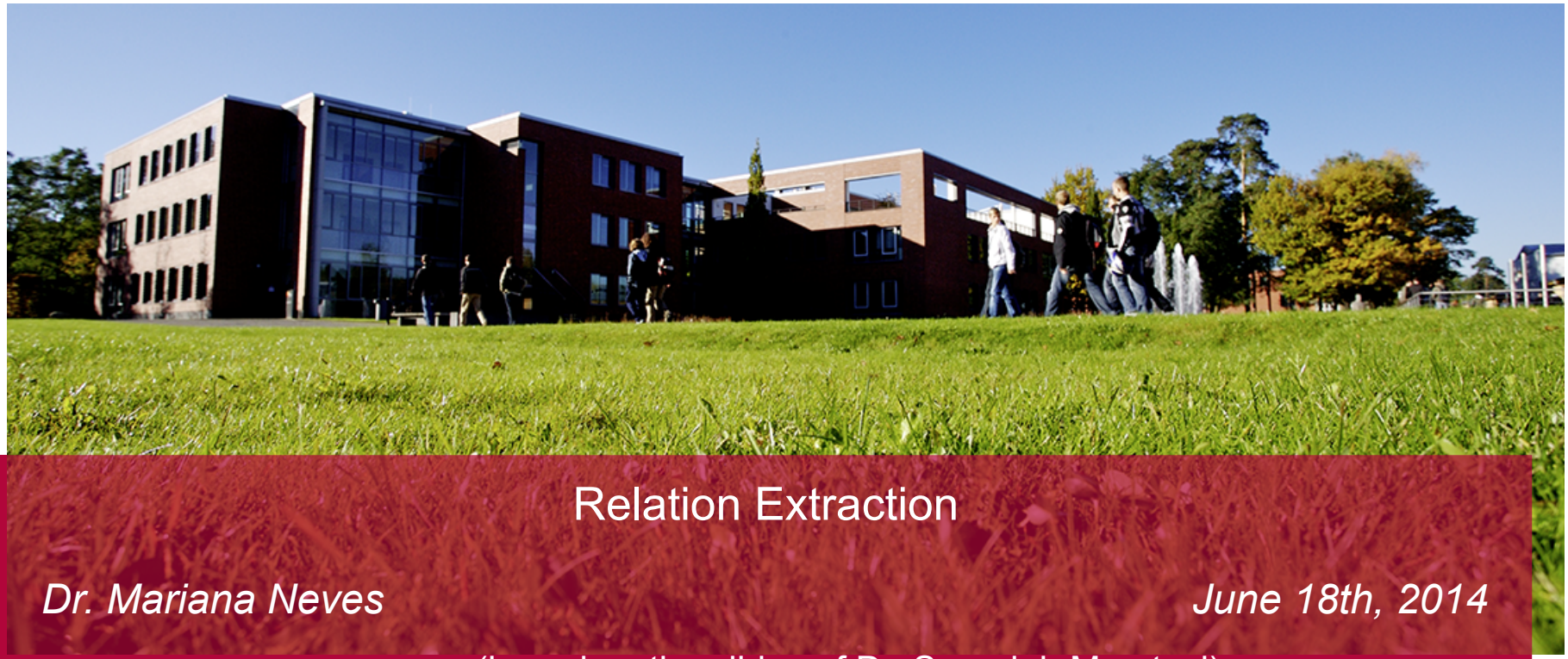
Natural Language Processing
SoSe 2014

HPI Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam



Relation Extraction

Dr. Mariana Neves

June 18th, 2014

(based on the slides of Dr. Saeedeh Momtazi)

Outline

- Introduction
- Task
- Pattern Extraction
- Supervised Learning
- Semi-supervised Learning

Outline

- Introduction
- Task
- Pattern Extraction
- Supervised Learning
- Semi-supervised Learning

Information Extraction

- Named entity recognition
- Relation Extraction

Named Entity Recognition

- HPI is affiliated to the Potsdam University and located in Potsdam near Berlin. It was founded in 1998 by Hasso Plattner, one of the co-founders of the European software company, SAP AG.
  - HPI (ORG)
  - Potsdam University (ORG)
  - Potsdam (LOC)
  - Berlin (LOC)
  - 1998 (DATE)
  - Hasso Plattner (PER)
  - SAP AG (ORG)

