Извлечение информации

Батыгин Владимир
vbatygin@yandex-team.ru
Computer Science Center

План

- Что такое Information Extraction
- Источники данных
- Подходы
- Заключение

Задачи

- Named Entity Recognition
- Disambiguation (= Conference resolution)
- Relationship Extraction (= Fact Extraction)

Named Entity Recognition

Первый мотоцикл с которого началась история знаменитого завода JAWA, был представлен на Пражском автосалоне в 1929-м году.

- Мотоцикл — объект
- JAWA — марка
- Пражском автосалоне — место
- 1929 — год

Named Entity Recognition

Part of speech tagging (POS)

По техническим характеристикам мотоциклы JAWA не уступают ведущим европейским и американским линиям.

- По {предлог}
- Техническим {прилагательное, дат, мн}
- Характеристикам {существительное, жен, дат, мн}
- Мотоциклы {существительное муж, им, мн}
- JAWA {JAWA??}
- Не {частица}
- Уступают {галгол, мн, изъяв, 3-л, несов}
- ...

Disambiguation

1 .bass как тип рыбы (окунь)

2. bass как звуки низкой частотности

- I went fishing for some sea bass.
- The bass line of the song is too weak.

Disambiguation (POS)

По техническим характеристикам мотоциклы JAWA не уступают ведущим европейским и американским линиям.

Ведущим:

- Глагол, несов, непрош, действ,
- Существительное ,муж,од, дат,мн
- Прилагательное, мн

Relationship Extraction (Fact Extraction)

- Михаил Афанасьевич Булгаков родился в Киеве 3 мая 1891 года.

- Отношения: тройки, вида <X, R, Y>

  <Булгаков, родился, 3 мая>

  <Булгаков, родился в, Киев>

Применение

- Сервисы, основанные на данных

Сортировать по: релевантности цене дате добавления году пробегу году и цене

399 000 руб.
торг

2008' Ford Focus

Седан, цвет красный, отличное сост., 70000 км

передний привод, 1.4 л, механика, инжектор, ABS, подушки безопасности (2), иммобилайзер, кассетная магнитола, бортовой компьютер, обогрев зеркал, сигнализация, центральный замок, передние электростеклоподъемники, электропривод зеркал

Санкт-Петербург, 1 ноября,
www.autolot24.ru копии в избранное вычеркнуть

465 000 руб.
торг

2008' Ford Focus II

Седан, цвет синий, отличное сост., 75000 км

передний привод, 1.8 л, механика, инжектор, ABS, г/у руля, иммобилайзер, ксенон, магнитола cd, бортовой компьютер, обогрев зеркал, подогрев сидений, сигнализация, центральный замок, электростеклоподъемники все, электропривод зеркал

