Ontology-Based Information Extraction: A machine learning approach

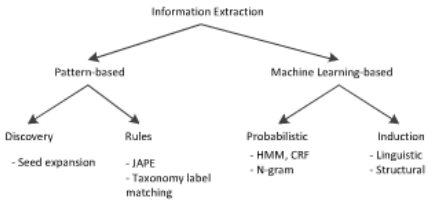
Slavko Žitnik
slavko.zitnik@fri.uni-lj.si

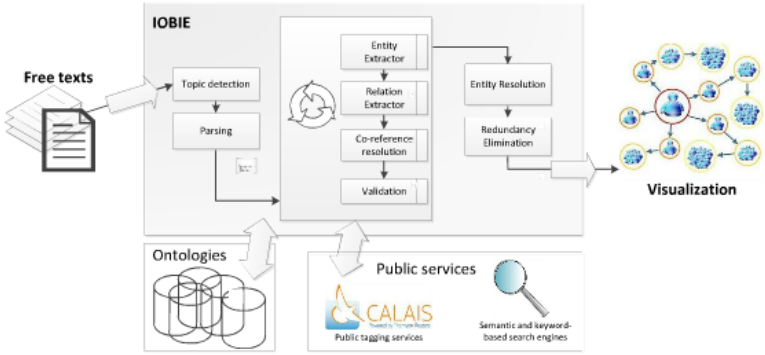
Optilab d.o.o.
Šuštarjeva 8
5270 Ajdovščina

University of Ljubljana
Faculty of computer and information science
Tržaška cesta 25
1000 Ljubljana

Information extraction

- As a task: Filling slots in a database from text
- As a family of techniques:
  segmentation+classification+association+clustering





Evaluation

- each component separately
- whole IOBIE system
  - MUC 1-7 (Message Understanding Conference)
  - RISE (Repository of Information Sources)
  - ACE (Automatic Content Extraction)
  - Enron Email Dataset
  - Wikipedia with DBPedia

Ontology-Based Information Extraction: A machine learning approach

Investing in your future
OPERATION PART FINANCED BY THE EUROPEAN UNION
European Regional Development Fund
Cohesion Fund
European Social Fund

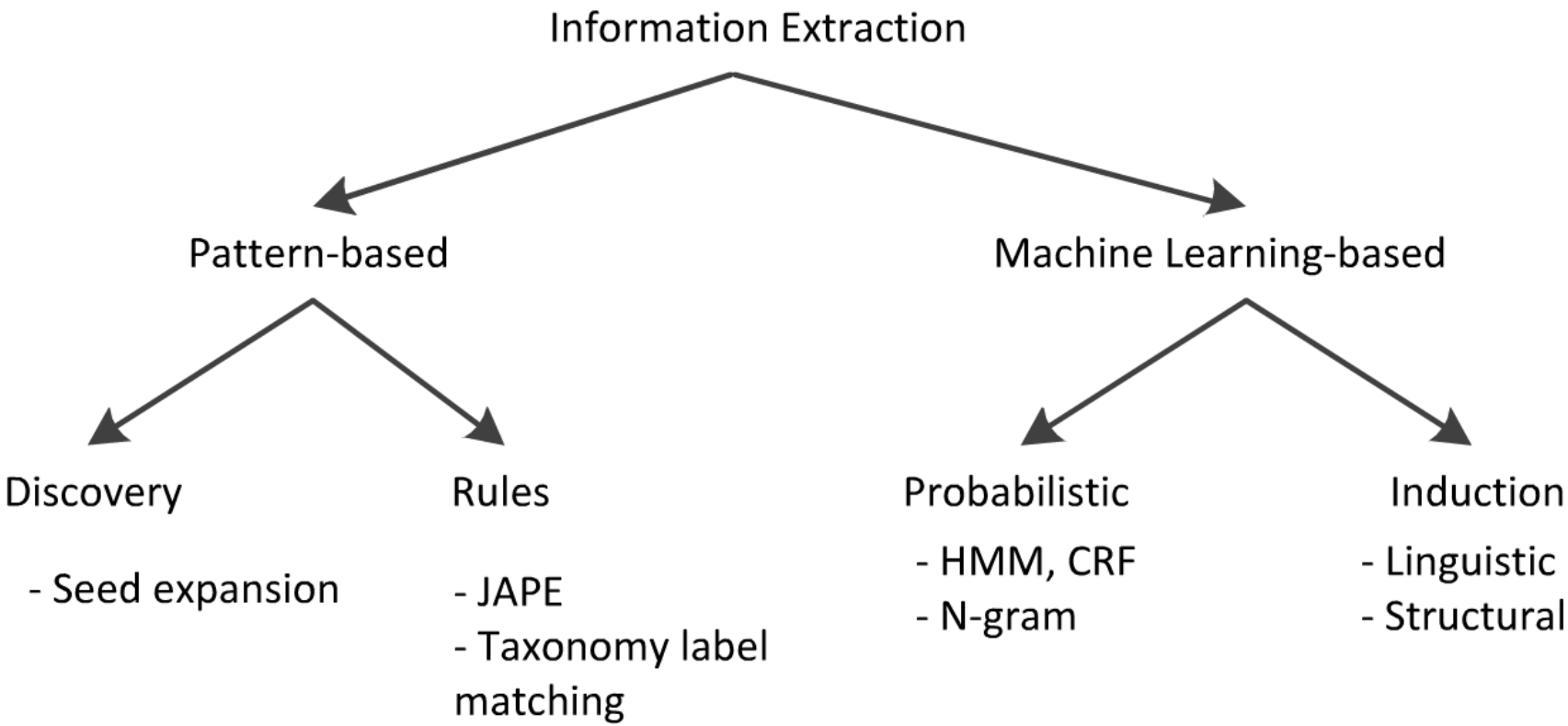
Slavko Žitnik
slavko.zitnik@fri.uni-lj.si

