Information Extraction: A short introduction

Rafael Carrascosa

Machinalis

rcarrascosa@machinalis.com

November 4, 2014

About the speaker

- Rafael Carrascosa.
- Background in Computer Science.
- Works as developer and Machine Leaning specialist for Machinalis.
- Machinalis (www.machinalis.com) is a software research and development company based on Argentina.
- Leaded the delopment of IEPY.

Overview

1. Information Extraction
    - What is IE?
    - Applications of IE
    - Limitations of IE
    - How IE is done, a sketch
    - Reasoning and Knowledge Bases
2. IE with IEPY
    - Overview
    - Relation extraction with IEPY
    - Corpora construction

What is Information Extraction

In one line

Is about getting information from text automatically

- Main activity: extract **relation** instances between **entities**.
- **Entity**: Any-thing: A person, a place, an organization, a date, etc.
- **Relation**: Can be any kind of linking concept between entities.
- Also posible but not common: IE over audio, images or video.

An example of entities and relations

Example text

Harvard University is a private Ivy League research university in **Cambridge**, Massachusetts, established 1636.

- `Harvard University`: Is an **entity** of kind organization.
- `Cambridge`: Is an **entity** of kind location.
- $Located(X, Y)$ is a **relation** that links organizations with places.

An example of entities and relations

Example text

Harvard University is a private Ivy League research university in **Cambridge**, Massachusetts, established 1636.

- `Harvard University`: Is an **entity** of kind organization.
- `Cambridge`: Is an **entity** of kind location.
- $Located(X, Y)$ is a **relation** that links organizations with places.

IE deals with detecting that

$$Located(\text{Harvard University}, \text{Cambridge})$$

is a **relation instance** in the text.

Example application: Protests

Example application: Protests

Use news text to build a map of events like politic acts or public protests.

Example application: Protests

Use news text to build a map of events like politic acts or public protests.

- News sites, social networks
- $happened\_at(protest, date)$
- $happened\_in(protest, location)$
- Data can end up in a map, or a timeline.

Example application: Foodborne outbreaks

Use social media comments to detect foodborne outbreaks.

Example application: Foodborne outbreaks

Use social media comments to detect foodborne outbreaks.

- $ate\_at(person, restaurant)$
- $has\_symptom(person, symptom)$
- True story: a colaboration between Yelp, Columbia University and NYC's Department of Health and Mental Hygiene.

More example applications

- Starting from business news build a network graph with deals, investments, acquisitions.
- With sanitary documents, build a timeline of diseases, symptoms and treatments.
- Reconstructing state terrorism victims' fate from military documents.
- Use social media to measure the growth of statup incubators and co-working spaces.
- . . .

What is not IE

What is not IE

Is not relation discovery, relations have to be known a-priori.

What is not IE

Is not relation discovery, relations have to be known a-priori.

Is not open domain, IE is domain specific.

A sketch of an IE system

Steps

- Basic natural language processing
- Entity recognition and linking
- Relation extraction

A sketch of an IE system

Steps

- Basic natural language processing
- Entity recognition and linking
    - Named entity recognition
    - Coreference resolution
    - Name linking
- Relation extraction

A sketch of an IE system

Steps

- Basic natural language processing
- Entity recognition and linking
  - Named entity recognition
  - Coreference resolution
  - Name linking
- Relation extraction
  - Rule-based
  - Corpus-based

Relation extraction

Rule-based

A regular expression to detect the presence of a relation.

- Handcrafted by an expert
- Suggested and validated by an algorithm

Corpus-based

For each pair of entities, a binary classification:

- Classifiers: Naive bayes, support vector machines, etc.
- Language models: Markov models, conditional random fields, etc.

Knowledge Bases and Reasoning

Knowledge bases

- Freebase
- Linked open data

Knowledge Bases and Reasoning

Knowledge bases

- Freebase
- Linked open data

Reasoning

Semantic reasoners, they exist.

IE summary

- Getting information out of large amounts of text documents.
- Extract **relation** instances between **entities**.
- Has many practical applications.
- Domain-specific.
- Solved using rules or statistic tools.

IE with IEPY

What is IEPY

- Tool for Information Extraction.
- Open source (BSD).
- Written in Python.
- Developed by Machinalis.
- Aimed at:
  - IE users
  - IE scientists
- `https://github.com/machinalis/iepy`

Features

For IE users

- NER and coreference resolution.
- **Active learning** relation extraction.
- **Rule-based relation extraction**.

For IE scientists

- **Corpus annotation** tool.
- Web-based UI.
- Easily hackable.

IEPY: Active learning core

Statistical classifier

- Support Vector Machine (by default).
- Features: Bag-of-words, POS tags, entity distance, etc.
- Active learning.
- Tunable to high-precision or high-recall.
- Web-based UI to interact with the expert.
- Easily hackable (`scikit-klearn` and `django`)

IEPY: Active learning core

Active learning goal

Minimize human effort

IEPY: Active learning core

Active learning goal

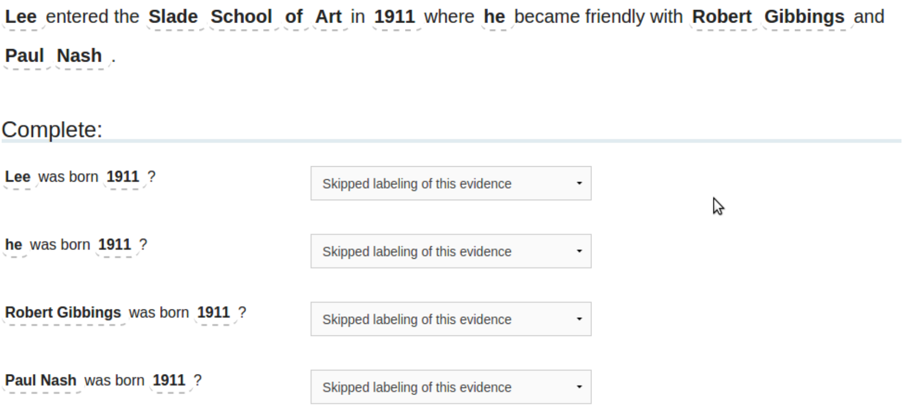
Minimize human effort

Idea

- Query a human expert.
- Use the answer to compute the next most useful questions to ask.
- Repeat.