Санкт-Петербург, 1 ноября, Реклама-ШАНС копии в избранное вычеркнуть

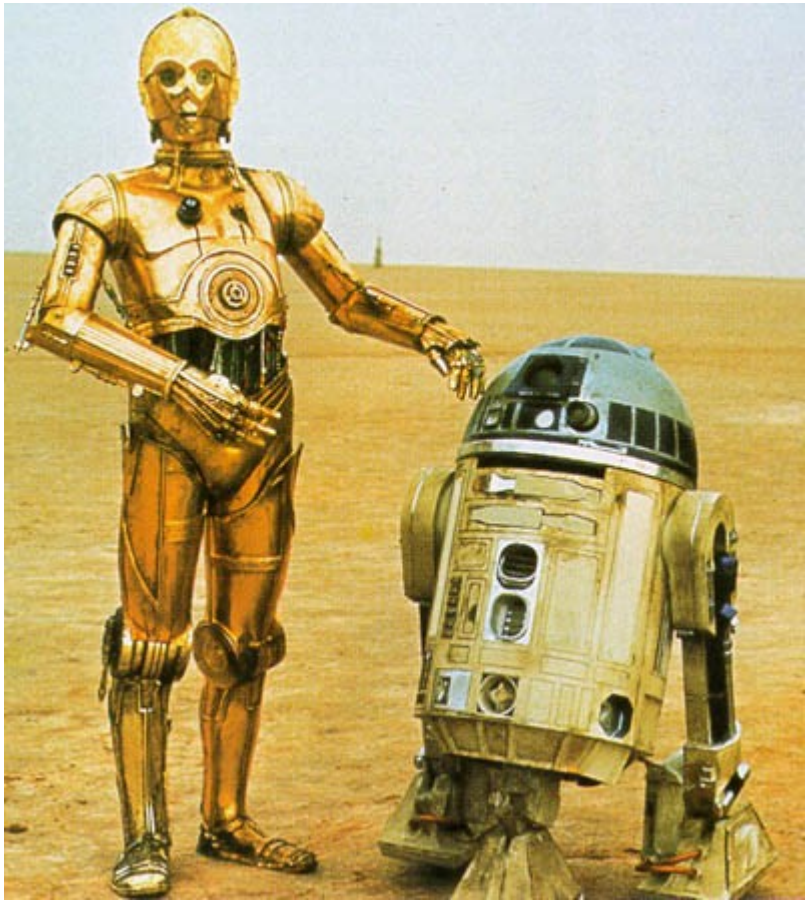
Точные ответы на вопросы

- В каком году родился текущий президент России?
- Какая футбольная команда победила в Лиге Чемпионов?
- Кто написал роман «Мастер и Маргарита»?

Пример: IBM Watson

Применение

- Понимание естественного языка



Применение

- Понимание естественного языка
- Сервисы, основанные на данных
- Точные ответы на вопросы
- Какие ещё варианты использования?

Применение