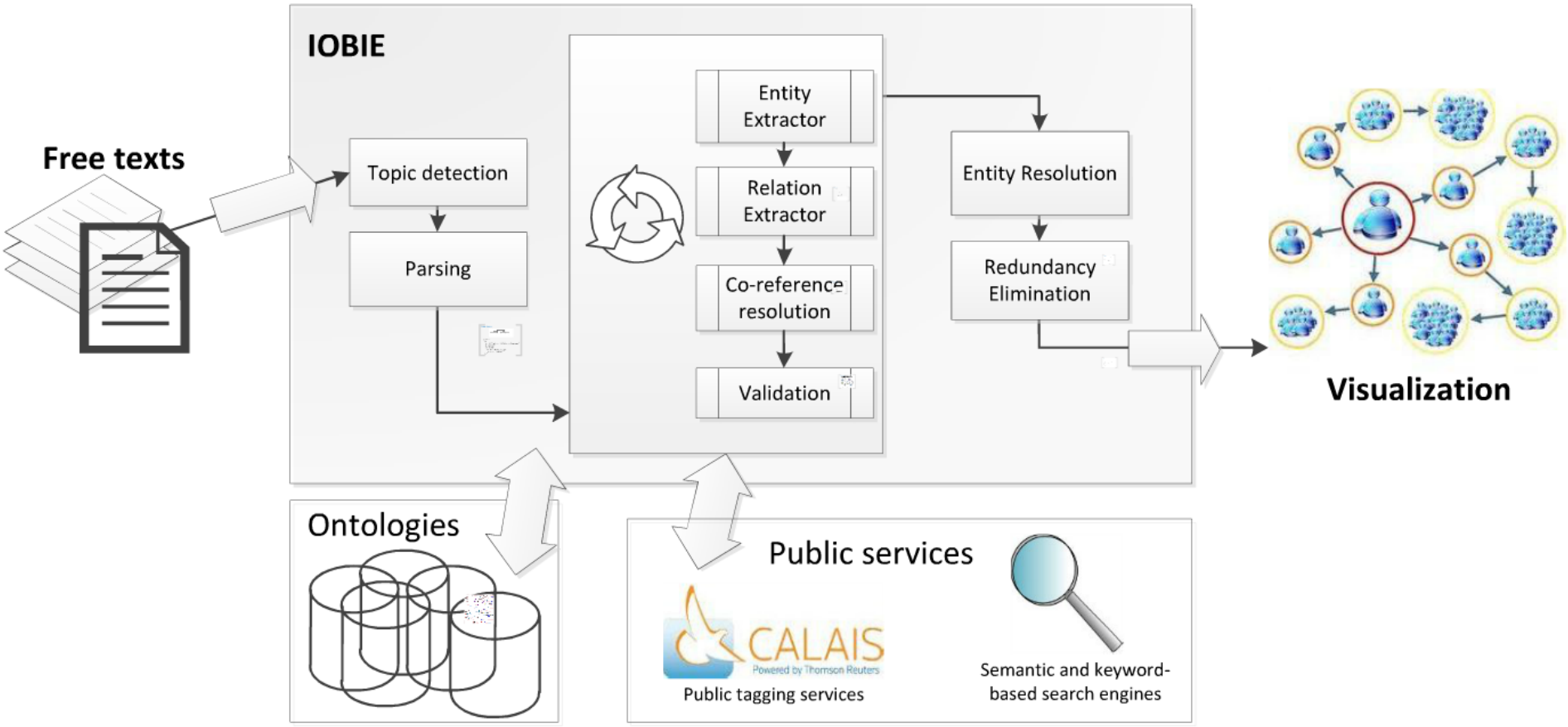
Optilab d.o.o.
Župančičeva 8
5270 Ajdovščina

University of Ljubljana
Faculty of computer and information science
Tržaška cesta 25
1000 Ljubljana

