Natural Language Processing

SoSe 2017

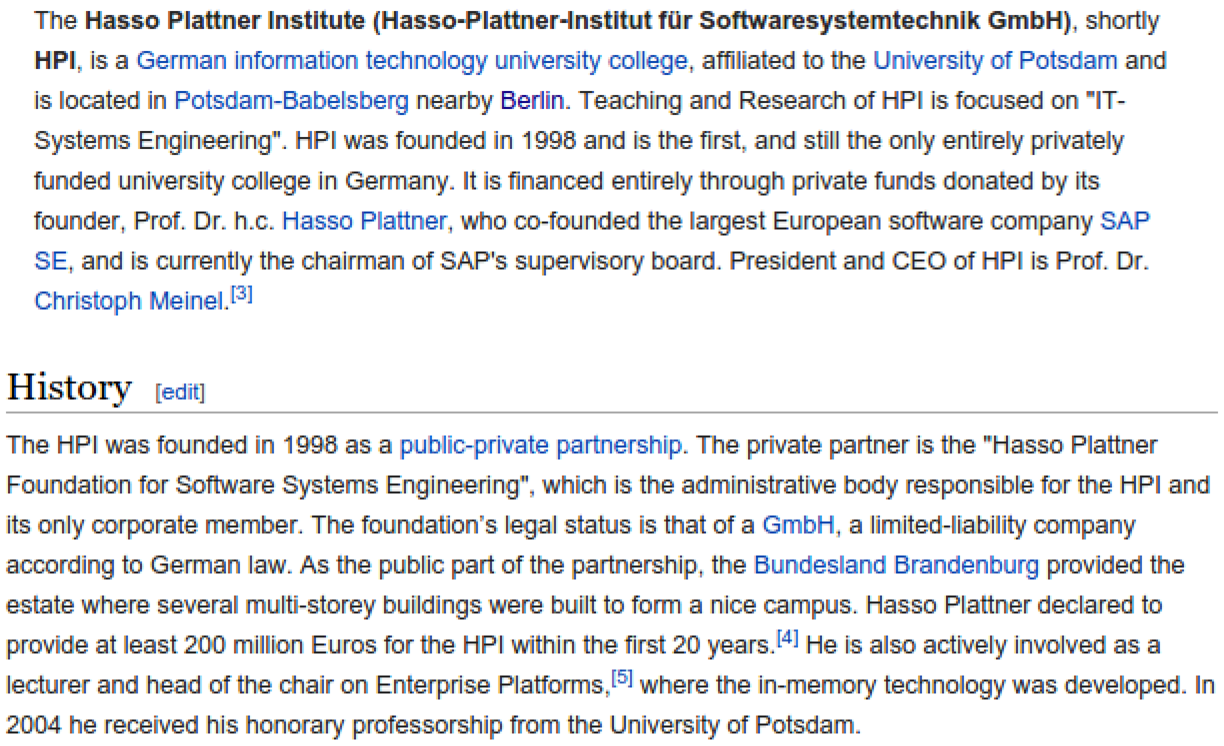
Hasso Plattner Institut
IT Systems Engineering | Universität Potsdam

Information Extraction

Dr. Mariana Neves

June 28th, 2017

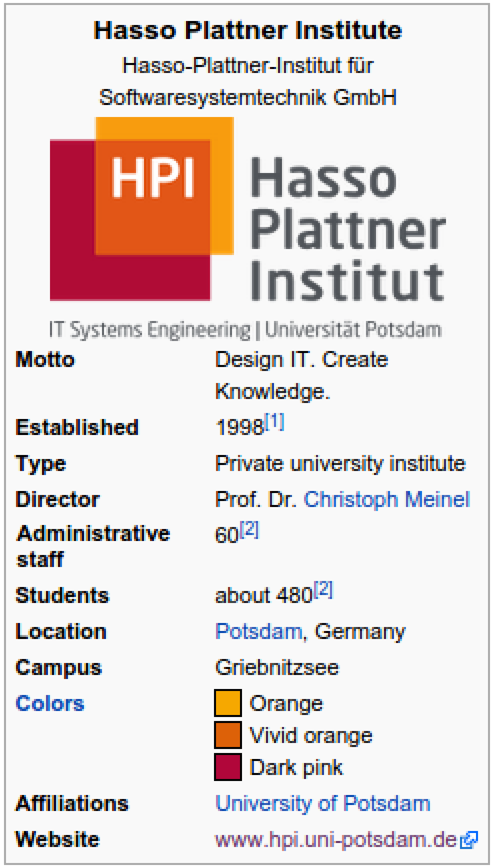
Information Extraction



The **Hasso Plattner Institute (Hasso-Plattner-Institut für Softwaresystemtechnik GmbH)**, shortly **HPI**, is a German information technology university college, affiliated to the University of Potsdam and is located in Potsdam-Babelsberg nearby Berlin. Teaching and Research of HPI is focused on "IT-Systems Engineering". HPI was founded in 1998 and is the first, and still the only entirely privately funded university college in Germany. It is financed entirely through private funds donated by its founder, Prof. Dr. h.c. Hasso Plattner, who co-founded the largest European software company SAP SE, and is currently the chairman of SAP's supervisory board. President and CEO of HPI is Prof. Dr. Christoph Meinel.[3]