- Понимание естественного языка
- Сервисы, основанные на данных
- Точные ответы на вопросы
- Извлечение мнений и отзывов
- Извлечение информации из ДНК

План

- Что такое Information Extraction
- Источники данных
- Подходы
- Заключение

Источники данных

- Энциклопедические и тематические ресурсы
- Веб страницы
- Логи поисковых машин

Энциклопедические и тематические ресурсы

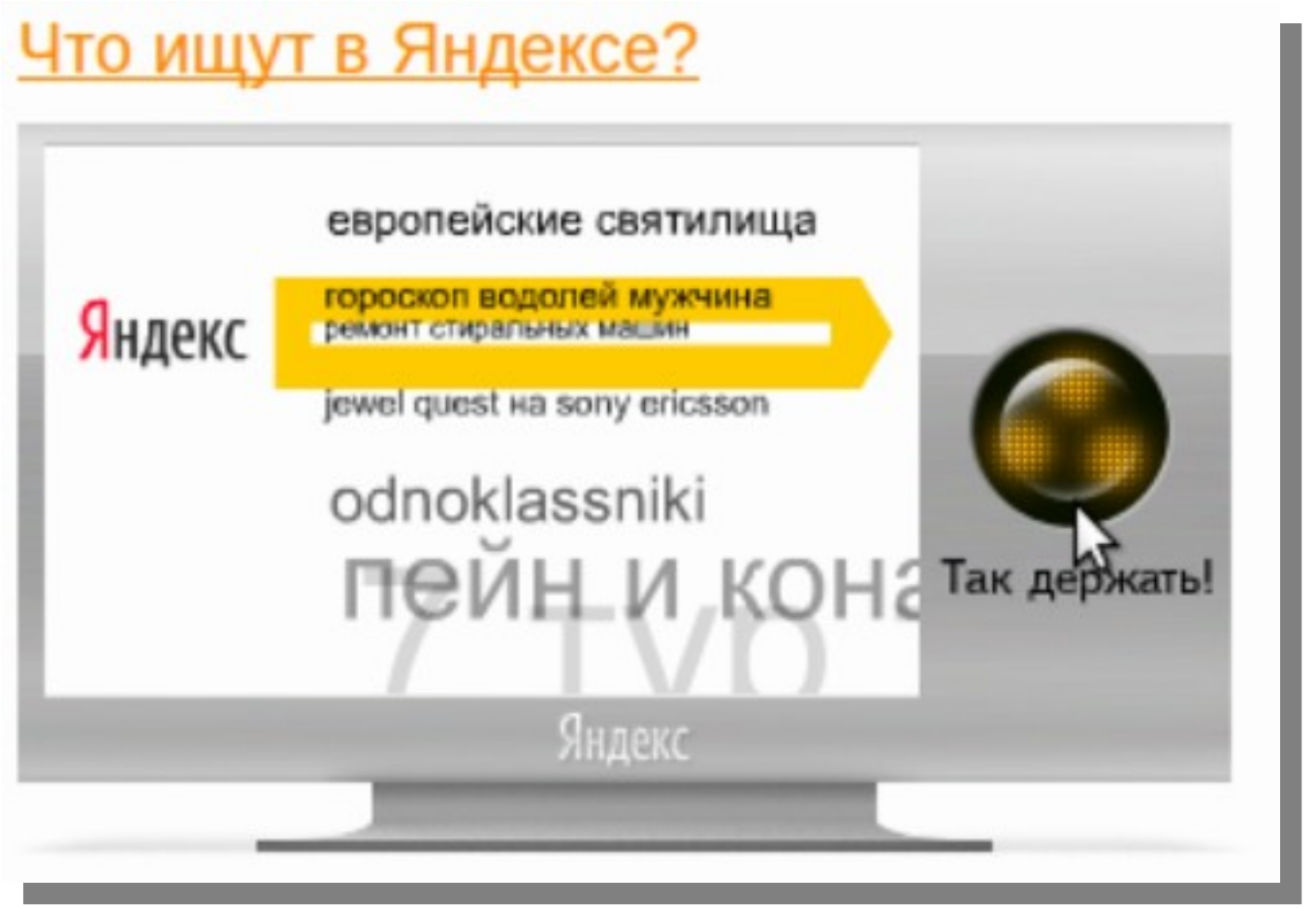
- Wikipedia
- Тематические ресурсы: IMDB, lastfm, ...
- Социальные сети: Facebook, twitter, ...

