Pig for Natural Language Processing

Max Jakob

neofonie*

Agenda

1. Introduction (speaker, affiliation, project)
2. Named Entities
3. pignlproc

Speaker: Max Jakob

- MSc in Computational Linguistics
- Software Developer at Neofonie GmbH
  - Dicode project
  - Text mining
  - Scalability
- Past year
  - DBpedia extraction framework
  - DBpedia Spotlight

Neofonie

neofonie*

- Spin-off out of TU Berlin (1998)
- 160 employees in Berlin (head quarters) and Hamburg
- State-of-the-art Search, Online Portals, Mobile Apps

- R&D: 14 successfully finished research projects
- Current focus:
  - Semantics
  - Question Answering
  - Recommendations
  - Cloud Computing

Alexandria

SPIGA

serum

dicode

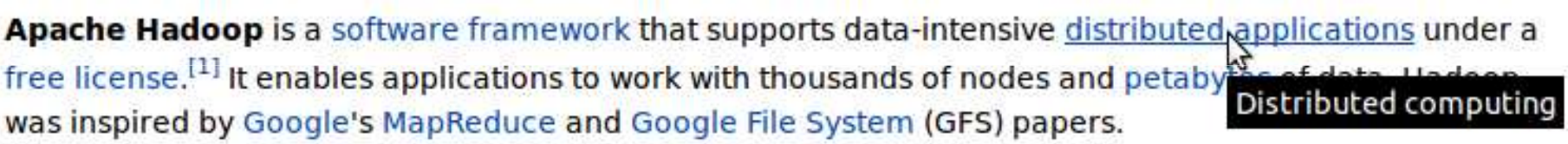
mia

Dicode

- EU-funded project (FP7)
- Goal: ``Augment collaboration and decision making in data-intensive and cognitively-complex settings.´´
  - Build scalable services for data mining and collaboration
- Example Use Case: **Social Media Monitoring**
  - Sources: blogs, news, Twitter, etc.
  - What are the key trends?
  - How is my brand perceived on the web?
- Problem: ambiguity of names and concepts
  - Need *Named Entity Disambiguation*

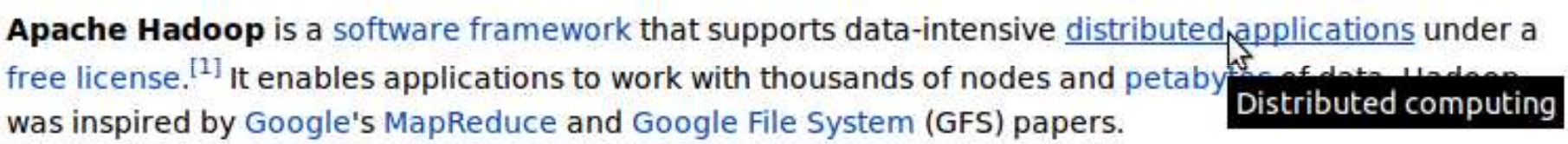
Named Entity Recognition/Disambiguation

- By example:
  - Input: plain text
  - Output: text with Wikipedia links



Named Entity Recognition/Disambiguation

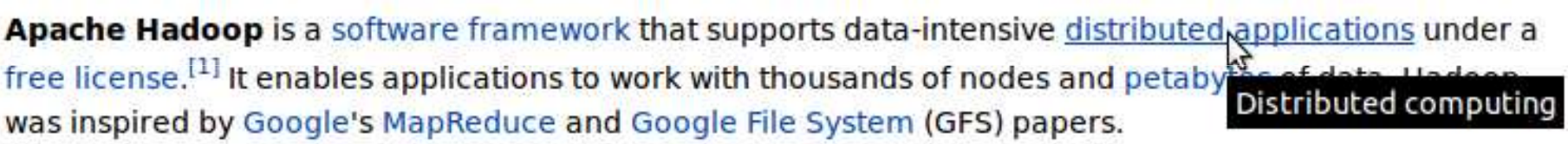
- By example:
  - Input: plain text
  - Output: text with Wikipedia links