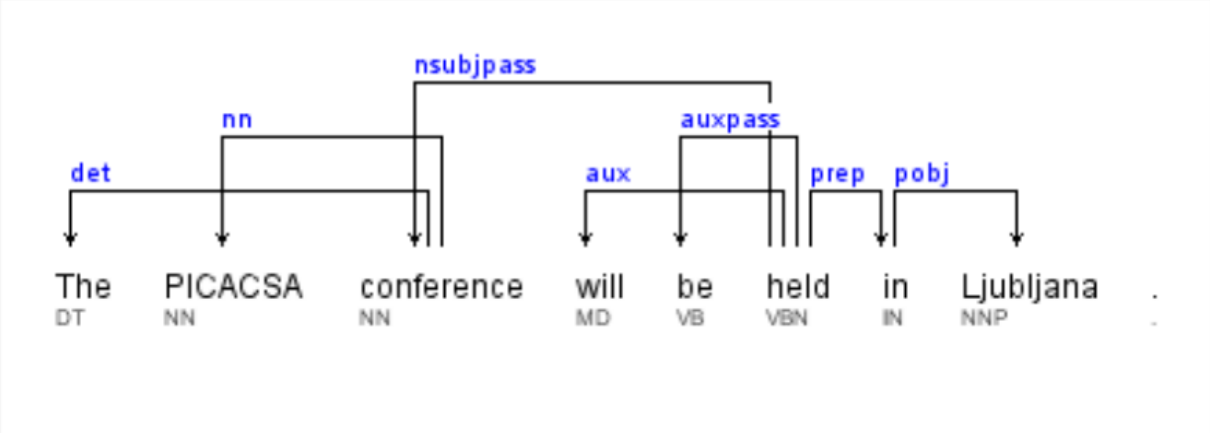
Information extraction

- As a task: Filling slots in a database from text
- As a family of techniques:
  segmentation+classification+association+clustering

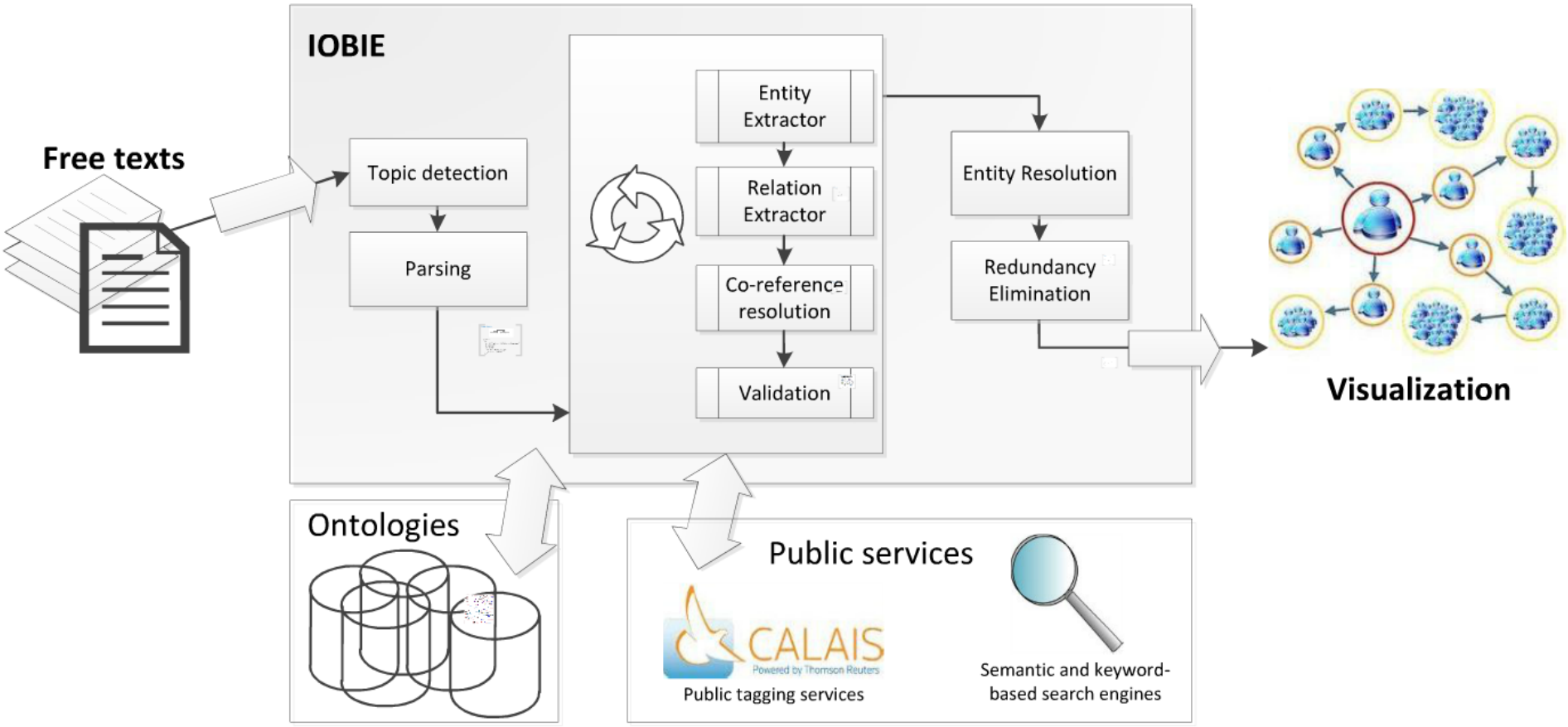



Free texts
IOBIE
Topic detection
Parsing
Entity Extractor
Relation Extractor
Co-reference resolution
Validation
Entity Resolution
Redundancy Elimination
Visualization
Ontologies
Public services
CALAIS
Powered by Thomson Reuters
Public tagging services
Semantic and keyword-based search engines

