IOBIE

Intelligent Ontology-based Information Extraction

Slavko Žitnik
Mentor: prof. dr. Marko Bajec
Laboratorij za podatkovne tehnologije

Motivation

local
www
text sources ->
no automatic extraction

relation
combining tasks
co-reference entity

driven
Ontology-
based

never ending learning
[Mitchell T. et. al., 2010]

insufficient classic architecture

Information Extraction (IE)

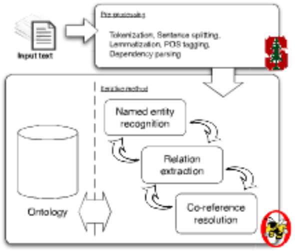
Definition

IE is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured machine-readable documents.

Common approaches

Pattern-based	Machine Learning-based
• Hand-written rules	• Induction
• Seed expansion	• Classifiers
	• Sequence models

IOBIE architecture



Conditional Random Fields

[Lafferty, J., McCallum, A., Pereira, F., 2001]

x = Mr. Janez Novak studied at FRI.
y = O PER PER O O ORG

Best labeling:

$$\hat{y} = \operatorname{argmax}_{\bar{y}} p(\bar{y}|\bar{x}; w)$$

Standard CRF model:

$$p(\bar{y}|\bar{x}; w) = \frac{\exp \sum_{j=1}^{J} w_j F_j(\bar{x}, \bar{y})}{\sum_{\bar{y}'} \exp \sum_{j=1}^{J} w_j F_j(\bar{x}, \bar{y}')}$$

(LC) Feature function:

$$F_j(\bar{x}, \bar{y}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \bar{x}, i)$$

Feature function example:

f(yi-1, yi, x, i) =
IF (x(i-1) == "Mr." && yi == "PER") THEN 1 ELSE 0

Applied in: NLP, bioinformatics, computer vision, pattern recognition, etc.

Named Entity Recognition (NER)

Feature functions classification:

- Preprocessed
- String
- Semantic
- Iteration:
  - NER
  - Relation
  - Co-reference

Relation (Mention) Extraction

[Li et. al., 2011]

x = Janez was born in Ljubljana.
y = ARG-1 O B-REL I-REL ARG-2

Label sequence constraint
Long range features
Features*:
- Semantic:
  - Relation property (arity, functional, reverse)
  - Unseen relation
- Iteration:
  - Repeated relation

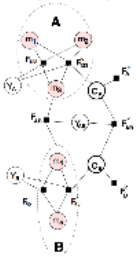
* - as defined before

Co-reference resolution

[Wick et. al., 2009]

- Non-linear CRF
- Relational clustering

Features:
- String:
  - intervening words, apposition, distance
- Iteration:
  - #mentions between, relation set feasible



Future Work

- Full implementation
- New features engineering
- Evaluation
- Slovene

Contribution

- Iterative method
- General IE framework
- Common algorithms for main tasks
- Extensive ontology use

Questions are guaranteed in life; Answers aren't.

IOBIE

Intelligent Ontology-based Information Extraction

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

Slavko Žitnik
Mentor: prof. dr. Marko Bajec
Laboratorij za podatkovne tehnologije

Information Extraction (IE)

Definition

IE is a type of information retrieval whose goal is to automatically extract structured information from unstructured and/or semi-structured machine-readable documents.

