

Text mining

5. Извлечение информации

Дмитрий Ильвовский, Екатерина Черняк

dilvovsky@hse.ru, echernyak@hse.ru

Национальный Исследовательский Университет – Высшая Школа Экономики
НУЛ Интеллектуальных систем и структурного анализа

March 10, 2016

Извлечение информации (Information extraction)

- Извлечение небольших и легко формализуемых фрагментов текста (текстов)
- Которые могут записаны в т.н. базы знаний (аналоги реляционных баз данных)
- Сама по себе такая формализация полезна для поиска и организации текстов
- И может быть использована для последующего автоматизированного анализа

Временные метки

Кузнецов Дмитрий To: Ильвовский Дмитрий Алексеевич Cc: Андрей Р Reply-To: Кузнецов Дмитрий Re: НА: занятия 10.03	занятия 10.03 Location Wednesday 9 March 2016 all-day
Ориентироваться лучше на пораньше (12.40 - 12.50), чт	

Парсеры на основе небольших словарей и регулярных выражений

Простые (одноместные) факты

Высшая школа экономики

[Веб-сайт](#) [Как проехать](#)

Университет в Москве, Россия

Национальный исследовательский университет «Высшая школа экономики» — один из ведущих и крупнейших университетов России. Федеральное государственное высшее учебное заведение, созданное в 1992 году. Находится в городе Москва. [Википедия](#)

Адрес: Мясницкая ул., 20, Москва, 101000

Основано: 27 ноября 1992 г.

Телефон: 8 (495) 771-32-32

Талисман: Ворона

Набор: 25 667 (21 апр. 2015 г.)

Выдающиеся выпускники

Ещё 3+

Тимати

Андрей Юрьевич Воробей

Сергей Пантеле...

Роза Рахимовна Сябитова

Алексей Евгеньевич Релик

После окончания средней школы (1979 год) Роза Сябитова поступила в Московский институт электронного машиностроения (wiki)

Простые правила “заголовок статьи – имя человека И заголовок статьи – название вуза”

Именованные сущности

Глава МИД России Сергей Лавров ответил Киеву на вопрос, почему к Надежде Савченко не были допущены украинские врачи для медицинского освидетельствования.

Граф объектов

☒ Большая картинка ☐ Средняя картинка ☐ Маленькая картинка ☐ Скрыть картинку

Министерство Иностранных
Дел (Мид) - объект

Россия

Киев

Савченко Н.

Лавров С.

Список объектов (можно выбирать текущий)

Тип	Краткое описание
Территориальное образование	Россия
Территориальное образование	город Киев
Организация	Министерство Иностранных Дел (Мид), Россия
Свойство персоны	глава; Министерство Иностранных Дел (Мид), Россия
Персона	Сергей Лавров
Персона	Надежда Савченко

Текущий объект

Атрибут	Значение
Наименование	РОССИЯ
Код страны	RU
Наименование	РОССИЙСКАЯ ФЕДЕРАЦИЯ
Тип	государство

<http://pullenti.ru/>

Задача извлечения именованных сущностей

Требуется:

- найти именованные сущности
- определить их тип

Глава МИД России Сергей Лавров ответил Киеву на вопрос, почему к Надежде Савченко не были допущены украинские врачи для медицинского освидетельствования.

Глава МИД России Сергей Лавров ответил Киеву на вопрос, почему к Надежде Савченко не были допущены украинские врачи для медицинского освидетельствования.

Location

Person

Извлечение именованных сущностей

Необходимо для

- умной индексации текстов
- анализа мнений (как пользователи относятся к какому-то бренду)
- последующего извлечения фактов
- создания вопросно-ответных систем (“Кто изобрел колесо?”)
- создания справочников и пресс-портретов

Извлечение именованных сущностей

- Именованная сущность состоит из нескольких токенов
- Задача: выделить именованные сущности из текста и определить их тип

[Вице-премьер [России] [Дмитрий Рогозин]] на заседании [президиума госкомиссии по вопросам развития [Арктики]] в [Мурманске] заявил, что санкции в отношении [России] в том или ином виде будут всегда.

Вице-премьер *PERSON*

России *PERSON, LOCATION*

Дмитрий *PERSON*

Рогозин *PERSON*

на

заседании

президиума *ORGANISATION*

госкомиссии *ORGANISATION*

по *ORGANISATION*

вопросам *ORGANISATION*

развития *ORGANISATION*

Арктики *ORGANISATION, LOCATION*

в

Мурманске *LOCATION*

...