Веб страницы

- Текст
  - Статьи
  - Новости
  - Отзывы
- Полуструктурированный текст
  - Списки, таблицы
  - Сайты с регулярной структурой

Запросы

- Маленький размер (обычно 1 — 3 слова)
- Много мусора



NER для запросов

Запрос: dell inspiron M301z 13'3

- Dell — производитель
- Inspiron — линейка
- M301z — модель
- 13'3 — размер экрана

Что можно извлечь из запросов?

- Столица Франции
- Альбомы Pink Floyd
- Angry birds для ipad

Что можно извлечь из запросов

- Столица Франции
  - У Франции есть столица
- Альбомы Pink Floyd
  - У Pink Floyd есть альбомы
- Angry birds для ipad
  - На ipad можно играть в Angry birds

План

- Что такое Information Extraction
- Источники данных
- Подходы
- Заключение

Подходы

- Основанные на онтологиях
- Статистические
- Основанные на правилах

Онтологии

Описывают множество объектов и связей между ними

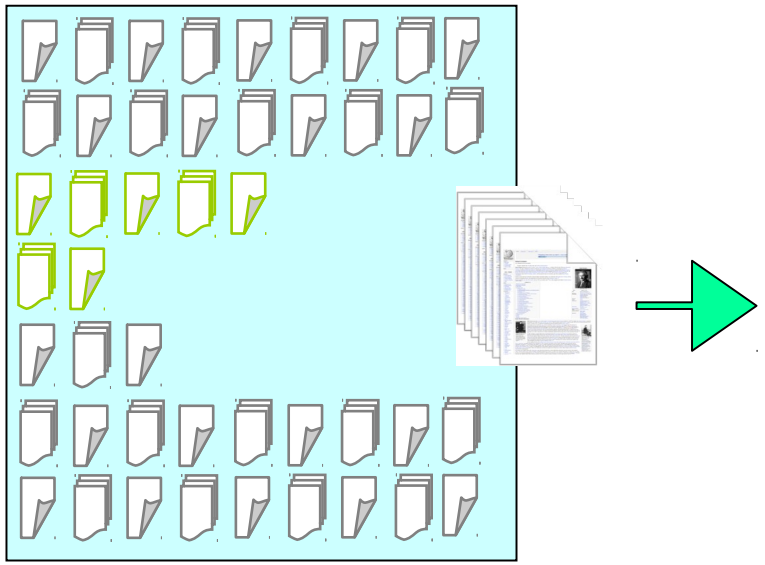
Формально: X = <S,R,O>, где

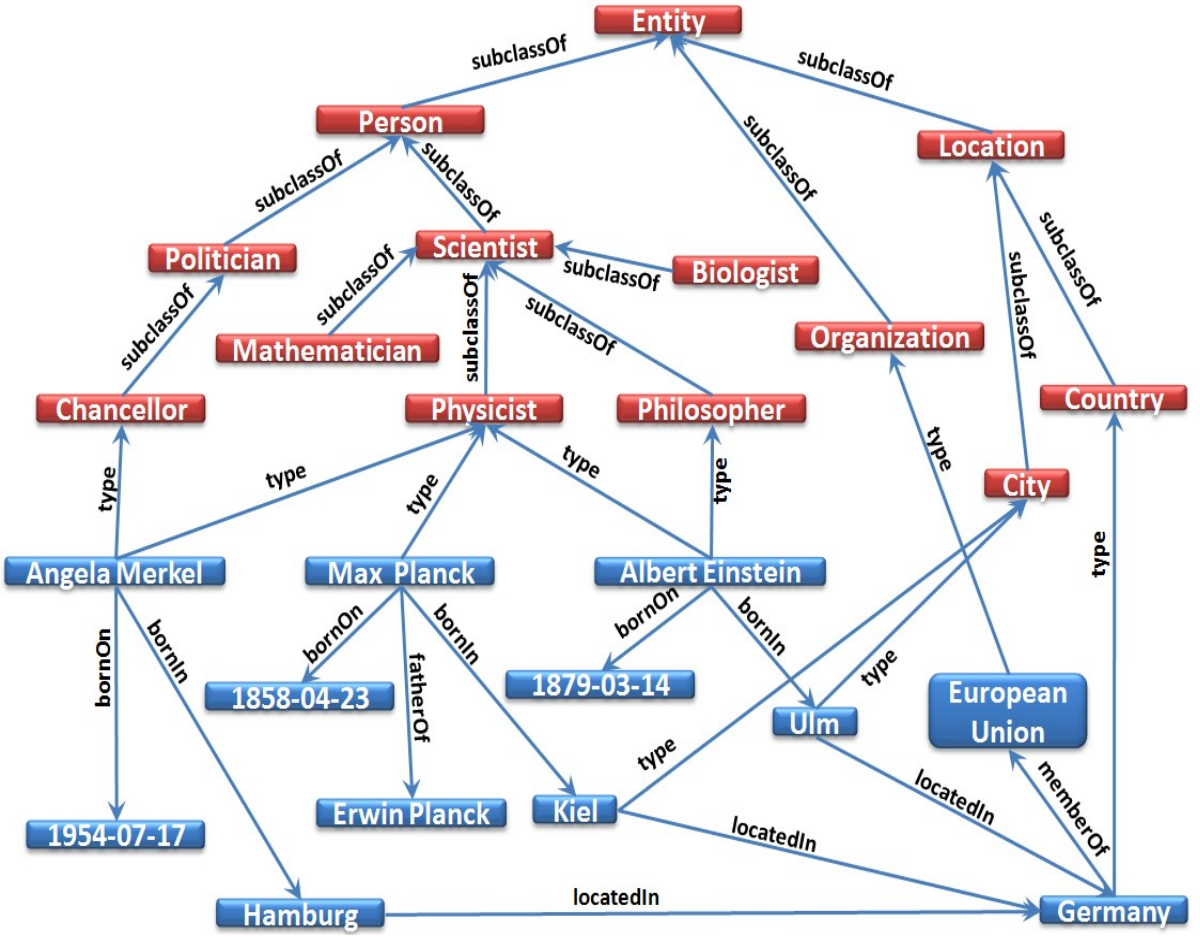
- S — объект
- R — отношение
- O — объект