Common approaches

Pattern-based

- Hand-written rules
- Seed expansion

Machine Learning-based

- Induction
- Classifiers
- Sequence models

Motivation

local

www

text sources ->

no automatic extraction

relation

driven

combining tasks

Ontology-

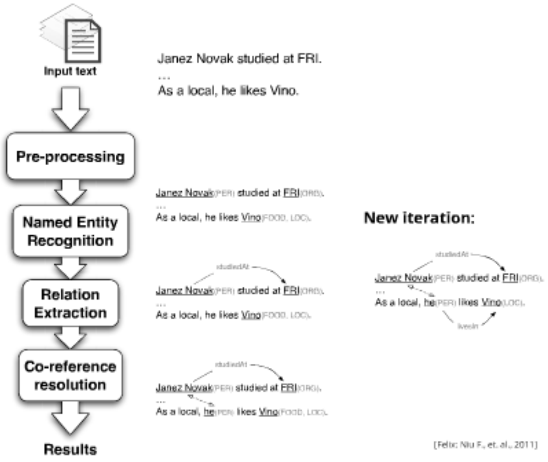
based

co-reference

entity

never ending learning

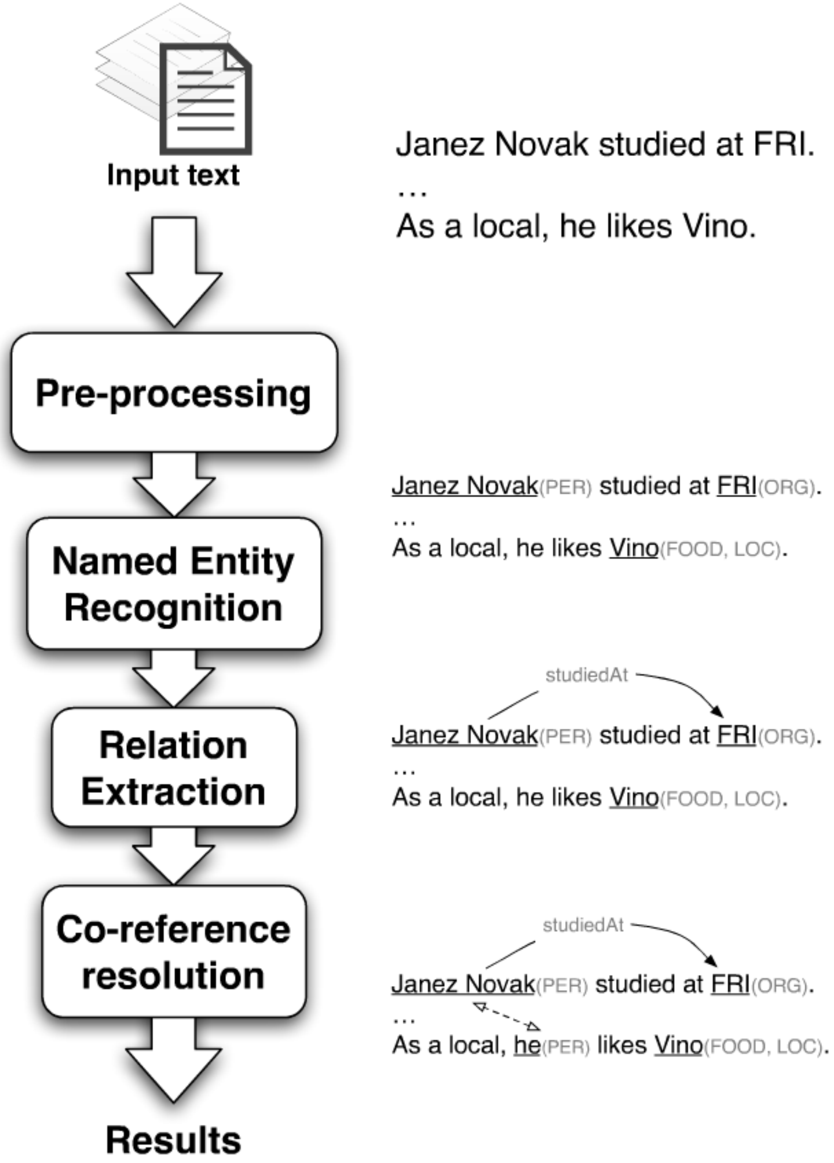
[Mitchell T. et. al., 2010]


Input text
Janez Novak studied at FRI.
...
As a local, he likes Vino.
Pre-processing
Janez Novak(PER) studied at FRI(ORG).
...
As a local, he likes Vino(FOOD, LOC).
Named Entity Recognition
studiedAt
Janez Novak(PER) studied at FRI(ORG).
...
As a local, he likes Vino(FOOD, LOC).
Relation Extraction
studiedAt
Janez Novak(PER) studied at FRI(ORG).
...
As a local, he(PER) likes Vino(FOOD, LOC).
Co-reference resolution
Results
New iteration:
studiedAt
Janez Novak(PER) studied at FRI(ORG).
...
As a local, he(PER) likes Vino(LOC).
livesIn

[Felix: Niu F., et. al., 2011]

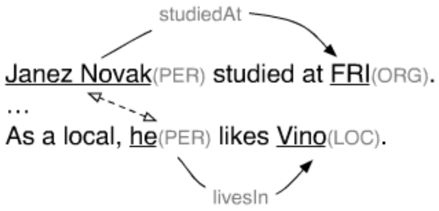
insufficient classic architecture

n


Input text
Pre-processing
Named Entity Recognition
Relation Extraction
Co-reference resolution
Results
Janez Novak studied at FRI.
…
As a local, he likes Vino.
Janez Novak(PER) studied at FRI(ORG).
…
As a local, he likes Vino(FOOD, LOC).
studiedAt
Janez Novak(PER) studied at FRI(ORG).
…
As a local, he likes Vino(FOOD, LOC).
studiedAt
Janez Novak(PER) studied at FRI(ORG).
…
As a local, he(PER) likes Vino(FOOD, LOC).

