец Microsoft, BG
- Нередко требуется установление *кореллеренции* имен (тождества обозначаемых объектов – *референтов*)
ГАИ, ГИБДД – это один референт или разные?
- В зависимости от контекста имен. сущность может относиться к разным видам (категориям): *Лена, ВВП*
В России прошли ... – географический объект
Россия отказалась от ... – страна

ОСОБЕННОСТИ ИЗВЛЕЧЕНИЯ ИМЕНОВАННЫХ СУЩНОСТЕЙ



- Опора на соответствующие словари :
личных имен, географических названий и т.п.
- Учет особенностей наименований:
 - ✓ регистр букв (первая или все большие)
 - ✓ определенные последовательности букв
(*-ов*, *-дзе* — окончания фамилий,
-банк, *-инвест* — окончания названий компаний)
 - ✓ внутренняя структура: *ООО*, *+1(23)45-67*
- Учет контекста (м.б. проверка по всему тексту)
...станция Зима..., ...увидел Зиму...

ОТНОШЕНИЯ И АТТРИБУТЫ СУЩНОСТЕЙ



- Атрибуты конкретных объектов
квартира (продажа/покупка): *адрес, этаж, метраж, количество комнат, лифт, газ, ...*
- Отношения (связи) конкретных объектов
Виды отношений:
 - ❖ Общие: *часть-целое, причина-следствие*)
 - ❖ Зависящие от ПО текста: *работать_в, быть_владельцем, вступить_в_реакцию*
- ◆ При извлечении учитываются типичные конструкции описания атрибутов и отношений
- ◆ Сложность: отношения непостоянны

ФАКТЫ И СОБЫТИЯ (*EVENTS*)



- При извлечении факта/события информация структурируется в виде *семантического фрейма*: (набора параметров-атрибутов события)
- Примеры:

Яндекс купил Кинопоиск за 80\$ млн. в октябре 2013 г.

Фрейм покупки:

атрибуты:	Сумма	Покупатель	Объект	Продавец
	80 млн.\$	Яндекс	Кинопоиск	?

Премьер-министр Казахстана Бакытжан Сагинтаев в апреле 2017 посетил офис Microsoft в Сан-Франциско

Фрейм делового визита:

Визитер	Принимающая сторона	Дата
Бакытжан Сагинтаев	офис Microsoft	04.2017

СЛОЖНОСТИ ИЗВЛЕЧЕНИЯ ФАКТОВ И СОБЫТИЙ



- Событие/факт в тексте может выражаться по-разному

Минобороны РФ ответило британскому министру обороны...

В Минобороны РФ ответили на обвинения британского министра...

- Часто сложно найти слово или словосочетание, которое выражает суть события
- Могут встречаться слова, меняющие суть (*почти, не*)
- Нередко необходимо слияние частичных описаний, полученных из разных предложений

ПОДХОДЫ К ИЗВЛЕЧЕНИЮ ИНФОРМАЦИИ



- **Машинное обучение:** необходим обучающий корпус +
 - ✓ деревья принятия решений (*DT*)
 - ✓ скрытая марковская модель (*HMM*)
 - ✓ модель максимальной энтропии (*MI*)
 - ✓ и др.
- **Инженерный подход:** применение лингвистических правил и шаблонов, содержащих лексическую и грамматическую информацию об извлекаемой конструкции
 - ✓ правила и шаблоны составляют эксперты
 - ✓ часто применяются специальные языки записи правил и поддерживающие их системы
- **Комбинирование** этих подходов

МАШИННОЕ ОБУЧЕНИЕ



Например, для именованных сущностей:

- Имена и названия размечаются:
*[Владислав]_{PERS} [Сурков]_{PERS} [встретится]_O [с]_O
[президентом]_O [Абхазии]_{LOC} [Раулем]_{PERS} [Хаджимба]_{PERS}*
- Определяются различные признаки слов:
 - признаки самого токена (слова): регистр букв, лемма, часть речи, длина и др.
 - словарные признаки (вхождение в опред. словарь)
- По этим частным данным выявляются общие закономерности
- Методы машинного обучения различаются способами учета признаков

МАШИННОЕ ОБУЧЕНИЕ: СОВРЕМЕННЫЕ ТЕНДЕНЦИИ



- Применение лингв. ресурсов (Wikipedia, Freebase): классификация сущностей, атрибутов и т.д.
- Поиск естественно размеченных данных
- Учет большого числа признаков разного вида
- Использование сложной разметки
- Применение мета-алгоритмов обучения, *bootstrapping*: обучение начинается с небольшого количества размеченных данных, итеративное расширение обучающего множества