History [edit]

The HPI was founded in 1998 as a public-private partnership. The private partner is the "Hasso Plattner Foundation for Software Systems Engineering", which is the administrative body responsible for the HPI and its only corporate member. The foundation's legal status is that of a GmbH, a limited-liability company according to German law. As the public part of the partnership, the Bundesland Brandenburg provided the estate where several multi-storey buildings were built to form a nice campus. Hasso Plattner declared to provide at least 200 million Euros for the HPI within the first 20 years.[4] He is also actively involved as a lecturer and head of the chair on Enterprise Platforms,[5] where the in-memory technology was developed. In 2004 he received his honorary professorship from the University of Potsdam.



Hasso Plattner Institute

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH

Motto	Design IT. Create Knowledge.
Established	1998[1]
Type	Private university institute
Director	Prof. Dr. Christoph Meinel
Administrative 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

(http://en.wikipedia.org/wiki/Hasso_Plattner_Institute)

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)

Motivation

- Creating new structured data sources (knowledge bases)
  - DBPedia
  - Freebase
  - Yago
  - Wikidata

Motivation

- 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")

Relation Representation

- Data can be represented as triples
  - (Argument1 RelationType Argument2)
  - (Subject Predicate Object)

("Messi" is-a "soccer player")

("Brad Pitt" married "Angelina Jolie")

("Messi" member "Barcelona FC")

Relation Types

- There are various relation types based on the type of arguments
  - PER-PER: Spouse, Parent, Child, Friendship, Colleague, …

("Brad Pitt" married "Angelina Jolie")

("Shiloh Nouvel Jolie-Pitt" child "Angelina Jolie")

("Messi" colleague "Neymar")

Relation Types

- There are various relation types based on the type of arguments
  - PER-LOC: Place of birth, Lives in, Place of death, Buried in, …

("Angela Merkel" place_of_birth "Hamburg")

("Angela Merkel" lives "Berlin")

("Beethoven" place_of_birth "Bonn")

("Beethoven" place_of_death "Vienna")

("Beethoven" buried "Vienna")

Approaches

- Manually created patterns
- Supervised machine 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|professor|researcher|lecturer) 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
  - Language is complex and creative

Supervised Classification

- Training data:
  - Define a set of relation types
  - Choose the corresponding named entities
  - Select a set of texts as training data
  - Recognize the named entities in the text
  - Label the relations between named entities manually

Classification Task

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

Classification Task

- „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

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- The (named entity) label of the target words (blind entities)
  - NE(T1): PER
  - NE(T2): LOC

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Bag-of-words
  - 1931 October died 18 , on , in
- Bag-of-bigrams
  - [1931 ,] [October 18] [died on] [18 ,] [, 1931] [on October] [, in]

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes."

- Bag-of-words, entities
  - YEAR MONTH died DATE , on , in
- Bag-of-bigrams, entities
  - [YEAR ,] [MONTH DATE] [died on] [DATE ,] [, YEAR] [on MONTH] [, in]

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Bag-of-words, entities, stems
  - YEAR MONTH die DATE , on , in
- Bag-of-bigrams, entities, stems
  - [YEAR ,] [MONTH DATE] [die on] [DATE ,] [, YEAR] [on MONTH] [, in]

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Distance in words between arguments
  - 6 words
  - 8 words (including punctuations)

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Number of entities between arguments
  - None?
  - Three (MONTH, DATE, YEAR)

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes."

- Surrounding words of target entities
  - For instance, [-1,+1]
    - $T1_{+1}$: died
    - $T2_{-1}$: in
    - $T2_{+1}$: due

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Bags of chunk heads
  - VP, PP, NP, NP

```
(ROOT
  (S
    (NP (NNP Thomas) (NNP Edison))
    (VP (VBD died)
      (PP (IN on)
        (NP (NNP October) (CD 18) (, ,) (CD 1931) (, ,)))
      (PP (IN in)
        (NP
          (NP (NNP New) (NNP Jersey))
          (ADJP (JJ due)
            (PP (TO to)
              (NP
                (NP (NNS complications))
                (PP (IN of)
                  (NP (NN diabetes))))))))
    (. .)))
```