Пример: Булгаков родился 3 мая в Киеве

- <Булгаков, родился, 3 мая>
- <Булгаков, родился в, Киев>

Онтологии




Entity
Person
Location
Politician
Scientist
Biologist
Mathematician
Organization
Country
Chancellor
Physicist
Philosopher
City
Angela Merkel
Max Planck
Albert Einstein
1858-04-23
1879-03-14
Ulm
European Union
1954-07-17
Erwin Planck
Kiel
Hamburg
Germany
subclassOf
type
bornOn
bornIn
fatherOf
locatedIn
memberOf

Извлечение информации их википедии

- Инфобоксы
- Категории

Категории: Персоналии по алфавиту | Писатели по алфавиту | Родившиеся 3 июля | Родившиеся в 1883 году | Родившиеся в Праге | Умершие 3 июня | Умершие в 1924 году | Умершие в Клостернойбурге | Писатели-экспрессионисты | Франц Кафка | Умершие от туберкулёза | Писатели Чехии | Писатели Австрии | Выпускники Карлова университета | Авторы знаменитых дневников | Писатели-модернисты | Похороненные на Ольшанском кладбище

Франц Кафка

Franz Kafka

Фотография писателя, 1906 г.

Дата рождения:	3 июля 1883
Место рождения:	Прага, Австро-Венгрия
Дата смерти:	3 июня 1924 (40 лет)
Место смерти:	Кирлинг, Первая Австрийская Республика
Гражданство:	Австро-Венгрия
Род деятельности:	прозаик
Направление:	модернизм, литература абсурда
Жанр:	притча, роман, малая проза
Подпись:	Franz Kafka

Произведения на сайте Lib.ru

Извлечение информации их википедии

Франц Ка́фка (нем. ***Franz Kafka***, 3 июля 1883, Прага, Австро-Венгрия — 3 июня 1924, Клостернойбург, Первая Австрийская Республика) — один из основных немецкоязычных писателей XX века, бо́льшая часть работ которого была опубликована посмертно. Его произведения, пронизанные абсурдом и страхом перед внешним миром и высшим авторитетом, способные пробуждать в читателе соответствующие тревожные чувства[1], — уникальное явление в мировой литературе.

Содержание [убрать]

Франц Кафка

Franz Kafka

Фотография писателя, 1906 г.

Дата рождения:	3 июля 1883
Место рождения:	Прага, Австро-Венгрия

Извлечение информации из википедии

- Дополнение инфобоксов википедии
- Извлечение связей из категорий википедии
- Извлечение фактов из статей

Примеры существующих систем

DBpedia

- 3.64 миллиона сущностей
- 97 различных языков
- 1 миллиард фактов (RDF triples)

About: Bastian Schweinsteiger

An Entity of Type : soccer player, from Named Graph : http://dbpedia.org, within Data Space : dbpedia.org

