Introduction to Information Extraction (hands on session)

Dr. Elena Demidova
L3S Research Center

3rd KEYSTONE Summer School
Vienna, Austria
August 22, 2017

The Goals

Get familiar with extraction tools

- NE extraction (Stanford NER)
- NE linking (DBpedia Spotlight)
- Temporal tagging (Stanford SUTime, Heideltime)

Compare performance of extractors

** Use extracted information to compare news articles

** Try Information Extraction tools for other languages

Hands On Session: Web Demos, APIs, Software

Stanford NER

- Stanford CoreNLP: https://stanfordnlp.github.io/CoreNLP/
- Demo: http://corenlp.run

Temporal Taggers Demos

- SUTime: https://nlp.stanford.edu/software/sutime.shtml
  Demo: http://nlp.stanford.edu:8080/sutime/process
- Heideltime: http://heideltime.ifi.uni-heidelberg.de/heideltime

Entity Linking

- DBpedia Spotlight demo: http://demo.dbpedia-spotlight.org/
- Babelfy demo: http://babelfy.org/

Hands On Session: Tasks

Download news articles from different news sources regarding Brexit

- http://www.theguardian.com/politics/eu-referendum
- http://www.nytimes.com/news-event/britain-brexit-european-union

Create annotations using (Stanford NER, DBpedia Spotlight, Babelfy, SUTime, Heideltime) for these articles (using online demo systems).

Analyse the differences between Stanford NER, DBpedia Spotlight and Babelfy annotations (precision and recall of extraction).

** Work in Groups of 2-3: Discuss methods how to use extracted information to compare representations of the Brexit debate in different news sources.

** Advanced: perform extraction in another language of your choice.

Hands On Session: Summary & Feedback

Got familiar with extraction tools

- NE extraction (Stanford NER)
- NE linking (DBpedia Spotlight)
- Temporal tagging (Stanford SUTime, Heideltime)

Performance of extractors in particular tasks

- How good were the extractors w.r.t precision and recall?
- How useful are the tools for comparing the news articles?
- Did you try other languages than English?
- Any other observations?

Hands On Session: Setup (Software)

- Requirements: Eclipse IDE, Git plugin
- Check out from GitHub https://github.com/edemidova/handson2016
- Import project “NLPtutorialMaven” in Eclipse

Hands On Session: NER and Temporal Annotations

- Stanford NER and SUTime:
  - Run uk.soton.examples.nlp.NLPParserDemo.java
  - Pipeline configuration example:

```
public NLPParserDemo() {

Properties props = new Properties();
props.setProperty("annotators",
"tokenize, ssplit, pos, lemma, ner, parse, dcoref");
props.setProperty("parse.model",
"edu/stanford/nlp/models/lexparser/englishPCFG.ser.gz");
props.setProperty("ssplit.isOneSentence", "false");

pipeline = new StanfordCoreNLP(props);
pipeline.addAnnotator(new TimeAnnotator("sutime", new Properties()));}
```

Hands On Session: Output

- Annotated text: POS, NER, LEMMA.

Annotations:

Dubrovnik(POS: NNP, NER: LOCATION, LEMMA: Dubrovnik)

Hands On Session: DBpedia Spotlight

Annotate: runs spotting and disambiguation. Takes text as input, recognizes entities/concepts to annotate and chooses an identifier for each recognized entity/concept given the context.

- Run uk.soton.examples.nlp.DBPediaReader
  - API_URL = "http://model.dbpedia-spotlight.org/en/";
  - CONFIDENCE = 0.6;

http://model.dbpedia-spotlight.org/en/annotate/?confidence=0.6&text=Dubrovnik+is+a+Croatian+city

Hands On Session: DBpedia Spotlight


```
{
    "@URI": "http://dbpedia.org/resource/Dubrovnik",
    "@support": "2423",
    "@types": "DBpedia:PopulatedPlace, DBpedia:Settlement, … Schema:City, DBpedia:City",
    "@surfaceForm": "Dubrovnik",
    "@offset": "0",
    "@similarityScore": "0.999980601335612 6",
    "@percentageOfSecondRank": "1.939904070672243E-5"
  }
```

Hands On Session: DBpedia Spotlight

https://github.com/dbpedia-spotlight/dbpedia-spotlight/wiki/Web-service

Candidates: Similar to annotate, but returns a ranked list of candidates instead of deciding on one. These list contains some properties as described below:

- **support**: how prominent is this entity, i.e. number of inlinks in Wikipedia;
- **priorScore**: normalized support;
- **contextualScore**: score from comparing the context representation of an entity with the text (e.g. cosine similarity with if-icf weights);
- **percentageOfSecondRank**: measure by how much the winning entity has won by takingcontextualScore_2ndRank / contextualScore_1stRank, which means the lower this score, the further the first ranked entity was "in the lead";
- **FinalScore**: combination of all of them;

Hands On Session: Relation Extraction

Run Ollie: uk.soton.examples.nlp.JavaOllieWrapper

Hands On Session: Load online News

- **Boilerpipe library to extract text from news articles / web pages.**
- uk.soton.examples.nlp.NLPParserDemo (**set online = true;)**

```
try {
text = ArticleExtractor.INSTANCE.getText
(new URL("http://www.bbc.co.uk/news/business-37220701"));

} catch (BoilerpipeProcessingException e) {

e.printStackTrace();
} catch (MalformedURLException e) {

e.printStackTrace();
}
Dubrovnik(POS: NNP, NER: LOCATION, LEMMA: Dubrovnik)
```

Thank you!

Questions, Comments?

Dr. Elena Demidova

L3S Research Center
Leibniz University of Hannover

email: demidova@L3S.de
www: https://demidova.wordpress.com

This presentation contains photos from www.fotolia.de