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

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Chunk base-phrase paths
  - VP→PP→NP→NP
  - Trigrams
    - VP→PP→NP,

      PP→NP→NP

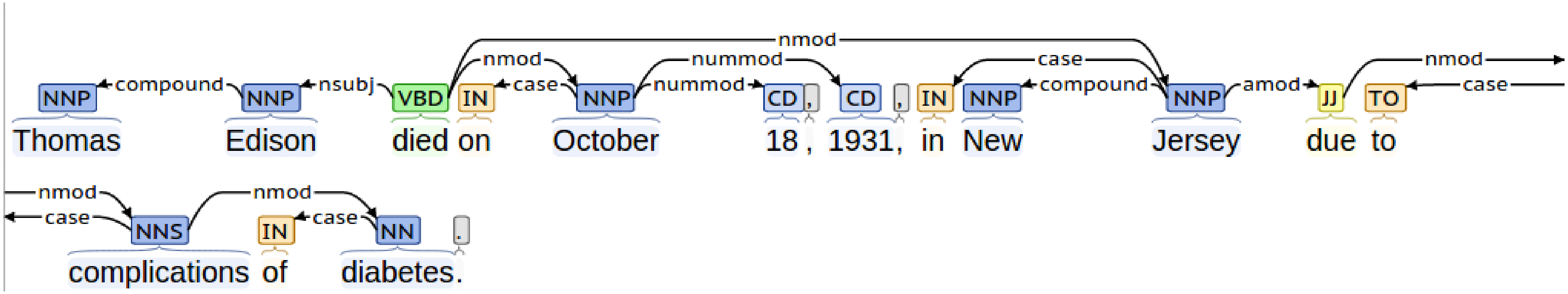
```
(ROOT
  (S
    (NP (NNP Thomas) (NNP Edison))
    (VP (VBD died)
      (PP (IN on)
        (NP (NNP October) (CD 18) (, ,) (CD 1931) (, ,)))
      (PP (IN in)
        (NP
          (NP (NNP New) (NNP Jersey))
          (ADJP (JJ due)
            (PP (TO to)
              (NP
                (NP (NNS complications))
                (PP (IN of)
                  (NP (NN diabetes)))))))))
    (. .)))
```

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

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Dependency-tree paths
  - nsubj-nmod
  - compound-nsubj-nmod-compound

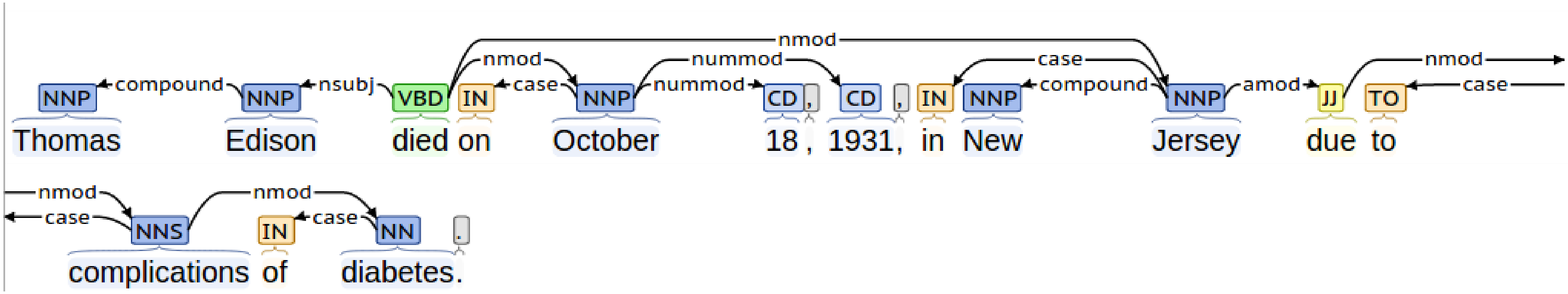


(http://nlp.stanford.edu:8080/corenlp/process)

Feature Selection

- „Thomas Edison died on October 18, 1931, in New Jersey due to complications of diabetes.“

- Tree distance between arguments
    - Two (nsubj-nmod)
    - Four (compound-nsubj-nmod-compound)



(http://nlp.stanford.edu:8080/corenlp/process)

Classification Algorithm

- Any of the usual ML classifiers
  - K Nearest Neighbor
  - Support Vector Machines
  - Naïve Bayes
  - Maximum Entropy
  - Logistic Regression
  - Neural Networks
  - …

Supervised Classification

- Advantages
  - Very good performance given
    - enough training data
    - test data similar to training data

- Disadvantages
  - Manual labeling of training data is labor expensive
  - Difficult to get good results for other domains and relations

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

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

Bootstrapping

- Use the collected seed data
  - (Thomas Edison Spouse Mina Mille)
  - (Brad Pitt Spouse Angelina Jolie)
  - ...