Stanford parser




```
(ROOT
 (S
  (NP (DT The) (NN PICACSA)  (NN Conference))
  (VP (MD  will)
   (VP (VB be)
    (VP (VBN held) (PP (IN  in)
     (NP (NNP  Ljubljana))))))
 (. .)))
```

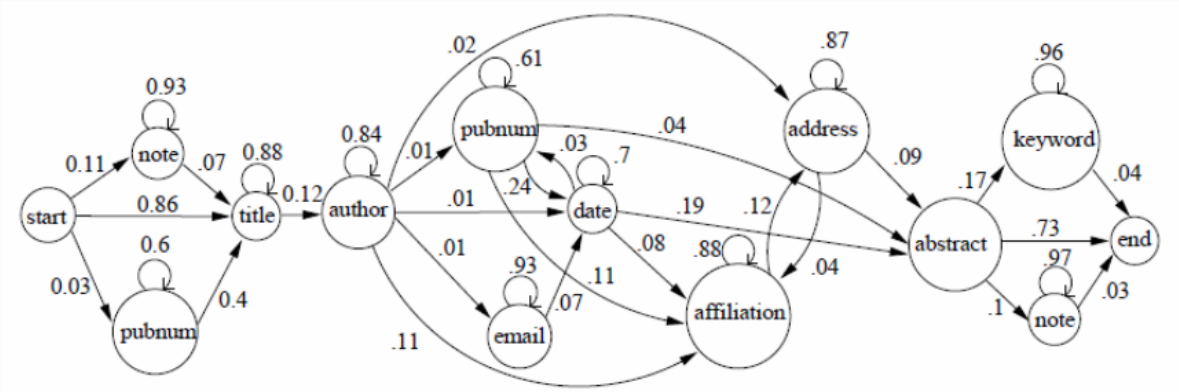

Free texts
IOBIE
Topic detection
Parsing
Entity Extractor
Relation Extractor
Co-reference resolution
Validation
Entity Resolution
Redundancy Elimination
Visualization
Ontologies
Public services
CALAIS
Powered by Thomson Reuters
Public tagging services
Semantic and keyword-based search engines

Entity Extraction

- Example

  <faculty>FRI</faculty>,
  located in
  <location>Ljubljana</location>
  will host
  <conference>PICACSA</conference>
  conference on
  <date>June 3-4 2011</date>.