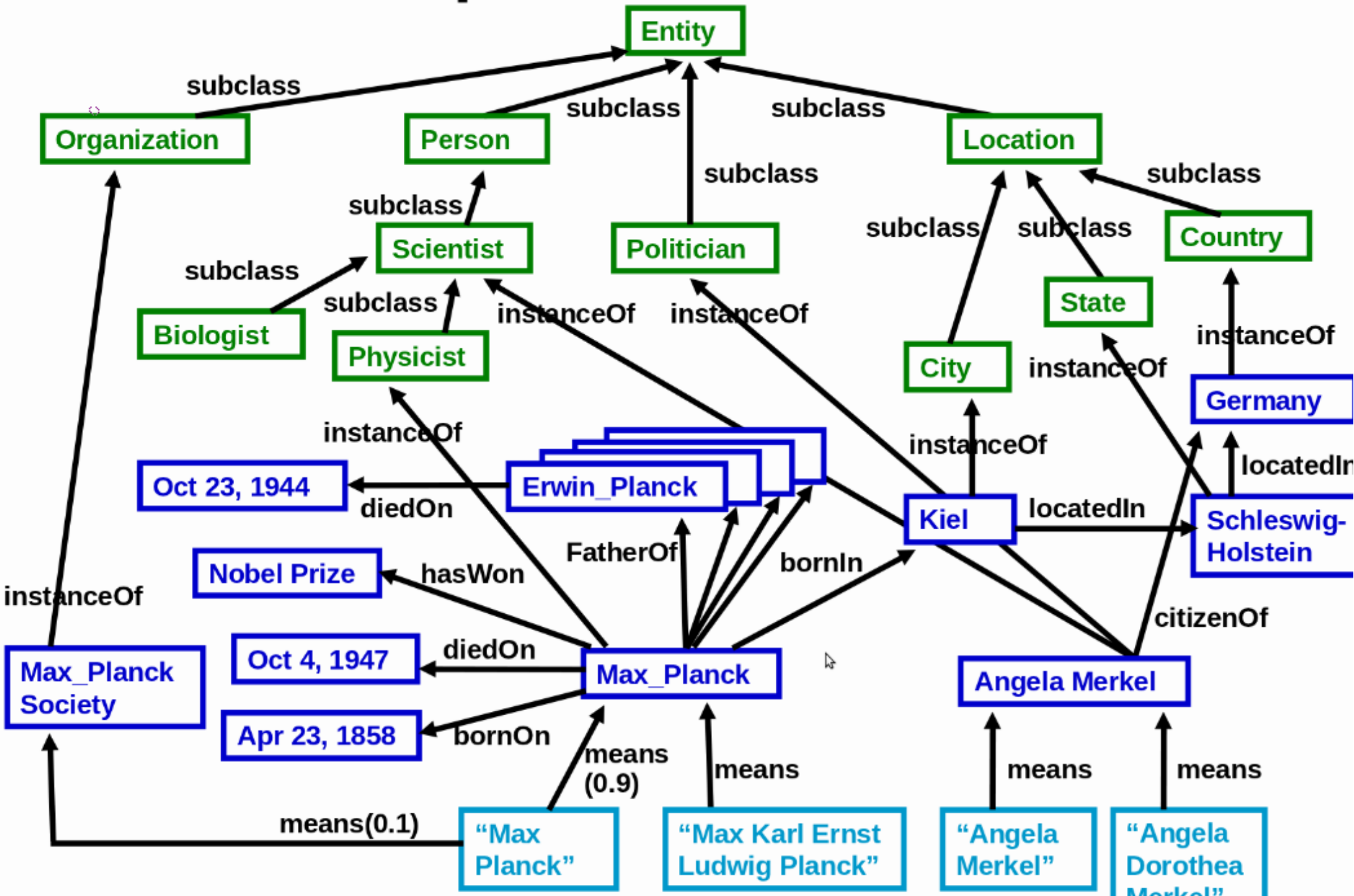
dbpedia-owl:birthPlace	▪ dbpedia:Kolbermoor ▪ dbpedia:West_Germany
dbpedia-owl:number	▪ 31 (xsd:integer)
dbpedia-owl:position	▪ Midfielder
dbpedia-owl:team	▪ dbpedia:FC_Bayern_Munich
dbpprop:hasPhotoCollection	▪ http://www4.wiwiss.fu-berlin.de/flickrwrappr/photos/Bastian_Schweinsteiger
dbpprop:reference	▪ http://www.fcbayern.t-com.de/en/teams/profis/00397.php?fcb_sid=a35deceb606f7dd80f23b7 ▪ http://www.bastian-schweinsteiger.de ▪ http://www.fussballdaten.de/spieler/schweinsteigerbastian/
dbpprop:wordnet_type	▪ http://www.w3.org/2006/03/wn/wn20/instances/synset-soccer_player-noun-1
rdf:type	▪ yago:PeopleFromBavaria ▪ yago:GermanyInternationalFootballers ▪ yago:BayernMunichPlayers ▪ owl:Thing ▪ yago:FirstBundesligaFootballers ▪ yago:GermanyUnder-21InternationalFootballers ▪ yago:BayernMunichIIPlayers ▪ yago:FIFAWorldCup2006Players ▪ yago:LivingPeople ▪ yago:UEFAEuro2004Players ▪ yago:GermanFootballers ▪ dbpedia-owl:Person ▪ dbpedia-owl:Athlete ▪ dbpedia-owl:SoccerPlayer

Примеры существующих систем

YAGO

- 2 Миллиона сущностей
- 200 тысяч классов
- 40 миллионов фактов (RDF triples)

KB's: Example YAGO (Suchanek et al.: WWW'07)



http://www.mpi-inf.mpg.de/yago-naga/

Тематические онтологии

- Автомобили
  - Название
  - Коробка передач
  - Цена
  - Тип кузова
  - Двигатель
  - Цвет
  - Пробег
  - ...