ЭТАПЫ ОБРАБОТКИ ТЕКСТА ПРИ ПОДХОДЕ НА ПРАВИЛАХ



- Графематика (токенизация)
- Морфологический анализ
- Лексический анализ
 - разрешение лексической многозначности
- Синтаксический анализ
 - при использовании *лингвистических шаблонов* часто достаточно частичного синтаксического анализа (*shallow approach, shallow parsing*)
- Дискурсивный и семантический анализ
 - анализ анафорических ссылок, кореференции
 - слияние (объединение) извлеченных атрибутов фактов/событий в единое описание

ЛИНГВИСТИЧЕСКИЕ ШАБЛОНЫ



- *Лингвистический шаблон* – описание языковой конструкции, ее лексического состава и грамматических свойств:

N «работает в» NP

N – существительное

N «купил» N

NP (Noun Phrase) –

группа существительного

- Основные элементы шаблонов:
 - Словоформы, леммы: возможно указание части речи/морфологических характеристик
 - Грамматические образцы: именные и др. группы
A + N – информационная система (A – Adjective)
- Шаблоны могут учитывать особенности слов:
регистр букв, последовательности букв

ЛИНГВИСТИЧЕСКИЕ ШАБЛОНЫ: ПРИМЕРЫ



- Объекты и атрибуты:
([А-Я] [а-я]+)банк – Собинбанк
A N Ngen – ведущая актриса театра
- Отношения между объектами (*NP* – именная группа)
NP1 «является частью» NP2
Процессор является частью компьютера
- Факты и события:
Loc «пошел ко дну» Ship
Ship «затонул» Loc
В Бенгальском заливе пошел ко дну корабль ВМС
Индии
Корабль ВМС Индии затонул в Бенгальском заливе

ИЗВЛЕЧЕНИЕ НА ПРАВИЛАХ: СОВРЕМЕННЫЕ ТЕНДЕНЦИИ



- Автоматическое расширение набора правил
- Автоматизация построения шаблонов и правил
 - Имеется множество пар сущность – отношение
 - В текстах выявляются упоминания этих пар, анализируется контекст (слова слева и справа)
 - Формируются наиболее вероятные шаблоны
 - Эти шаблоны проверяются на других текстах
- Применение инструментальных систем со специализированными языками правил

ИНСТРУМЕНТЫ ПОСТРОЕНИЯ СИСТЕМ ИЗВЛЕЧЕНИЯ



Свободно доступны

- Построение на основе обучения:
 - OpenNLP* (именованные сущности)
 - StanfordCore NLP* (все виды *IE*)
 - Есть встроенные машинные классификаторы и возможность построить свои
 - Не ориентированы на русскоязычные тексты
- Построение на основе правил:
 - GATE*, Томита-парсер, *LSPL*
 - Формальный язык правил и шаблонов, с их помощью – настройка на конкретную задачу *IE*
 - Возможность работы с русскими текстами

СИСТЕМА *GATE*



Среда построения *IE*-приложений на любом ЕЯ
(<https://gate.ac.uk>) Архитектурные особенности:

- Есть набор стандартных компонентов (процессоры, словари) для обработки текста
- Приложение собирается из этих компонентов (с помощью среды пользователя)
- В ходе обработки текста информация хранится в форме *аннотаций* – наборов атрибутов вида:
имя_атрибута == значение_атрибута
- Анализ текста – последовательное создание и переработка аннотаций по заданным правилам
- Язык лингвистических правил *Jape*

СИСТЕМА *GATE*: *JARE*-ПРАВИЛА



- Правила на языке *Jare* состоят из двух частей:
 - ✓ левая – шаблон для текстового фрагмента
 - ✓ правая – действия с его аннотациями
- Правило для выявления города рождения во фразах вида *Иван родился в Самаре*:

Rule: BornPlace

*(({Token.kind == word, Token.orth == upperInitial}):
person*

{Token.string == "родился"}

{Token.string == "в"}

({Lookup.majorType == "City"}): city)