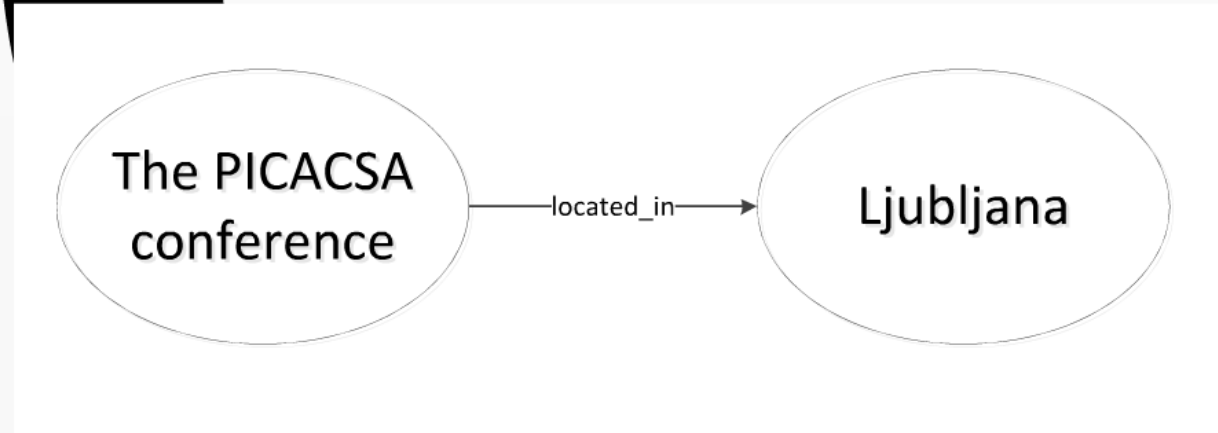
- Hidden Markov Models



Relation extraction

- Example

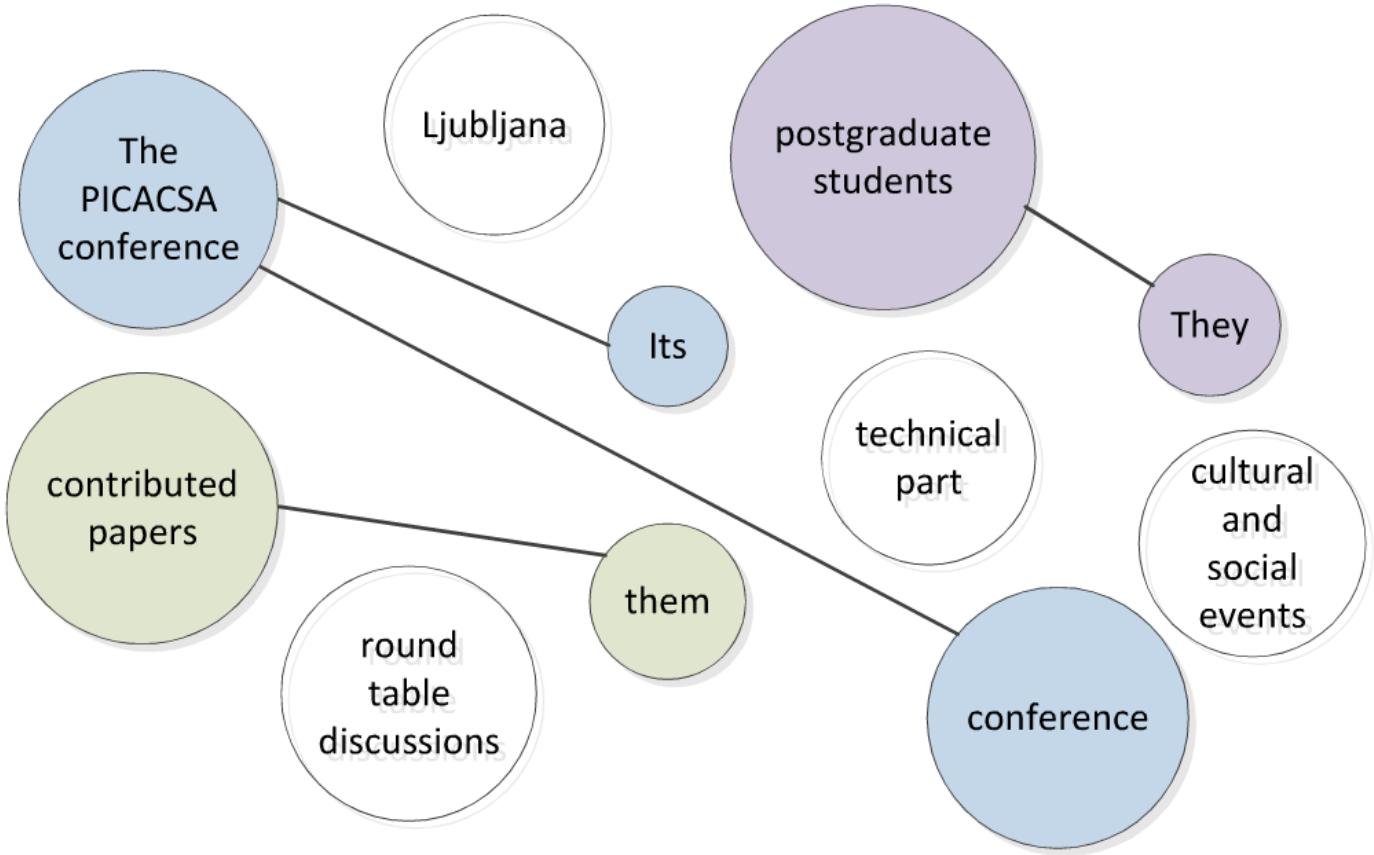
<conference>The PICACSA conference</conference>
will be held in
<location>Ljubljana.</location>