- Определение границ
на заседании [президиума госкомиссии по вопросам развития
[Арктики]] в [Мурманске]
- Неоднозначность
Сергей Олегович Кузнецов – историк и искусствовед, архитектор,
заведующий дАДИИ ФКН
- Оценка качества: точности и полноты мало

Методы извлечения именованных сущностей

- По внешнему источнику: онтологиям, словарям, Википедии
- По правилам
- С использованием машинного обучения
 - ▶ алгоритмы классификации
 - ▶ модели последовательностей

Извлечение именованных сущностей по дереву категорий Википедии

Президент РФ
Владимир Путин
считает, что
высказывания в ЕС
по поводу решения
Киева
приостановить
процесс интеграции
с Евросоюзом
оказывают
давление на
Украину

```
http://ru.wikipedia.org/wiki/Президент
.../wiki/Президент_Российской_Федерации
.../wiki/Россия
.../wiki/Владимир
.../wiki/Владимир_Путин
.../wiki/Высказывание
.../wiki/В
.../wiki/Европейский_союз
.../wiki/По
...wiki/Решение
...wiki/Киев
.../wiki/Процесс
.../wiki/Интеграция
.../wiki/С
.../wiki/Европейский_союз
.../wiki/Давление
.../wiki/На
.../wiki/Украина
```

Нужно отсекать лишних кандидатов!

Извлечение именованных сущностей по правилам

Можно использовать Томита парсер. Попробуем извлечь имена в виде
“И. О. Фамилия”

Initial -> Word<wff=/[А-Я]/>;

Initials -> Initial Initial;

FIO -> Initials Word<kwtype=surname>;

S -> FIO;

Извлечение именованных сущностей с использованием машинного обучения

Нужны размеченные данные

- CoNLL 2003 shared task 2003
- FactRuEval — соревнование по выделению именованных сущностей и извлечению фактов. Разметка корпуса осуществляется с помощью краудсорсинга. 4 класса: Person, Org, Location, LocOrg

The screenshot displays the FactRuEval interface. On the left, a text snippet is shown with various entities highlighted in green and red. On the right, a table lists the identified entities with their corresponding labels and types.

Text snippet:

Руководитель НИИ транспорта и дорожного хозяйства Михаил Бликин утверждает, что может помочь Сергею Собянину решить транспортные проблемы столицы, в частности разгрузить МКАД. « Сама постановка задачи — вернуть МКАД исходное значение — мне чрезвычайно импонирует. Я все последние годы писал и говорил о том, что МКАД превратилась у нас в просёлочную дорогу, подъезд к магазинам, даже в городскую улицу межквартальную », — цитирует РБК Бликина.

Annotations table:

СПАНЫ	УПОМИНАНИЯ
✖ Н/И	org_desc
✖ Н/И транспорта и дорожного хозяйства	org_name
✖ Михаил	name
✖ Бликин	surname
✖ Руководитель	job
✖ Сергею	name
✖ Собянину	surname
✖ МКАД	loc_name
✖ МКАД	loc_name
✖ МКАД	loc_name
✖ РБК	org_name
✖ Бликина	surname

<https://geektimes.ru/post/267774/>

<http://opencorpora.org/ner.php>

Использование алгоритмов классификации

- Кандидаты
 - ▶ Именные группы
 - ▶ Морфологические шаблоны
 - ▶ n -граммы
- Признаки для классификации
 - ▶ Частота не важна
 - ▶ Форма кандидата: Xxx Xxxx Xxxxx, xXxxx, Xxx-xxx, XXX, XXXd, Xxxxd
 - ▶ Специфические подстроки: “бург”, “река”, “евич”
 - ▶ Положение в предложении
 - ▶ Является ли заголовком в Википедии
 - ▶ Встречается ли в кандидат в заголовке текста
 - ▶ Контекст: POS-теги слов до и после кандидата
 - ▶ Контекст: специфические слова до и после кандидата (заявил, постановило, городе)
- Алгоритм классификации
 - ▶ На два класса (именованная сущность или нет)
 - ▶ На несколько классов (именованная сущность или нет, типы именованных сущностей)

Использование моделей последовательностей

- Разметка токенов:
 - ▶ IO (inside-outside) разметка
 - ▶ IOB (inside-outside-begins) разметка
- Признаки для классификации
 - ▶ Текущий токен
 - ▶ Контекст: предыдущий и следующий токен
 - ▶ POS тег
 - ▶ и другие
- Обучение классификатора последовательностей (Марковские модели максимальной энтропии, условные случайные поля)
- Тестирование: IO или IOB разметка входного текста