- 1. Find „interesting“ strings (recognition)
  - Surface forms

Named Entity Recognition/Disambiguation

- By example:
  - Input: plain text
  - Output: text with Wikipedia links



- 1. Find „interesting“ strings (recognition)
  - Surface forms
- 2. Choose appropriate Wikipedia page (disambiguation)
  - Each Wikipedia page represents an entity
  - Every surface form can have multiple candidate entities for linking

„Michael Jackson died in 2007.“

„Michael Jackson died in 2007.“

- Named Entity Recognition
  - Find surface forms
  - [**Michael Jackson**] died in 2007.

„Michael Jackson died in 2007.“

- Named Entity Recognition
  - Find surface forms
  - [**Michael Jackson**] died in 2007.

- Named Entity Disambiguation
  - Choose entity from candidates
  - [**Michael Jackson**]

„Michael Jackson died in 2007.“

- Named Entity Recognition
  - Find surface forms
  - [**Michael Jackson**] died in 2007.

- Named Entity Disambiguation
  - Choose entity from candidates
  - [**Michael Jackson**]
    - Michael Jackson (singer)

neofonie*

„Michael Jackson died in 2007.“

- Named Entity Recognition
  - Find surface forms
  - [**Michael Jackson**] died in 2007.

- Named Entity Disambiguation
  - Choose entity from candidates
  - [**Michael Jackson**]
    - Michael Jackson (singer)
    - Michael Jackson (writer)

Michael Jackson

Michael Jackson

Born	Michael, James Jackson 27 March 1942 Wetherby, West Yorkshire
Died	30 August 2007 (aged 65) London
Nationality	British
Known for	Beer and whisky reviewing and journalism

Michael Jackson

Jackson at the White House in 1984

Background information

Birth name	Michael Joseph Jackson[1]
Also known as	Michael Joe Jackson
Born	August 29, 1958 Gary, Indiana, U.S.
Died	June 25, 2009 (aged 50) Los Angeles, California, U.S.

„Michael Jackson died in 2007.“

- Named Entity Recognition
  - Find surface forms
  - [**Michael Jackson**] died in 2007.

- Named Entity Disambiguation
  - Choose entity from candidates
  - [**Michael Jackson**]
    - Michael Jackson (singer)
    - Michael Jackson (writer)
  - *Context*: died in 2007

Michael Jackson

Michael Jackson

Born	Michael, James Jackson 27 March 1942 Wetherby, West Yorkshire
Died	30 August 2007 (aged 65) London
Nationality	British
Known for	Beer and whisky reviewing and journalism

Michael Jackson

Jackson at the White House in 1984

Background information

Birth name	Michael Joseph Jackson[1]
Also known as	Michael Joe Jackson
Born	August 29, 1958 Gary, Indiana, U.S.
Died	June 25, 2009 (aged 50) Los Angeles, California, U.S.

neofonie*

„Michael Jackson died in 2007.“

- Named Entity Recognition
  - Find surface forms
  - [**Michael Jackson**] died in 2007.

- Named Entity Disambiguation
  - Choose entity from candidates
  - [**Michael Jackson**]
    - Michael Jackson (singer)
    - **Michael Jackson (writer)**
  - *Context*: died in 2007

Michael Jackson

Michael Jackson

Born	Michael, James Jackson 27 March 1942 Wakefield, West Yorkshire
Died	30 August 2007 (aged 65) London
Nationality	British
Known for	Beer and whisky reviewing and journalism

Michael Jackson

Jackson at the White House in 1984

Background information

Birth name	Michael Joseph Jackson[1]
Also known as	Michael Joe Jackson
Born	August 29, 1958 Gary, Indiana, U.S.
Died	June 25, 2009 (aged 50) Los Angeles, California, U.S.

neofonie*

„Michael Jackson died in 2007.“

- Named Entity Recognition
  - Find surface forms
  - [**Michael Jackson**] died in 2007.