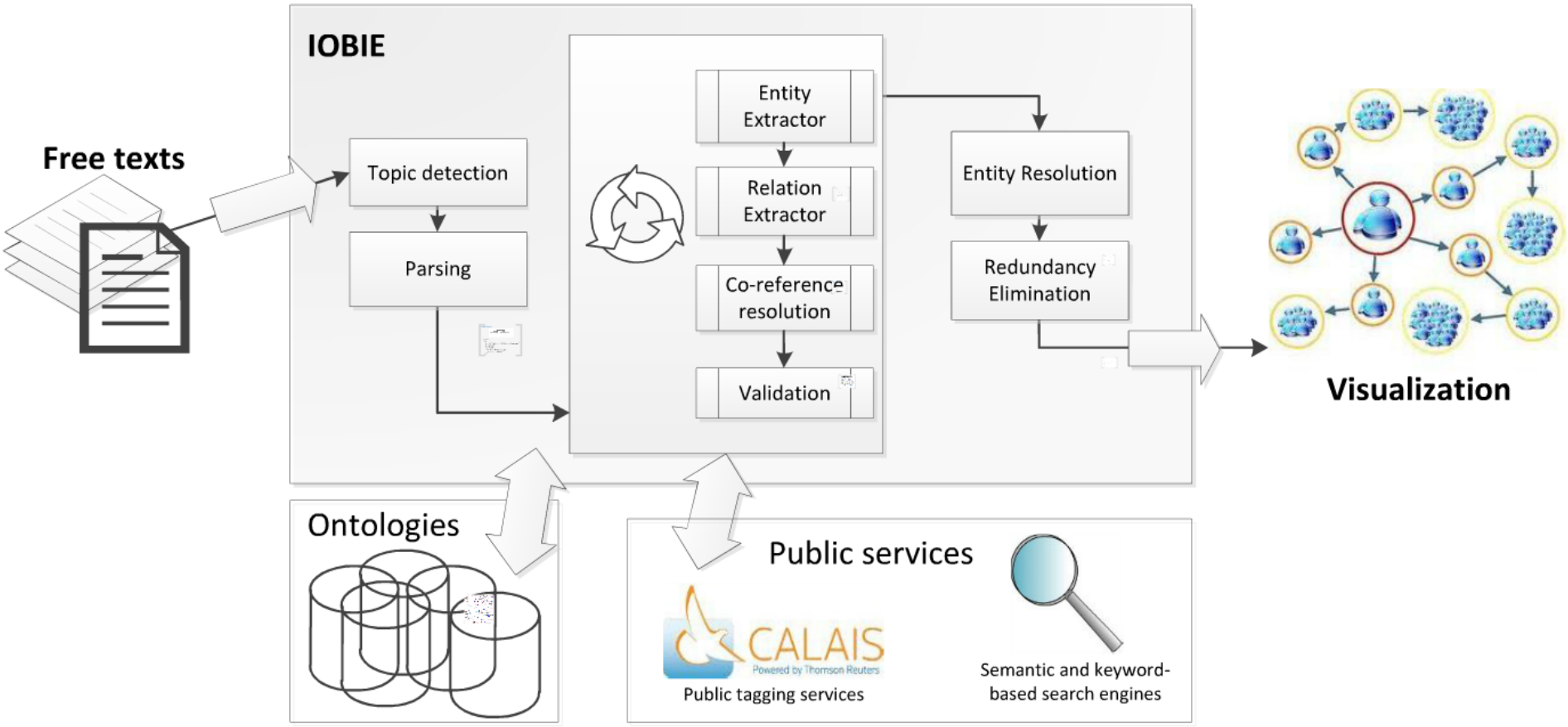
- Feature-based methods

Co-reference Resolution

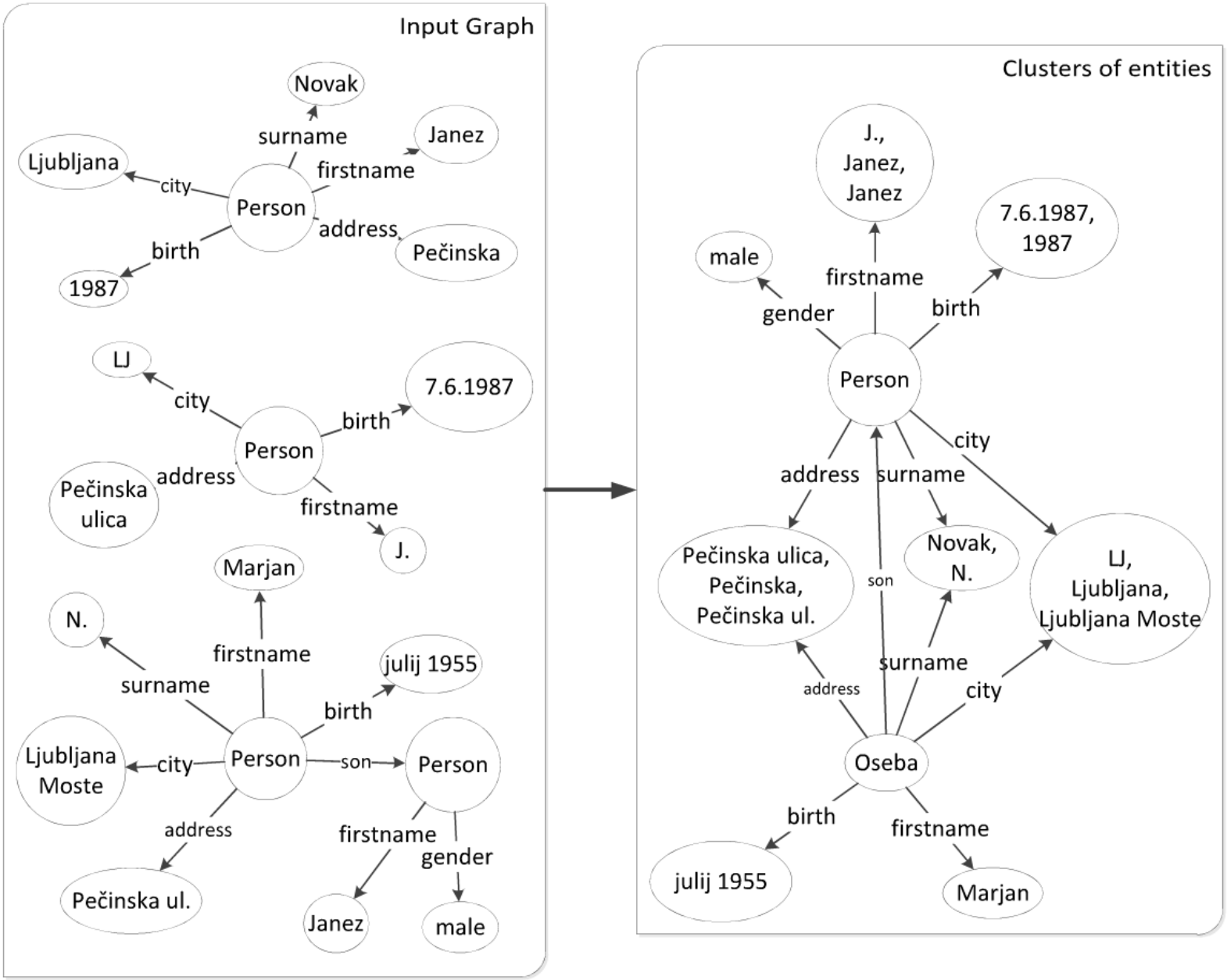
The PICACSA Conference1 will be held in Ljubljana2. Its1 purpose is to bring together a panel of postgraduate students4. They4 will present contributed papers5 and have round table discussions6 about them5. The technical part7 of the conference1 will be accompanied by cultural and social events8.