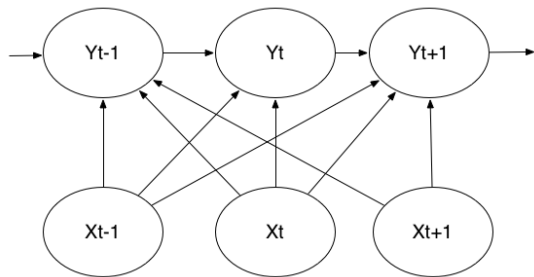
Глава	МИД	РФ	С.	Лавров	ответил	Киеву	на	вопрос	,	почему	к	Н.
Per	Per	Per	Per	Per	O	Loc	O	O	O	O	O	Per
B-Per	I-Per	I-Per	I-Per	I-Per	O	B-Loc	O	O	O	O	O	B-Per

Марковская модель максимальной энтропии

Алгоритм восстановления условного распределения

$$P(y_{t+1}|y_t, x),$$

где y – IO или IOB разметка, x – токены. Существует несколько режимов вывода в этой модели, например, вывод Витерби .



Извлечение именованных сущностей

История успеха: достигается точность порядка 95%!.

Извлечение фактов

НИУ ВШЭ — один из ведущих и крупнейших университетов России. Федеральное государственное высшее учебное заведение (национальный исследовательский университет), созданное в 1992 году. Находится в городе Москва. Высшая школа экономики (ВШЭ) была создана 27 ноября 1992 года в соответствии с Постановлением Правительства Российской Федерации, подписанном Егором Гайдаром, исполнявшим в то время обязанности Председателя Правительства.

Сложный факт “Создание НИУ ВШЭ”:

- где: Москва
- когда: 27 ноября 1992 года
- приказ: Егор Гайдар
- первое название: ВШЭ

Простой факт “Создание НИУ ВШЭ”:

- где: Москва
- когда: 1992

Простой факт — это кортеж из **трех** элементов: (НИУ ВШЭ, Москва, 1992)

Нужно для:

- Создания и пополнения баз знаний
- Вопросно-ответных систем
- Навигации по большим коллекциям текстов

FactRuEval – соревнование по выделению именованных сущностей и извлечению фактов

- Occupation (работа персоны в организации): who, where, position
- Deal (сделка между несколькими сторонами без указания её предмета и условий): participant1, participant2, (participant3):
- Ownership (владение организацией): owner, property, phase
- Meeting (встреча нескольких персон): participant1, (participant2)

<http://www.dialog-21.ru/adx/aspx/adxGetMedia.aspx?DocID=6258b1c6-642e-418f-b63a-0797db6b3521>

- Is-A отношение: кошка – это млекопитающее
- Part-Of отношение: двигатель – это часть машины
- Instance-Of: Москва – город

Методы извлечения фактов

- По правилам и шаблонам
- С использованием машинного обучения

Целевое назначение Томиита парсера

```
S -> merge Company interp(PurchaseFact.Company1) "с" Company  
interp(PurchaseFact.Company2);
```

```
S -> merge Company interp(PurchaseFact.Company1) "и" Company  
interp(PurchaseFact.Company2);
```

Извлечения фактов по шаблонам

Лексико-семантические шаблоны Хирст

- such NP as NP, NP[,] and/or NP;
- NP such as NP, NP[,] and/or NP;
- NP, NP [,] or other NP;
- NP, including NP, NP [,] and/or NP;
- NP, especially NP, NP [,] and/or NP.
- Поиск в Google по запросу "such * as *, * and *"
In the dissociative type, alterations may occur in the patient's state of consciousness or in his identity, to produce such symptoms as amnesia, somnambulism, fugue, and multiple personality.
- Поиск в Google по запросу "такие * как *, * и *"
Такие слова как честь, совесть или достоинство просто не входят в круг их понятий. Миньоны преданы только злему хозяину...

Извлечения фактов по шаблонам

Плюсы:

- Высокая точность
- Для каждой предметной области можно создать свой набор шаблонов

Минусы:

- Низкая полнота
- Трудозатратно в высшей мере

Извлечения фактов с использованием машинного обучения

Есть корпус, в котором размечены именованные сущности и отношения между ними. Задача: обучить классификатор выделять и определять типы отношения между парами или тройками сущностей

- Кандидаты: пары или тройки именованных сущностей, встречающихся в пределах окна
- Признаки:
 - ▶ тип1, тип2, тип1+тип2
 - ▶ мешки слов: объединенный мешок слов из именованных сущностей-кандидатов, мешок слов перед именованной сущностью 1 и после именованной сущности 2
 - ▶ синтаксические пути и типы синтаксической связи
- Любой алгоритм классификации
- Стандартные меры качества: точность и полнота
- Нужно много размеченных данных
- Классификатор может быть чувствителен к предметной области