in

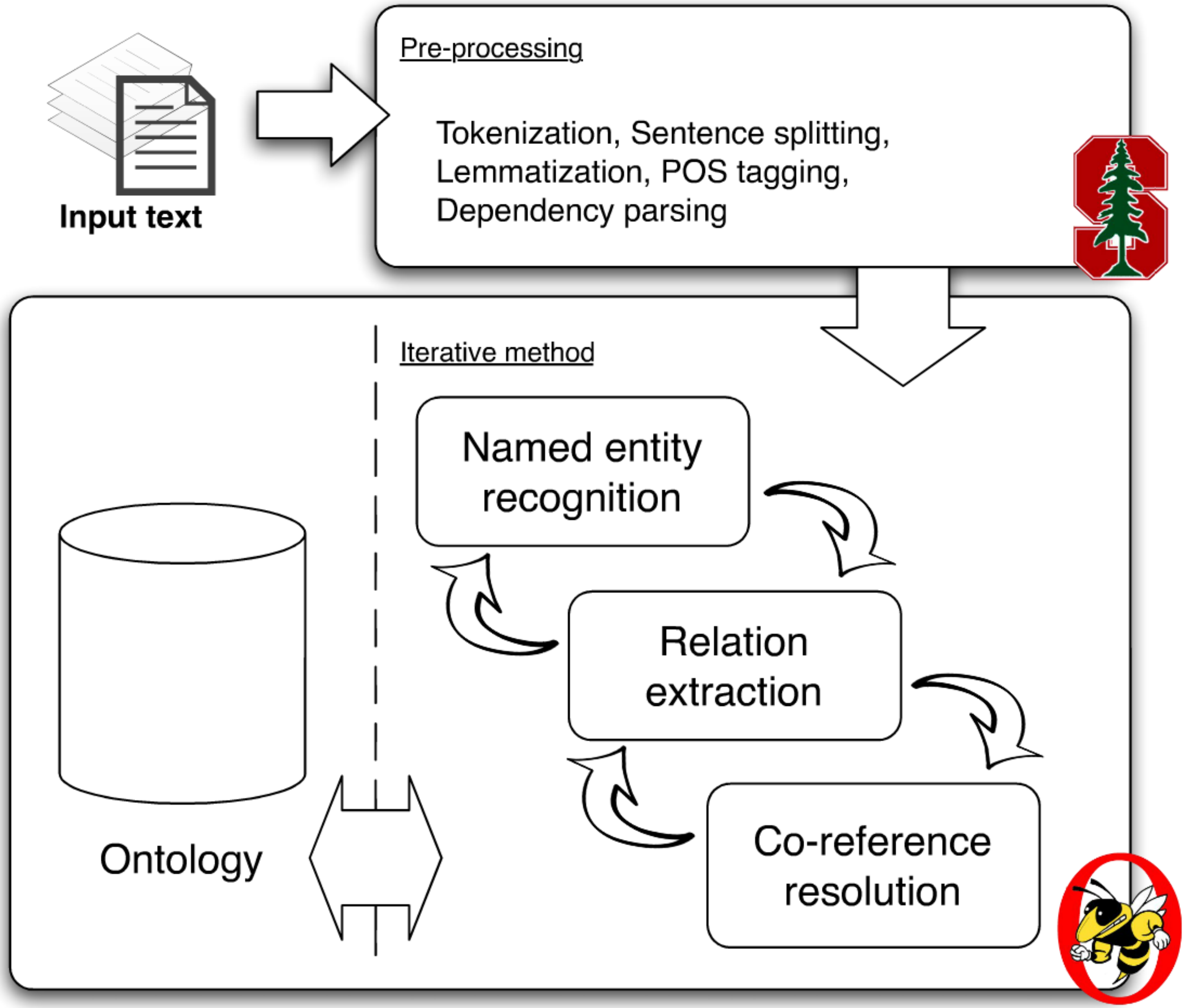
New iteration:


studiedAt
Janez Novak(PER) studied at FRI(ORG).
…
As a local, he(PER) likes Vino(LOC).
livesIn

a

[Felix: Niu F., et. al., 2011]

IOBIE architecture



Input text

Pre-processing

Tokenization, Sentence splitting, Lemmatization, POS tagging, Dependency parsing

Iterative method

Named entity recognition

Relation extraction

Co-reference resolution

Ontology

Conditional Random Fields

[Lafferty, J., McCallum, A., Pereira, F., 2001]

x = Mr. Janez Novak studied at FRI.
y = O PER PER O O ORG

Best labeling:

$$\hat{y} = \text{argmax}_{\bar{y}} p(\bar{y}|\bar{x}; w)$$

Standard CRF model:

$$p(\bar{y}|\bar{x}; w) = \frac{\exp \sum_{j=1}^{J} w_j F_j(\bar{x}, \bar{y})}{\sum_{\bar{y}'} \exp \sum_{j=1}^{J} w_j F_j(\bar{x}, \bar{y}')}$$

(LC) Feature function:

$$F_j(\bar{x}, \bar{y}) = \sum_{i=1}^{n} f_j(y_{i-1}, y_i, \bar{x}, i)$$

Feature function example:

```
f(yi1, yi, x, i) =
        IF (x(i-1) == "Mr." && yi == "PER") THEN 1 ELSE 0
```

Applied in: NLP, bioinformatics, computer vision, pattern recognition, etc.