- Named Entity Disambiguation
  - Choose entity from candidates
  - [**Michael Jackson**]
    - Michael Jackson (singer)
    - **Michael Jackson (writer)**
  - *Context*: died in 2007
  - Context may not be distinctive

Michael Jackson

Michael Jackson

Born	Michael, James Jackson 27 March 1942 Wetherby, West Yorkshire
Died	30 August 2007 (aged 65) London
Nationality	British
Known for	Beer and whisky reviewing and journalism

Michael Jackson

Jackson at the White House in 1984

Background information

Birth name	Michael Joseph Jackson[1]
Also known as	Michael Joe Jackson
Born	August 29, 1958 Gary, Indiana, U.S.
Died	June 25, 2009 (aged 50) Los Angeles, California, U.S.

neofonie*

Probabilities

- P(entity | surface form)
  - Which entity is typically meant by a name?
  - For example, given [**Michael Jackson**] (and ignoring the context), what are the probabilities of the candidates?
    - Michael Jackson (singer) 0.75
    - Michael Jackson (writer) 0.25

Probabilities

- P(entity | surface form)
  - Which entity is typically meant by a name?
  - For example, given [**Michael Jackson**] (and ignoring the context), what are the probabilities of the candidates?
    - Michael Jackson (singer) 0.75
    - Michael Jackson (writer) 0.25
- Other useful probabilities:
  - P(surface form | entity), P(entity), P(surface form)

Probabilities

- P(entity | surface form)
  - Which entity is typically meant by a name?
  - For example, given [**Michael Jackson**] (and ignoring the context), what are the probabilities of the candidates?
    - Michael Jackson (singer) 0.75
    - Michael Jackson (writer) 0.25
- Other useful probabilities:
  - P(surface form | entity), P(entity), P(surface form)
- Estimate using Wikipedia page links
  - `In 1994 the beer journalist [[Michael Jackson (writer)|Michael Jackson]] described Webster's beers as "light" and "faintly oily".`

Previous method

- Sequential processing on one machine
- Process includes
  - Parsing the Wikipedia articles
  - Resolving redirects
  - Tokenizing and counting 3-grams*
  - Aggregating counts

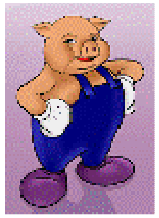
* Word sequences with length <= 3

Previous method

- Sequential processing on one machine
- Process includes
  - Parsing the Wikipedia articles
  - Resolving redirects
  - Tokenizing and counting 3-grams*
  - Aggregating counts
- Extremely long runtime: more than 1 week ☹
  - Tedious update process
    - Hard to improve pipeline
    - New concepts do not have probabilities
  - Going beyond 3-grams seems impossible

* Word sequences with length <= 3

Apache Pig



- Framework for analyzing large datasets on top of Apache Hadoop
- High-level scripting language **PigLatin**
  - Data-flow language
  - Think in tuples, bags and maps
  - load, filter, join, group by, store, …
- Pig analyzes the PigLatin script and derives a MapReduce plan
  - No need to dive deep into MapReduce
  - High development productivity
  - Automatic optimizations
- Simple interface for *user defined functions* (UDF)

pignlproc

- Open source project started by Olivier Grisel
  - Pig-Loader for Wikipedia articles
  - Several UDFs, e.g. to extract links
  - Example scripts, e.g. to build a training corpus for Named Entity Recognition in OpenNLP format

- Our Extensions:
  - Loader: Parse Wikipedia page ID
  - UDF: Resolve redirects
  - UDF: N-gram generator
  - PigLatin script for probability estimation

Probability estimation

- $P(\text{entity} \mid \text{surface form}) = \dfrac{\text{count}(\text{surface form, entity})}{\text{count}(\text{surface form})}$

Probability estimation

- $P(\text{entity} \mid \text{surface form}) = \dfrac{\text{count}(\text{surface form, entity})}{\text{count}(\text{surface form})}$