Relation Extraction

- HPI is affiliated to the Potsdam University and located in Potsdam near Berlin. It was founded in 1998 by Hasso Plattner, one of the co-founders of the European software company, SAP AG.
  - HPI – Potsdam: located (ORG-LOC)
  - HPI – Berlin: near (ORG-LOC)
  - Potsdam – Berlin: near (LOC-LOC)
  - HPI – 1998: founded (ORG-DATE)
  - HPI - Hasso Plattner: founder (ORG-PER)
  - SAP AG - Hasso Plattner: co-founder (ORG-PER)

Information Extraction

Potsdam

View of the rebuilt Potsdam City Palace with St. Nicholas' Church in the background

Flag

Coat of arms

Coordinates: 52°24'N 13°4'E

Country	Germany
State	Brandenburg
District	Urban district
Government	
• **Lord Mayor**	Jann Jakobs (SPD)
Area	
• **Total**	187.28 km² (72.31 sq mi)
Elevation	32 m (105 ft)
Population (2012-12-31)[1]	
• **Total**	159,456
• **Density**	850/km² (2,200/sq mi)
Time zone	CET/CEST (UTC+1/+2)
Postal codes	14467–14482
Dialling codes	0331
Vehicle registration	P
Website	www.potsdam.de

Hasso Plattner Institute

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH

HPI Hasso Plattner Institut

IT Systems Engineering | Universität Potsdam

Motto	Design IT. Create Knowledge.
Established	1998[1]
Type	Private university institute
Director	Prof. Dr. Christoph Meinel
Admin. staff	60[2]
Students	about 480[2]
Location	Potsdam, Germany
Campus	Griebnitzsee
Colors	Orange Vivid orange Dark pink
Affiliations	University of Potsdam
Website	www.hpi.uni-potsdam.de

Motivation

- Creating new structured data sources (knowledge bases)
  - DBPedia
  - Freebase
  - Yago
- Answering complex questions using multiple sources
  - Which soccer player married a Spice Girls star?

    ("?x" is-a "soccer player")

    ("?x" married "?y")

    ("?y" member "Spice Girls")

Outline

- Introduction
- Task
- Pattern Extraction
- Supervised Learning
- Semi-supervised Learning

Relation Representation

- Representing data as triples
  - (Argument1 RelationType Argument2)
  - (Subject Predicate Object)

- Resource Description Framework (RDF)

Relation Types

- Having various relation types based on the type of arguments
  - PER-PER: Spouse, Parent, Child, Friendship, Colleague, ...
  - PER-LOC: Place of birth, Lives in, Place of death, Buried in, ...
  - PER-ORG: Founder, Co-founder, Owner, Employee, Student/Alum, Professor, ...
  - ORG-LOC: Located, Near, Founded-location, Headquarter, ...
  - PER-DATE: Date of Birth, Date of Marriage, Date of Death, ...

Approaches

- Manually created patterns
- Supervised machine learning
- Semi-supervised learning

Outline

- Introduction
- Task
- Pattern Extraction
- Supervised Learning
- Semi-supervised Learning

Pattern Extraction

- What are the potential words to express a relation type?
  - (PER Member ORG)
  - ("?x" Member "?y")

  - x is a member of y.
  - x is an employee of y.
  - x works at y.
  - x is a staff of y.

  - x is (a|an) (member|employee|staff) of y.
  - x (works) at y.

Pattern Extraction

- Advantages
  - Having high precision results

- Disadvantages
  - Having low recall
  - Finding all possible patterns is labor intensive
  - Covering all relations is very difficult

Outline

- Introduction
- Task
- Pattern Extraction
- Supervised Learning
- Semi-supervised Learning

Supervised Classification

- Training data
  - Defining a fix set of relation types
  - Choosing the corresponding named entities
  - Selecting a set of texts as training data
  - Recognizing the named entities in the text
  - Labeling the relations between named entities manually

Task

- Input
  - A pair of entities (NER)
  - A context in which this pair appears
  - Possible relation types
- Output
  - Type of relation between two entities, if there exist any