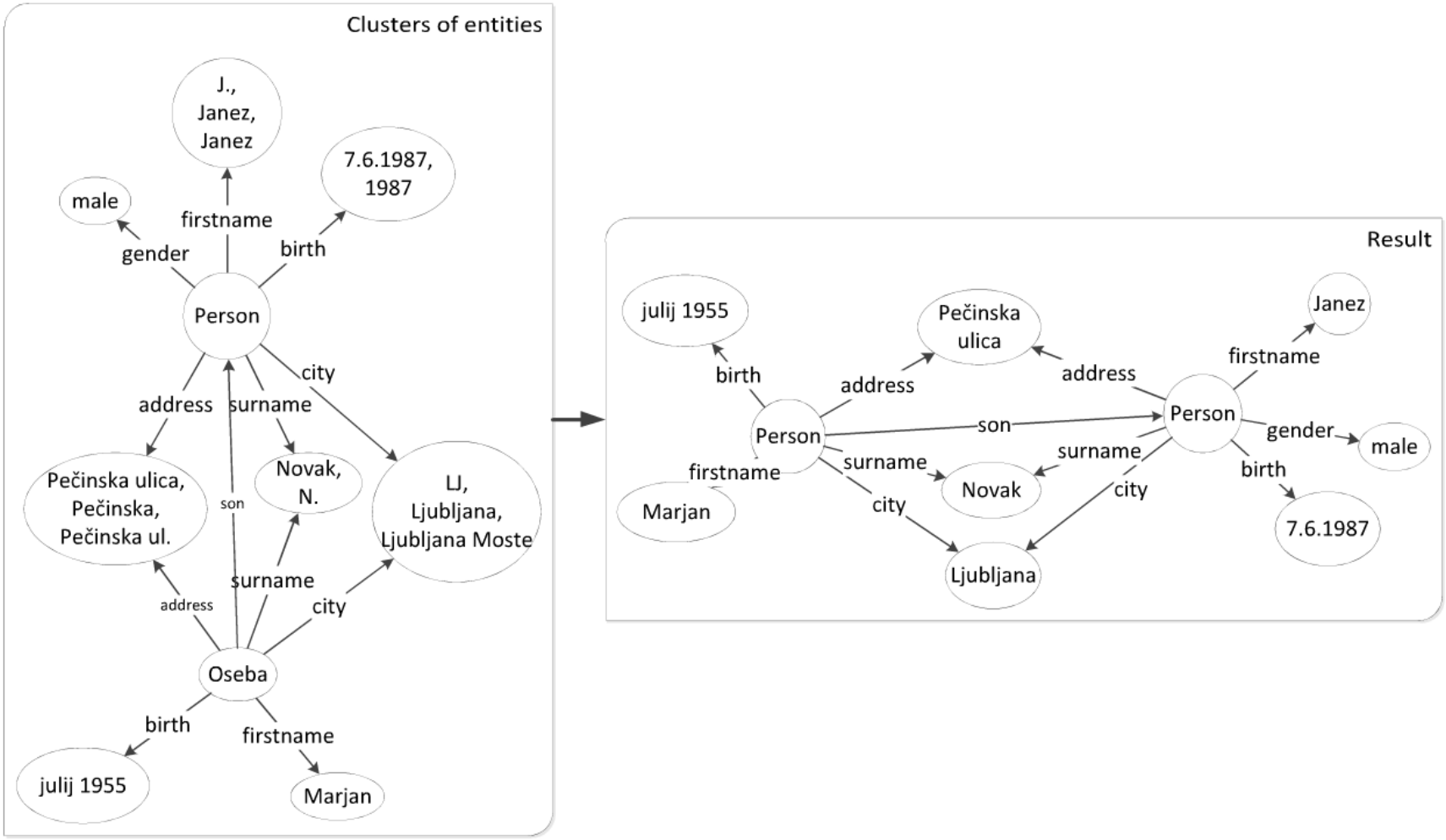
- Pairwise coreference model

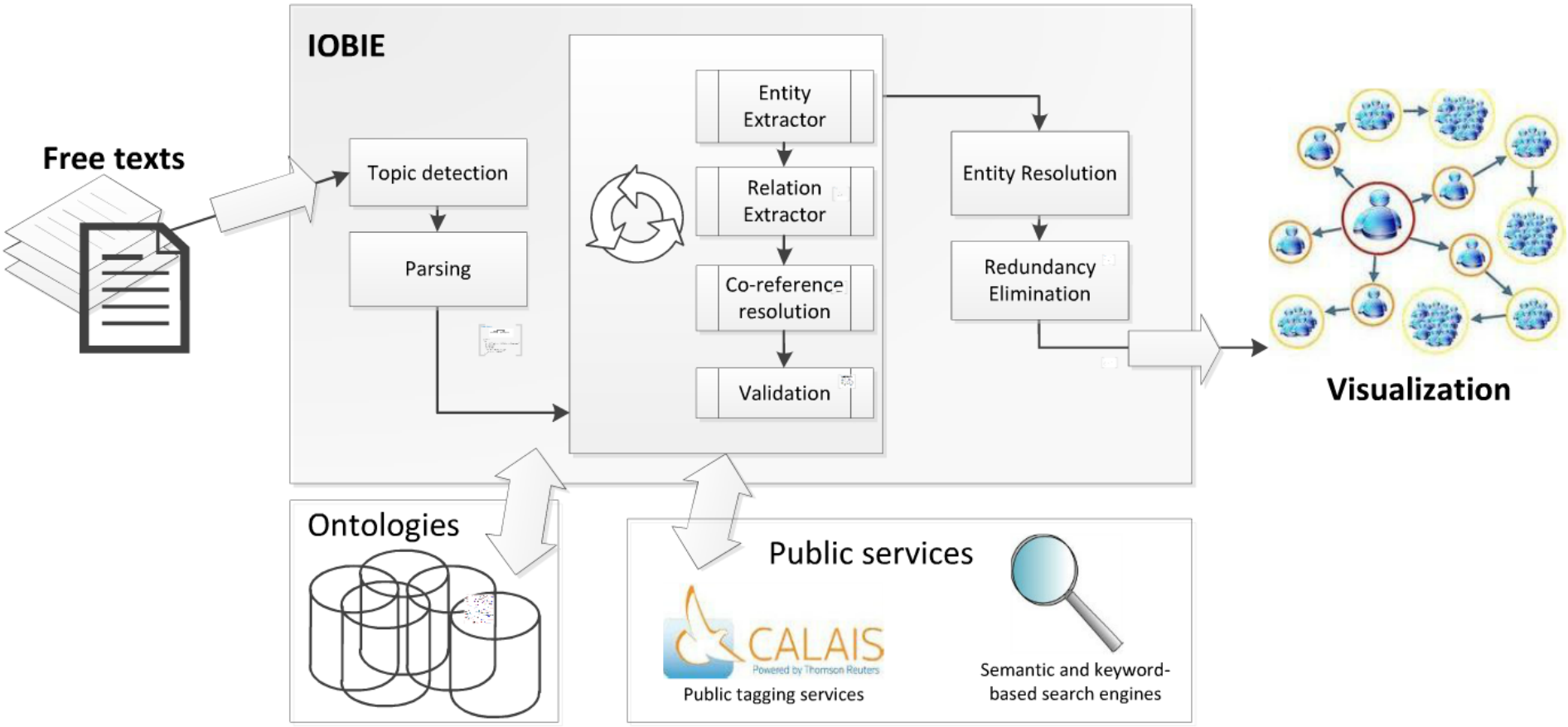

Free texts
IOBIE
Topic detection
Parsing
Entity Extractor
Relation Extractor
Co-reference resolution
Validation
Entity Resolution
Redundancy Elimination
Visualization
Ontologies
Public services
CALAIS
Powered by Thomson Reuters
Public tagging services
Semantic and keyword-based search engines

Entity Resolution



Redundancy elimination




Free texts
IOBIE
Topic detection
Parsing
Entity Extractor
Relation Extractor
Co-reference resolution
Validation
Entity Resolution
Redundancy Elimination
Visualization
Ontologies
Public services
CALAIS
Powered by Thomson Reuters
Public tagging services
Semantic and keyword-based search engines

Evaluation

- each component separately
- whole IOBIE system
  - MUC 1-7 (Message Understanding Conference)
  - RISE (Repository of Information Sources)
  - ACE (Automatic Content Extraction)
  - Enron Email Dataset
  - Wikipedia with DBPedia


Thanks