- count(Michael Jackson) = 4

Probability estimation

- $P(\text{entity} \mid \text{surface form}) = \frac{\text{count}(\text{surface form, entity})}{\text{count}(\text{surface form})}$

- count(Michael Jackson) = 4
- count(Michael Jackson, Michael Jackson (singer)) = 3
- count(Michael Jackson, Michael Jackson (writer)) = 1

Probability estimation

- $P(\text{entity} \mid \text{surface form}) = \frac{\text{count}(\text{surface form, entity})}{\text{count}(\text{surface form})}$

- count(Michael Jackson) = 4
- count(Michael Jackson, Michael Jackson (singer)) = 3
- count(Michael Jackson, Michael Jackson (writer)) = 1

- P(Michael Jackson (singer) | Michael Jackson) = ¾ = 0.75
- P(Michael Jackson (writer) | Michael Jackson) = ¼ = 0.25

Probability estimation

- $P(\text{entity} \mid \text{surface form}) = \dfrac{\text{count}(\text{surface form, entity})}{\text{count}(\text{surface form})}$

- count(Michael Jackson) = 4
- count(Michael Jackson, Michael Jackson (singer)) = 3
- count(Michael Jackson, Michael Jackson (writer)) = 1

- P(Michael Jackson (singer) | Michael Jackson) = ¾ = 0.75
- P(Michael Jackson (writer) | Michael Jackson) = ¼ = 0.25

- Check the project web for estimation of other probabilities

PigLatin code example (1)

```
parsed = LOAD 'enwiki-20111207-pages-articles.xml',
    USING pignlproc.storage.ParsingWikipediaLoader('en')
    AS (title, id, pageUrl, text, redirect, links, headers, paragraphs);
```

… more pignlproc magic …

PigLatin code example (1)

```
parsed = LOAD 'enwiki-20111207-pages-articles.xml',
    USING pignlproc.storage.ParsingWikipediaLoader('en')
    AS (title, id, pageUrl, text, redirect, links, headers, paragraphs);

… more pignlproc magic …

DESCRIBE pageLinks;
pageLinks: {
    surfaceForm: chararray,
    entity: chararray
}
```


PigLatin code example (1)

```
parsed = LOAD 'enwiki-20111207-pages-articles.xml',
    USING pignlproc.storage.ParsingWikipediaLoader('en')
    AS (title, id, pageUrl, text, redirect, links, headers, paragraphs);
```

… more pignlproc magic …

```
DESCRIBE pageLinks;
pageLinks: {
    surfaceForm: chararray,
    entity: chararray
}
```

- Bag of tuples, e.g. {
  (Micheal Jackson, Michael Jackson (singer)),
  (Micheal Jackson, Michael Jackson (singer)),
  (Micheal Jackson, Michael Jackson (writer)),
  (King of Pop, Michael Jackson (singer)), … }

PigLatin code example (2)

```
groupedBySurfaceForms = GROUP pageLinks BY surfaceForm;
```


PigLatin code example (2)

```
groupedBySurfaceForms = GROUP pageLinks BY surfaceForm;
surfaceFormCounts = FOREACH groupedBySurfaceForms GENERATE
    group AS surfaceForm,
    COUNT(pageLinks) AS surfaceFormCount;
```


PigLatin code example (2)

```
groupedBySurfaceForms = GROUP pageLinks BY surfaceForm;
surfaceFormCounts = FOREACH groupedBySurfaceForms GENERATE
     group AS surfaceForm,
     COUNT(pageLinks) AS surfaceFormCount;

DESCRIBE surfaceFormCounts ;
surfaceFormCounts : {
     surfaceForm : chararray,
     surfaceFormCount: long
}
```


PigLatin code example (2)

```
groupedBySurfaceForms = GROUP pageLinks BY surfaceForm;
surfaceFormCounts = FOREACH groupedBySurfaceForms GENERATE
    group AS surfaceForm,
    COUNT(pageLinks) AS surfaceFormCount;

DESCRIBE surfaceFormCounts ;
surfaceFormCounts : {
    surfaceForm : chararray,
    surfaceFormCount: long
}
```