Bootstrapping

- Use the collected seed data
- Find sentences which contain at least one entity pair

- 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

- Use the collected seed data
- Find sentences which contain at least one entity pair
- Extract the common contexts of the pair
- Create 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

- Use the collected seed data
- Find sentences which contain at least one entity pair
- Extract the common contexts of the pair
- Create patterns (or models) from the extracted context
- Use the pattern (or model) to get more pairs and add them to seed data
  - (Albert Einstein Spouse “?”)

Bootstrapping

- 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

- Use the collected seed data (start over again)
  - (Thomas Edison Spouse Mina Mille)
  - (Brad Pitt Spouse Angelina Jolie)
  - …
  - (Albert Eistein Spouse Elsa Löwenthal)

Bootstrapping

- Use the collected seed data
- Find sentences which contain at least one entity pairs
- Extract the common contexts of the pair
- Create 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.

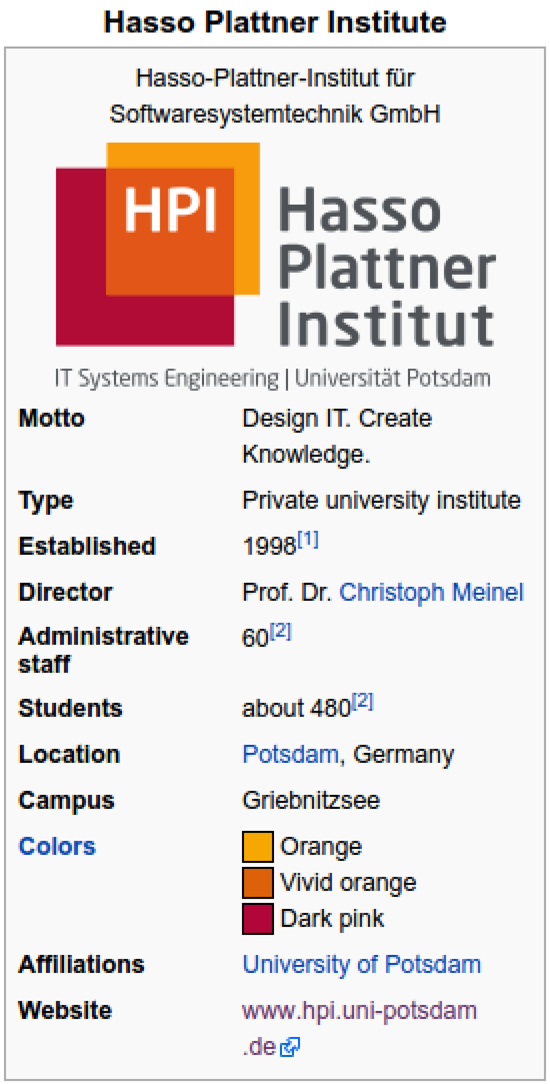
Semantic drift

- Erroneous patterns → introduction of erroneous tuples → problematics patterns

- Brad Pitt married the daughter of Jon Voigth

Template filling

- Template
  - slots



Hasso Plattner Institute

Hasso-Plattner-Institut für Softwaresystemtechnik GmbH

IT Systems Engineering | Universität Potsdam

Motto	Design IT. Create Knowledge.
Type	Private university institute
Established	1998[1]
Director	Prof. Dr. Christoph Meinel
Administrative 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

Statistical template filling

- Train separate classifiers, one for each slot

Name	INSTITUTION	Hasso-Plater Institute
Year foundation	YEAR	1998
Director	PEOPLE	Prof. Meinel
Location	CITY COUNTRY	Potsdam Germany
Affiliation	INSTITUTION UNIVERSITY	University of Potsdam

Statistical template filling

- Train separate classifiers, one for each slot
- Challenges
  - Multiple text segments labeled with the same slot label
    - Christoph Meinel, Prof. Meinel

Statistical template filling

- Train separate classifiers, one for each slot
- Challenges
  - Multiple entities of the expected type for a given slot
    - Potsdam, Germany, Berlin, Haifa, etc.
    - University of Potsdam, Stanford University, Cape Town University, Nanjing University, etc.

Statistical template filling

- Train one large classifier, usually Hidden Markov Model
  - Sequential labeling
  - Potsdam, Berlin, Germany (location) → University of Potsdam (university)

Summary

- Information extraction
  - Relation extraction
  - Slot filling
- Very task-specific and domain adaptation is difficult
- Approaches
  - Rule-based
  - (Semi-)Supervised learning

Further Reading

- „Speech and Language Processing“ book
  - Chapter 22