Особенности тематических онтологий

- Высокая точность (+)
- Низкая полнота (-)
  - Извлекается только то, что есть в онтологии
- Хорошо использовать для закрытых классов
  - Кинобазы, автомобили, товары ...

Подходы

- Основанные на онтологиях
- Статистические
- Основанные на правилах

Статистический подход

- Использует машинное обучение
- Требуется большой размеченный корпус
- Не требуется знаний лингвистики ;)
- Подходит для любого типа текстов (статьи, новости, сообщения в соц. Сетях)
- Сложно контролировать и настраивать

Схема

- Составление корпуса
- Предобработка
- Построение шаблона или обучение классификатора
- Подтверждение полученных данных для повышения точности
- Bootstrapping!

Составление корпуса

- Для английского языка можно найти готовые корпуса
- Для русского надо размечать вручную

Предобработка

Предложение: По техническим характеристикам мотоциклы JAWA не уступают ведущим европейским и американским линиям

- POS tagging
  - По {предлог}
  - Техническим {прилагательное, дат, мн}
  - Характеристикам {существительное, жен, дат, мн}
- Chunking
  - Анализ предложений с целью выявления составляющих частей (групп существительных(noun phrases), глаголов)
  - По техническим характеристикам мотоциклы JAWA не уступают ...

Построение шаблона

- Шаблоны
  - NP1 «such as» NP2 (NP = noun phrase)
  - NP1 «,» «CEO of» NP2
  - * «such as» I «and» J (I и J уже известны)

Классификаторы

- Альтернативный шаблонам метод — построение классификатора
- Классификатор обучается для извлечения отношений между найденными noun phrase.
- Используется маркирование последовательностей (sequence labelling)

Маркирование последовательностей

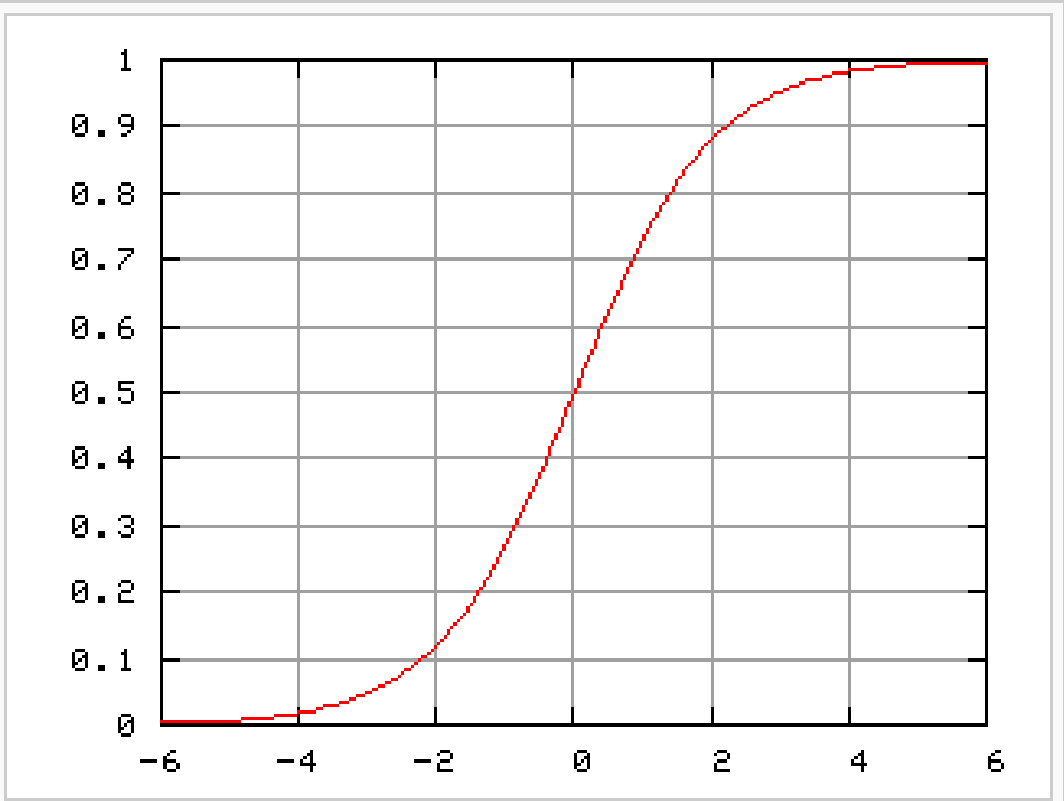
- В основном используются вероятностные алгоритмы
- Важно: Вероятность следующей метки зависит от предыдущей
  - HMM (Hidden Markov Model)