--> person.Name = {BirthCity = city.Token.string}



ТОМИТА-ПАРСЕР

- Система ориентирована на извлечение фактов из текстов на русском языке
(<https://tech.yandex.ru/tomita/>)
- Последовательность этапов обработки текста фиксирована
- Правила извлечения записываются на языке расширенных КС-грамматик (с условиями)
- Широкий набор встроенных помет-ограничений на значения грамматических и неграмматических характеристик слов
- Встроенные средства описания структуры извлекаемых фактов

ТОМИТА-ПАРСЕР: ИЗВЛЕЧЕНИЕ ФАКТОВ



- Пример правила:

$S \rightarrow \text{Person} \langle \text{gn-agr}[1] \rangle \text{interp}(\text{BornFact.Person})$
 $\text{Born} \langle \text{gn-agr}[1] \rangle \text{"в"} \text{City interp}(\text{BornFact.City});$

где $\langle \text{gn-agr}[1] \rangle$ – согласование по роду и числу,
interp – в какой факт нужно извлечь слово

- Из фразы *Иван родился в Самаре* извлекается факт $\text{BornFact} \{ \text{Person} = \text{Иван}; \text{City} = \text{Самара} \}$
- Типы фактов описываются отдельно от грамматики

$\text{message BornFact: NFactType.TFact} \{$
 $\quad \text{required string Person} = 1;$
 $\quad \text{required string Place} = 2; \}$

LSPL И ЕГО ПРОГРАММНЫЕ СРЕДСТВА



- Декларативный язык для описания конструкций, распознаваемых в РЯ текстах, в виде лексико-синтаксических шаблонов, (www.lspl.ru)
- Визуальная среда анализа текстов по шаблонам
- Программные средства поддержки языка позволяют
 - находить в тексте конструкции по их шаблонам
 - преобразовывать найденные конструкции в некоторый текст

Правило для извлечения города рождения из фраз вида
Иван родился в Самаре:

BornPlace = N V<родиться> "в" City =text> #City

СРАВНЕНИЕ ИНСТРУМЕНТАЛЬНЫХ СИСТЕМ



	<i>GATE</i>	<i>Томита-парсер</i>	<i>LSPL</i>
Извлечение информации	Нет	Есть	Есть
Визуальная среда	Есть	Нет	Есть
Поддержка РЯ	Недостаточная	Достаточная	Достаточная
Язык шаблонов	Сложный и громоздкий	Мощный для извлечения фактов	Удобный, но менее мощный
Прилагательное в именительном падеже: <i>JAPE:</i> {Morph.SpeechPart="Adjective", Morph.Case="Nominative"} Томита-парсер: Adj<gram="nom"> <i>LSPL:</i> A<c=nom>			
Особенности	Можно выстроить свою посл-сть этапов анализа	Извлечение фактов	Возможность автоматической генерации шаблонов

ОЦЕНКИ КАЧЕСТВА ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ



Результаты работы системы

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

ДРУГИЕ МЕРЫ ЭФФЕКТИВНОСТИ ИЗВЛЕЧЕНИЯ ИНФОРМАЦИИ



- *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



- ❖ *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
- ❖ *FactRuEval 2016* (в рамках конференции Диалог)



MUC-6 и MUC-7

- 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: $P=90\%$, $R=20\%$

FactRuEval 2016

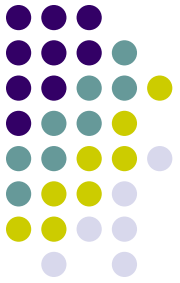


- Новостная коллекция
- Три подзадачи выделения:
 - ✓ именованных сущностей (организация, персона, географический объект)
 - ✓ сущностей и их атрибутов
 - ✓ фактов из текстов (найм, сделка, владение, встреча)
- 13 участников, большинство использовали инженерный подход с элементами статистики
- Автоматическая система оценки результатов, в открытом доступе
- Лучшие значения – для 1 и 2 дорожки ($F1 = 93\%$), худшие – для фактов ($F1=66\%$)

ЗАКЛЮЧЕНИЕ



- Задача *IE* актуальна, много приложений
- Подходы к решению:
 - машинное обучение: хорошо работает для извлечения сущностей
 - подход на правилах обычно лучше для выделения событий и фактов
- Направления развития:
 - проведение более глубокого синтаксического анализа и использование синтаксических признаков при машинном обучении
 - визуализация извлеченной и структурированной информации, удобная для человека-аналитика



СПАСИБО ЗА ВНИМАНИЕ!