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“
  - PER-LOC (Thomas Edison, New Jersey)
  - Place of birth, Place of death, Buried in

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“
- The target entities
  - T1: Thomas Edison
  - T2: New Jersey
- Surrounding words of target entities
  - $T1_{+1}$: died
  - $T2_{-1}$: in
  - $T2_{+1}$: due
- All words between the target entities (bag-of-word)
  - 1931 October died 18 , on , in

Feature Selection

- The named entity label of the target words
  - NE(T1): PER
  - NE(T2): LOC
- The syntactic structure of the sentence
  - Shortest dependency path
    - nsubj-prep-pobj

```
  nn(Edison-2, Thomas-1)
1 nsubj(died-3, Edison-2)
  root(ROOT-0, died-3)
  prep(died-3, on-4)
  pobj(on-4, October-5)
  num(October-5, 18-6)
  num(October-5, 1931-8)
2 prep(died-3, in-10)
  nn(Jersey-12, New-11)
3 pobj(in-10, Jersey-12)
  amod(Jersey-12, due-13)
  prep(due-13, to-14)
  pobj(to-14, complications-15)
  prep(complications-15, of-16)
  pobj(of-16, diabetes-17)
```

http://nlp.stanford.edu:8080/parser/index.jsp

Classification Algorithm

- Applying any of the classifiers
  - K Nearest Neighbor
  - Support Vector Machines
  - Naïve Bayes
  - Maximum Entropy
  - Logistic Regression
  - ...

Supervised Classification

- Advantages
  - Very good performance if
    - Having enough training data
    - Having test data similar to training data
- Disadvantages
  - Manual labeling of training data is labor expensive
  - Difficult to get good results for other domains

Outline

- Introduction
- Task
- Pattern Extraction
- Supervised Learning
- Semi-supervised Learning

Semi-supervised Learning

- Having no large training data
  - but a large collection of documents
- Producing a small training data (seed data)
  - A set of triples
- Bootstrapping
  - Using the seed data to find further entity pairs with the same relation

Bootstrapping

- Using the collected seed data
- Finding sentences which contain at least one entity pair
- Extracting the common contexts of the pair
- Creating patterns (or models) from the extracted context
- Using the pattern (or model) to get more pairs and add them to seed data

Bootstrapping

- Using the collected seed data
  - (Thomas Edison Spouse Mina Mille)

Bootstrapping

- Finding sentences which contain at least one entity pair (normalized entities)

- Thomas Edison married Mina Mille.
- Edison married a young woman named Mina Mille.
- In 1871, Thomas Edison married Mina Mille.
- Thomas Edison marries Mina Mille on December 25.

Bootstrapping

- Extracting the common contexts of the pair
- Creating patterns (or models) from the extracted context

- Thomas Edison married Mina Mille.
- Edison married a young woman named Mina Mille.
- In 1871, Thomas Edison married Mina Mille.
- Thomas Edison marries Mina Mille on December 25.

Bootstrapping

- Using the pattern (or model) to get more pairs and add them to seed data
  - (Albert Einstein Spouse "?")

- Einstein marries his cousin Elsa Löwenthal on June 2.
- Einstein married Elsa Löwenthal in Berlin.
- Einstein married Elsa Löwenthal on 2 June 1919.
- After their divorce in 1919, Einstein married Elsa Löwenthal in the same year.
- Albert Einstein was married to Elsa Löwenthal for 17 years.
- Einstein marries Elsa Löwenthal.
- In the same year Albert Einstein married Elsa Löwenthal.

⇒ (Albert Eistein Spouse Elsa Löwenthal)

Bootstrapping

- Using the collected seed data
  - (Thomas Edison Spouse Mina Mille)
  - (Albert Eistein Spouse Elsa Löwenthal)

Bootstrapping

- Using the collected seed data
- Finding sentences which contain at least one entity pairs
- Extracting the common contexts of the pair
- Creating patterns (or models) from the extracted context

- Albert Einstein's wife, Elsa Löwenthal, was his first cousin.
- Elsa Löwenthal was the wife of Albert Einstein.
- Einstein's wife was named Elsa Löwenthal.

Further Reading

- Speech and Language Processing
  - Chapters 22.2