Information Extraction
Seminar WS 2013 / 2014

Session 1, Wednesday October 23, 2013
(Introduction, Organization, Topic Assignment)

Sabine Storandt, Elmar Haußmann

Chair for Algorithms and Data Structures
Department of Computer Science
University of Freiburg

Topic of this Seminar (1/2)

- Information Extraction (IE)
  - "automatic extraction of information from unstructured sources"
  - For example:
    - Named Entity Recognition
    - Relationship Extraction (RE)
      - Typically triples: (subject) (predicate) (object)
    - Open Information Extraction (OpenIE)
      - Extraction of triples with arbitrary predicate

Topic of this Seminar (2/2)

- IE is needed by lots of applications
  - E.g. semantic full-text search (Broccoli) or Question Answering
- This seminar in more detail
  - Information Extraction is rooted in Natural Language Processing (NLP) and often utilizes Machine Learning (ML)
    - We will need to learn some basic NLP an ML in order to understand the approaches to IE
  - Based on this we try to understand approaches for RE and OpenIE
    - Starting with early systems (2007/2008)
    - Up to very recent systems

Organization of this Seminar

- One or two presentations per session
  - Today: introduction of topics + topic assignment
  - After today, we have 13 sessions left
  - First talk is on November 13 (three weeks from now)
  - Time schedule for your presentation: next slide

Time Schedule for your Presentation

- >=3 weeks before your presentation
  - Start reading material and make a plan of what you want to talk about
- 2 weeks before your presentation
  - Meet with us (Sabine + Elmar) and present your plan
  - **Please do not waste our time** by coming unprepared
  - In the week that follows, work out all the necessary details and play around (extensively) with software
  - Prepare an outline of your presentation
- 1 week before your presentation
  - Meet with us again, and present your findings and the outline of your presentation (tentative slides)
  - In the week that follows, finish your work and the presentation

Challenges

- Research
  - You have to collect yourself interesting and relevant material, we provide the general topic and give you an initial paper
- Understand
  - Get a decent overview of your topic and understand what you will be talking about
- Presentation
  - Present your material in an interesting manner, don't forget that you have an audience
- Insight
  - Understand the advantages and disadvantages of the approach, what works well and where are problems
  - If code or a demo is available, check for efficiency and quality

Your Presentation

- Guidelines
  - You have 30 minutes for your talk + discussion
  - Use slides in PPT or PDF
  - Your talk will be recorded

Our Role

- We will help you, don't worry
  - In your first meeting with us (two weeks before your talk) we will help you focus on a good selection of material for your talk
  - We will also give you feedback and advice on the structure and contents of your slides
  - And, of course, we try to help when you have problems understanding something
  - However, the **initiative** has to come from you !

Evaluation

- Feedback
  - Anonymous from the audience
  - Final grade at the end of the seminar
  - Grade consists of two parts:
    - understanding of the paper/system
    - presentation (slides, time management ...)

List of Topics 1/2

UNI FREIBURG

NLP Basics

1. Shallow NLP (POS tagging, noun phrase chunking, named entity recognition, entity linking) (Code)
2. Deep NLP (constituent / dependency parsing, semantic role labeling) (Code)

Relation Extraction

1. Distant supervision for RE without labeled data
2. Relation Extraction With Matrix Factorization (Code)

Open Information Extraction (shallow NLP)

1. TextRunner + KnowItAll
2. ReVerb + R2A2 (Code)

List of Topics 2/2

Open Information Extraction (deep NLP)

1. Open IE using Wikipedia
2. SRL-IE
3. OLLIE (Code)
4. Dependency based OpenIE (Code)
5. ClausIE (Code)
6. Integrating Syntactic and Semantic Analysis into Open IE

Systems using IE / Tasks related to IE

1. PATTY (Demo)
2. Never Ending Language Learning / ReadTheWeb (Demo / Data)
3. TREC Question Answering Track
4. Lymba (QA System)

Thank you

- ~~Next session in one week (30 November 2013):~~
- Next session (6 November 2013):
  - **Machine Learning Introduction** (Sabine, Elmar)

- Check our Wiki:
  - http://ad-wiki.informatik.uni-freiburg.de/teaching/InformationExtractionWS1314

Backup List of Topics 1/2

NLP Basics

1. **Shallow NLP** (POS tagging, noun phrase chunking, named entity recognition , entity linking) (Code)
2. **Deep NLP** (constituent / dependency parsing, semantic role labeling) (Code)

Relation Extraction

1. Distant supervision for RE without labeled data
2. **Relation Extraction With Matrix Factorization (Code)**

Open Information Extraction (shallow NLP)

1. **TextRunner +** KnowItAll
2. **ReVerb** + R2A2 **(Code)**

Backup List of Topics 2/2

Open Information Extraction (deep NLP)

1. Open IE using Wikipedia
2. SRL-IE
3. **OLLIE (Code)**
4. Dependency based OpenIE (Code)
5. **ClausIE (Code)**
6. Integrating Syntactic and Semantic Analysis into Open IE

Systems using IE / Tasks related to IE

1. PATTY (Demo)
2. Never Ending Language Learning / ReadTheWeb (Demo / Data)
3. TREC Question Answering Track
4. Lymba (QA System)