University of Economics
Prague

Czech Technical University
in Prague

Recognizing, Classifying and Linking Entities with Wikipedia and DBpedia

Milan Dojchinovski[1], Tomas Kliegr[2]

[1] Faculty of Information Technology
Czech Technical University in Prague

[2]Faculty of Informatics and Statistics
University of Economics, Prague

Milan Dojchinovski
milan.dojchinovski@fit.cvut.cz - @m1ci - http://dojchinovski.mk

The 7th Workshop on Intelligent and Knowledge Oriented Technologies (WIKT 2012)
November 22-23, 2012, Smolenice, SK

Overview

- Introduction
- Entity Recognition, Classification and Publication
- Experiments
- Conclusion and Future Work

Introduction

- Unsupervised and fully-automated:
  - entity recognition - rule based lexico-syntactic patterns
  - entity classification by extraction of hypernyms - targeted hypernym extraction
  - entity linking to DBpedia concepts
- Publication as Linked Data
  - results in NLP Interchange Format (NIF)

Overview

- Introduction
- Entity Recognition, Classification and Publication
- Experiments
- Conclusion and Future Work

Tool Architecture

- Available as Web 2.0 application at: http://ner.vse.cz/thd
- Web API available at: http://ner.vse.cz/thd/docs

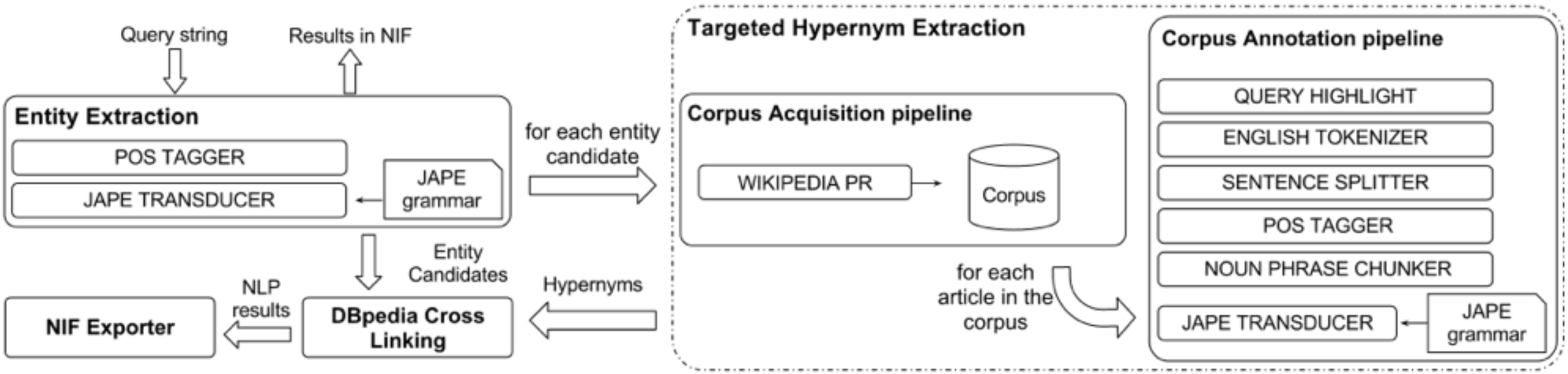

Fig 1. Architecture overview

Entity Recognition and Classification

- Entity Recognition
  - 2 JAPE grammars: 1) NNP+ 2) JJ* NN+
  - input: free text
  - output: Named (e.g., “Diego Maradona”) or Common Entities (e.g., “hockey player”)
- Entity Classification
  - supported by the Targeted Hypernym Discovery algorithm
  - lexico-syntactic patterns, e.g. “_x_ is a _y_ “

Entity Linking and Publication

- Entity Linking
  - linking with concepts from DBpedia
  - used Wikipedia Search API
  - mapping Wikipedia article URL to its DBpedia representation
- Publication in NIF
  - NLP Interchange Format (RDF-based representation)
  - each processed document (context) has unique identifier
  - each entity and hypernym as offset-based string

Overview

- Introduction
- Entity Recognition, Classification and Publication
- Experiments
- Conclusion and Future Work

Experiments

- Question addressed
  - How well our tool recognizes, classifies and links Named and Common Entities?
- Experiment setup
  - manually created dataset, “Czech Traveler Dataset”
  - 101 Named Entities, 85 Common Entities
  - comparison with 3 other systems: DBpedia Spotlight, Open Calais, Alchemy API
- Results
  - Named Entities,
    - f-score: recognition 0.66, classification 0.66, linking 0.58
  - Common Entities
    - f-score: recognition 0.60, classification 0.51, linking 0.61
  - better results in all tasks
    - overtaken only by DBpedia Spotlight - linking of common entities with f-score 0.69

Overview

- Introduction
- Entity Recognition, Classification and Publication
- Experiments
- Conclusion and Future Work

Conclusion and Future Work

- Tool for Entity Recognition, Classification and Publication
- Future directions
  - multilingual support - Dutch, German and Czech language
  - grammar improvements
  - evaluation on a standard benchmark

Feedback

Thank you!

Questions, comments, ideas?

demo at: http://ner.vse.cz/thd

Milan Dojchinovski
milan.dojchinovski@fit.cvut.cz
@m1ci
http://dojchinovski.mk