Named Entity Recognition (NER)

Feature functions classification:

- Preprocessed [illegible]
- String [illegible]
- Semantic [illegible]
- Iteration:
    - NER
    - Relation
    - Co-reference

- Coreference NER types
- Relation subject/object
- Coreferent [illegible]
- Iteration number
- Type change during iteration
- Multiple labelings

- Lemma
- POS tag
- Chunk tag
- Parse tree

- **Prefix**
- **Suffix**
- **Words**
- **Word shape**
- **Position**
- **N-gram**
- **TF-IDF**
- **String distance**

C

- Gender match
- Number match
- Property match
- Relation match
- Predefined features with weights
- Rules (inference)
- Regular expressions
- Constraints
- "Gazetteer" lists

- **Co-reference NER types**
- **Is relation subject/object**
- **Co-referent relations set**
- **Iteration number**
- **Type change during iterating**
- **Multiple labelings**

Relation (Mention) Extraction

[Li et. al., 2011]

x =	Janez	was	born	in	Ljubljana.
y =	ARG-1	O	B-REL	I-REL	ARG-2

Label sequence constraint

Long range features

Features*:

- Semantic:
    - Relation property (arity, functional, reverse)
    - Unseen relation
- Iteration:
    - Repeated relation

* - as defined before

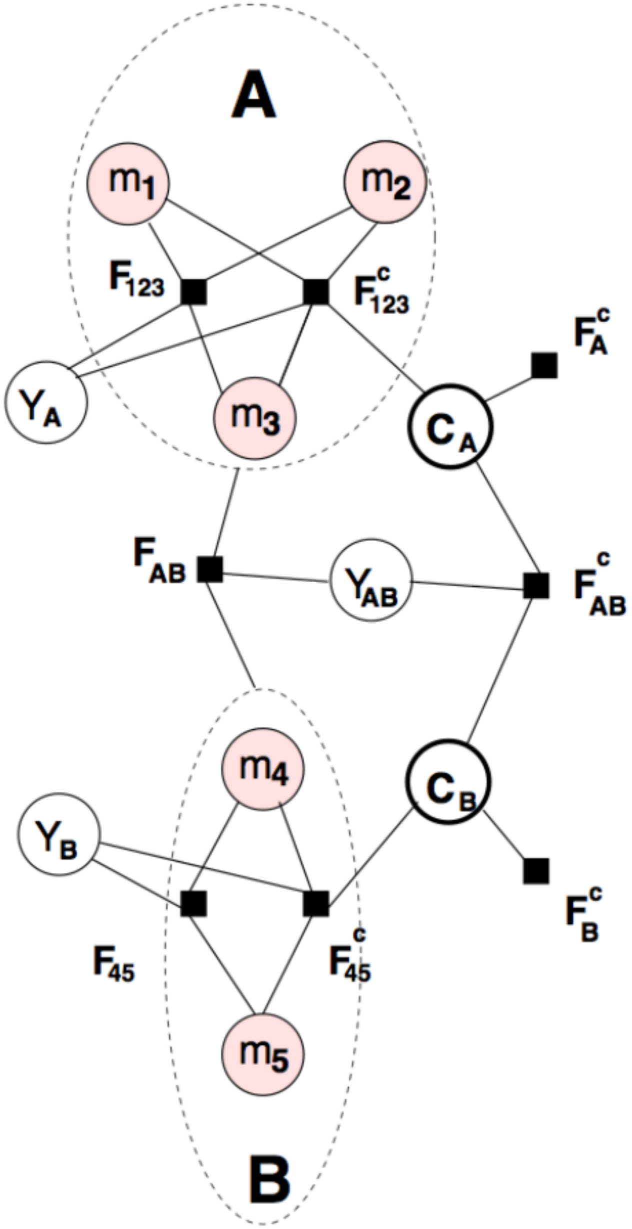
Co-reference resolution

- Non-linear CRF
- Relational clustering

Features:

- String:
    - intervening words, apposition, distance
- Iteration:
    - #mentions between, relation set feasible

[Wick et. al., 2009]



Future Work

- Full implementation
- New features engineering
- Evaluation
- Slovene

Contribution

- Iterative method
- General IE framework
- Common algorithms for main tasks
- Extensive ontology use

Questions
are
guaranteed in
life;
Answers
aren't.