- Bag of tuples, e.g. {
  (Micheal Jackson, 4),
  (King of Pop, 1),
  … }

PigLatin code example (3)

```
groupedByPairs = GROUP pageLinks BY (surfaceForm, entity);
```


PigLatin code example (3)

```
groupedByPairs = GROUP pageLinks BY (surfaceForm, entity);
pairCounts = FOREACH groupedByPairs GENERATE
    group AS pair,
    COUNT(pageLinks) AS pairCount;
```


PigLatin code example (3)

```
groupedByPairs = GROUP pageLinks BY (surfaceForm, entity);
pairCounts = FOREACH groupedByPairs GENERATE
     group AS pair,
     COUNT(pageLinks) AS pairCount;

DESCRIBE pairCounts ;
pairCounts : {
     pair: (chararray, chararray),
     pairCount: long
}
```


PigLatin code example (3)

```
groupedByPairs = GROUP pageLinks BY (surfaceForm, entity);
pairCounts = FOREACH groupedByPairs GENERATE
     group AS pair,
     COUNT(pageLinks) AS pairCount;

DESCRIBE pairCounts ;
pairCounts : {
     pair: (chararray, chararray),
     pairCount: long
}
```

- Bag of tuples, e.g. {
  ((Micheal Jackson, Michael Jackson (singer)), 3),
  ((Micheal Jackson, Michael Jackson (writer)), 1),
  ((King of Pop, Michael Jackson (singer)), 1), … }

PigLatin code example (4)

```
joined = JOIN
    surfaceFormCounts BY surfaceForm,
    pairCounts BY pageLinks::surfaceForm;
```


PigLatin code example (4)

```
joined = JOIN
    surfaceFormCounts BY surfaceForm,
    pairCounts BY pageLinks::surfaceForm;

probEntityGivenSf = FOREACH joined GENERATE
    surfaceForm,
    pairCount/surfaceFormCount,
    pairUri;
```

PigLatin code example (4)

```
joined = JOIN
    surfaceFormCounts BY surfaceForm,
    pairCounts BY pageLinks::surfaceForm;

probEntityGivenSf = FOREACH joined GENERATE
    surfaceForm,
    pairCount/surfaceFormCount,
    pairUri;
```

- Bag of tuples, e.g. {
  (Micheal Jackson, ¾, Michael Jackson (singer)),
  (Micheal Jackson, ¼, Michael Jackson (writer)),
  (Jacko, 1, Michael Jackson (singer)), … }

Runtimes for probability estimation

- Runtime single-threaded
  - 3-grams: > 1 week

Runtimes for probability estimation

- Runtime single-threaded
  - 3-grams: > 1 week
- Hadoop Cluster
  - 3 nodes, 2 hexacores each, hyper-threaded

Runtimes for probability estimation

- Runtime single-threaded
  - 3-grams: > 1 week
- Hadoop Cluster
  - 3 nodes, 2 hexacores each, hyper-threaded
- Runtime pignlproc
  - 3-grams: 1.5 hours

Runtimes for probability estimation

- Runtime single-threaded
  - 3-grams: > 1 week
- Hadoop Cluster
  - 3 nodes, 2 hexacores each, hyper-threaded
- Runtime pignlproc
  - 3-grams: 1.5 hours
  - 5-grams: 2.5 hours
    - „Karl Theodor zu Guttenberg“, „Ursula von der Leyen“

Dicode	http://dicode-project.eu
Pig	http://pig.apache.org
pignlproc	https://github.com/dicode-project/pignlproc
DBpedia Spotlight	http://spotlight.dbpedia.org

Jobs at Neofonie http://www.neofonie.de/karriere/jobs

neofonie*

Max Jakob
Software Developer

Neofonie GmbH
Robert-Koch-Platz 4
10115 Berlin
Germany

T +49.30 24627 - 290
F +49.30 24627 120
max.jakob@neofonie.de
www.neofonie.de