Uses less human time to achieve the same performance.

IEPY: Active learning core



Lee entered the **Slade** **School** **of** **Art** in **1911** where **he** became friendly with **Robert** **Gibbings** and **Paul** **Nash** .

Complete:

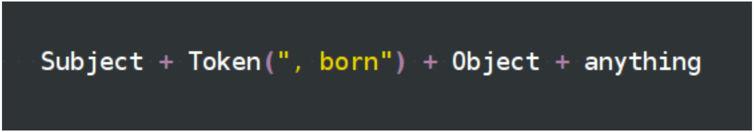
Lee was born **1911** ?	Skipped labeling of this evidence
he was born **1911** ?	Skipped labeling of this evidence
Robert Gibbings was born **1911** ?	Skipped labeling of this evidence
Paul Nash was born **1911** ?	Skipped labeling of this evidence

IEPY: Rule core

Rule-based

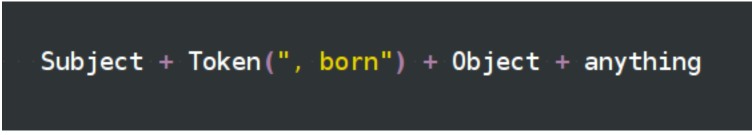
- Enhanced regular expressions.
- Expresiveness: Words, POS tags, entities, entity kinds, etc.
- Positive or negative rules.
- Have to be written in Python.
- Simplified syntax

Rule-based



```
Subject + Token(", born") + Object + anything
```

Rule-based

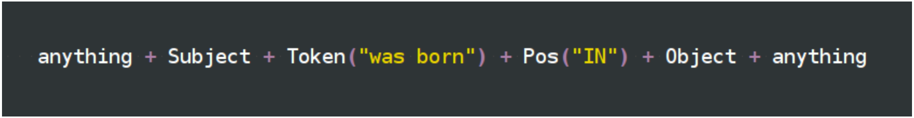


```
Subject + Token(", born") + Object + anything
```

Matches something like

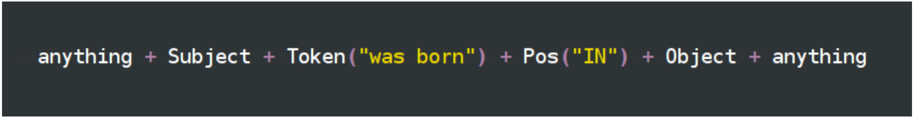
Lyle Eugene Hollister, born **6 July 1923** in Sioux Falls...

Rule-based



```
anything + Subject + Token("was born") + Pos("IN") + Object + anything
```

Rule-based



```
anything + Subject + Token("was born") + Pos("IN") + Object + anything
```

Matches something like

. . . **Shamsher M. Chowdhury** was born in **1950** . . .

Corpora construction

Evaluation?

Eventually you'll need to evaluate your IE performance.

Corpora construction

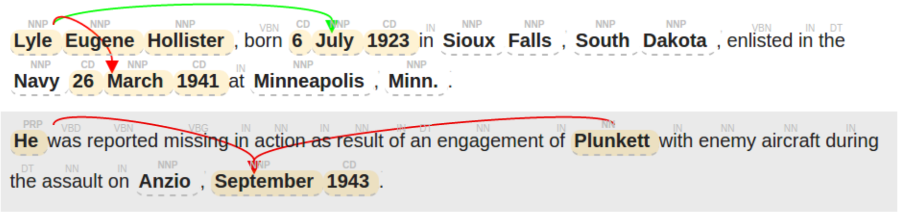
Evaluation?

Eventually you'll need to evaluate your IE performance.

- How much noise has the information you extracted?
- How many relevant facts you are leaving behind?

Corpora construction

IEPY has a corpora construction tool



Lyle Eugene Hollister , born 6 July 1923 in Sioux Falls , South Dakota , enlisted in the Navy 26 March 1941 at Minneapolis , Minn. .

He was reported missing in action as result of an engagement of Plunkett with enemy aircraft during the assault on Anzio , September 1943 .

NNP NNP NNP VBN CD NNP CD IN NNP NNP NNP NNP VBN IN DT
NNP CD NNP CD IN NNP NNP
PRP VBD VBN VBG IN NN IN NN IN DT NN IN NN IN NN NN IN
DT NN IN NNP NNP CD

Since is a web-based it allows multiple experts to work simultaneously

Some performance figures

Active learning

On an easy relation (date of birth)

86% prec. and 80% rec.

with an hour's worth of labeling.

Rule based

On an easy relation (date of birth)

98.65% prec. and 38% rec.

investing 4 hours (11 rules).

Corpora construction

Rate of 50 documents per hour (we labeled 4300 docs).

The end

That's it, thank you!