$$P(Y) = \sum_X P(Y \mid X) P(X),$$

    - Y — неизвестная последовательность
    - X — последовательность из тестового множества
  - CRF (Conditional Random Fields)

Поддтверждение

- Также правила или классифкатор. Например Логистическая регрессия



Логистическая функция: $f(x) = \frac{1}{1 + e^{-x}}$.

Список фич для классификатора

- 1.16 (x; r; y) covers all words in s
- 0.50 The last preposition in r is for
- 0.49 The last preposition in r is on
- 0.46 The last preposition in r is of
- 0.43 len(s) 10 words
- 0.43 There is a WH-word to the left of r
- 0.42 r matches VW*P from Figure 1
- 0.39 The last preposition in r is to
- 0.25 The last preposition in r is in
- 0.23 10 words < len(s) 20 words
- 0.21 s begins with x
- 0.16 y is a proper noun
- 0.01 x is a proper noun
- -0.30 There is an NP to the left of x in s
- -0.43 20 words < len(s)
- -0.61 r matches V from Figure 1
- -0.65 There is a preposition to the left of x in s
- -0.81 There is an NP to the right of y in s
- -0.93 Coord. conjunction to the left of r in

Особенности статистического подхода

- Обработка больших объёмов данных, высокая полнота
- Недостаточно развиты NLP инструменты. Например нет chunker'a для русского языка.
- Тяжело настраивать точечно

Подходы

- Основанные на онтологиях
- Статистические
- Основанные на правилах

Построение правил

много лингвистики

Построение правил

- Используется синтаксический анализ
- Словари по каждой тематике составляются вручную лингвистами с учётом всех правил языка.

Оценка

		Condition (as determined by "Gold standard")		
		Positive	***Negative***	
Test outcome	***Positive***	**True Positive**	**False Positive** (Type I error)	→ **Positive predictive value** $= \frac{\Sigma \text{ True Positive}}{\Sigma \text{ Test outcome Positive}}$
	Negative	**False Negative** (Type II error)	**True Negative**	→ Negative predictive value $= \frac{\Sigma \text{ True Negative}}{\Sigma \text{ Test outcome Negative}}$
		↓ Sensitivity $= \frac{\Sigma \text{ True Positive}}{\Sigma \text{ Condition Positive}}$	↓ Specificity $= \frac{\Sigma \text{ True Negative}}{\Sigma \text{ Condition Negative}}$	

Полнота и точность

$$presicion = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + false\ negatives}$$

$$recall = \frac{number\ of\ true\ positives}{number\ of\ true\ positives + false\ positives}$$

План

- Что такое Information Extraction
- Источники данных
- Подходы
- Заключение

Тенденции

- Извлечение информации из соц. Сетей (настроения, мнения)
- Извлечение мнений из отзывов (рекомендательные системы)
- Большинство систем только для английского языка

Инструменты

- OpenNLP http://incubator.apache.org/opennlp/
- Gate platform http://gate.ac.uk/

NLP семинар

- Семинар: Natural Language Processing http://mathlingvo.ru/nlpseminar

Что узнали

- Задачи information extraction
- Применение извлечённой информации
- Как построить базу знаний на основе википедии
- Как извлекать данные с помощью онтологий, машинного обучения и правил