A Graph-based Method for Entity Linking

David Soares Batista

Disciplina de Recuperação de Informação, Instituto Superior Técnico

December 9, 2011

What is Entity-Linking ?

Entity Linking is the process of associating an **entity mentioned in a text** to an entry, representing that entity, in a **knowledge base**

What is Entity-Linking ?

Entity Linking is the process of associating an **entity mentioned in a text** to an entry, representing that entity, in a **knowledge base**

"There are hardly any countries here which said they were ready to go along with the EFSF (euro zone rescue fund)," German Chancellor Angela Merkel told a news conference.

Potential sovereign investors such as China and Brazil wanted to see more detail before they made any firm commitment to put money into the bailout fund.

Global stocks and the euro fell as doubts resurfaced about Europe's financial rescue package.

What is Entity-Linking ?

Entity Linking is the process of associating an **entity mentioned in a text** to an entry, representing that entity, in a **knowledge base**

"There are hardly any countries here which said they were ready to go along with the EFSF (euro zone rescue fund)," German Chancellor Angela Merkel told a news conference.

Angela Dorothea Merkel (; née **Kasner**, born 17 July 1954 in Hamburg) is the current Chancellor of Germany (since 22 November 2005).

96% probability of being a link

Brazil wanted to see

nt to put money into the

ced about Europe's financial rescue package.

What is Entity-Linking ?

Entity Linking is the process of associating an **entity mentioned in a text** to an entry, representing that entity, in a **knowledge base**

"There are hardly any countries here which said they were ready to go along with the EFSF (euro zone rescue fund)," German Chancellor Angela Merkel told a news conference.

Potential sovereign investors such as China and Brazil wanted to see m t to put money into the ba

G ed about Europe's fir

Brazil (; ,), officially the **Federative Republic of Brazil** (,), is the largest country in South America.

94% probability of being a link

Ambiguity

the same name can refer more than one entity *(*"James Cook")

- 4 organizations (Universities, Institutes, etc.)
- 11 persons (British explorer, former NFL player, UK boxer, etc)

the same entity can be referred to by more than one name

- *"Praça do Comércio"* and *"Terreiro do Paço"*
- *"Roger Bacon"* also known as *"Doctor Mirabilis"*

Entity Linking task in the Text Analysis Conference

- Participants are provided with: a set of **queries**, an **external knowledge** base and a **collection of documents**
- **Goal**: For every query, return the identifier of the entity in the knowledge base, or NIL if the entity is not part of it

Evaluation:

$$Accuracy = \frac{Number\ of\ Correct\ Queries}{Total\ Number\ of\ Queries}$$

Typical Approach

- **Candidates Generation**: generate sub-set of the Knowledge Base with potential candidates
- **Candidates Ranking**: rank the candidates according to some criteria
- **NIL clustering:** queries that had a NIL answer and refer to the same entity are clustered

Outline

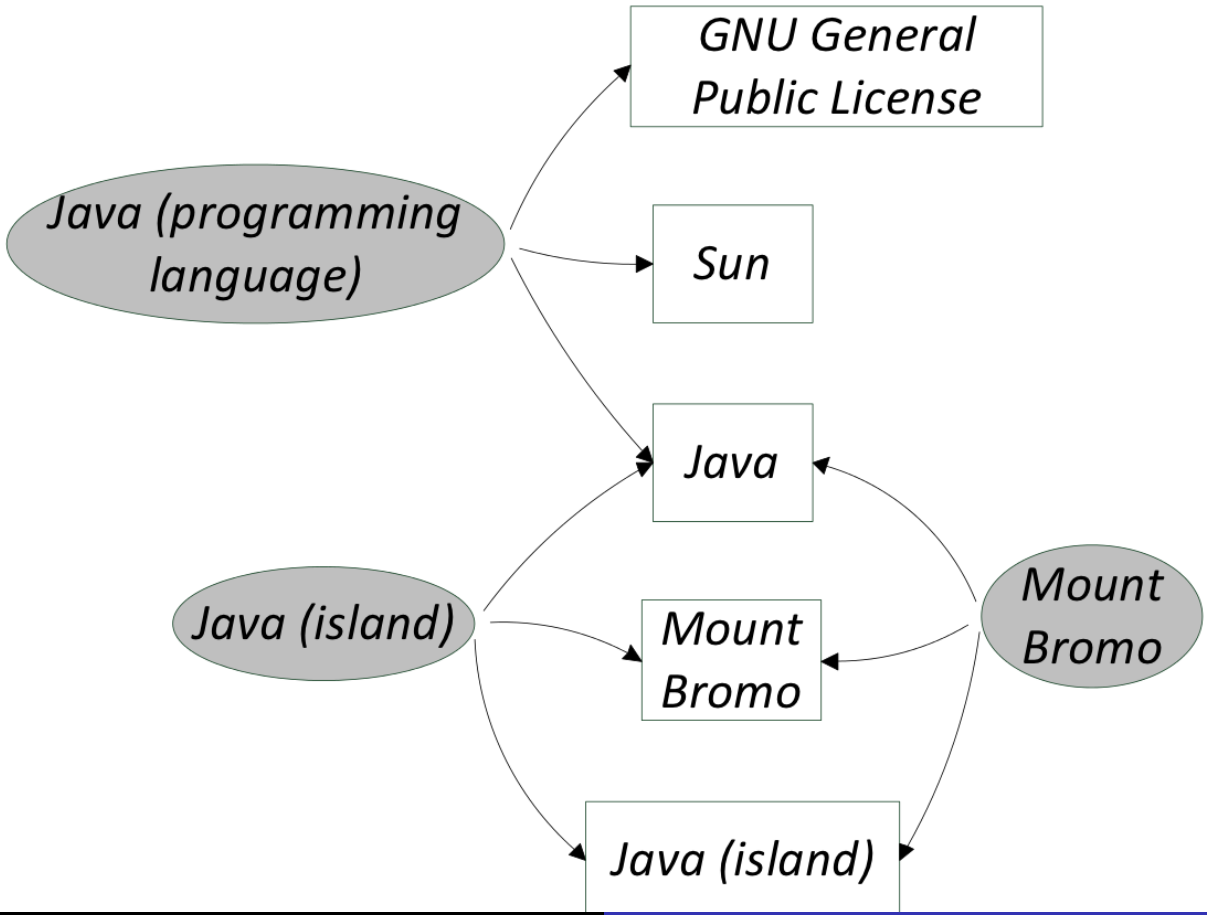
"A Graph-based Method for Entity Linking"

5th International Joint Conference on Natural Language Processing,2011

- Introduction: Wikipedia as a graph
- Disambiguation method: out-degree, in-degree
- Entity Linking System: Process
- Experiments and Results
- Conclusion

Wikipedia as a graph

"Sun relicensed most of its Java technologies under the [[GNU General Public License]]."



Query Example

- Query String: *Java*
- Context: *"Mount Bromo is on of Java's most popular tourist attractions."*

Query Example

- Query String: *Java*
- Context: *"Mount Bromo is on of Java's most popular tourist attractions."*

Extract all the names from the context that are also wikipedia article titles

- Context: "**Mount Bromo** is on of **Java**'s most popular tourist attractions."

Query Example

- Query String: *Java*
- Context: *"Mount Bromo is on of Java's most popular tourist attractions."*

Extract all the names from the context that are also wikipedia article titles

- Context: "**Mount Bromo** is on of **Java**'s most popular tourist attractions."

Candidates from the KB (based on query string):

- Java (island)
- Java (programming language)

Information gathered from query

Wikipedia articles:

- Java (island)
- Java (programming language)

Context names:

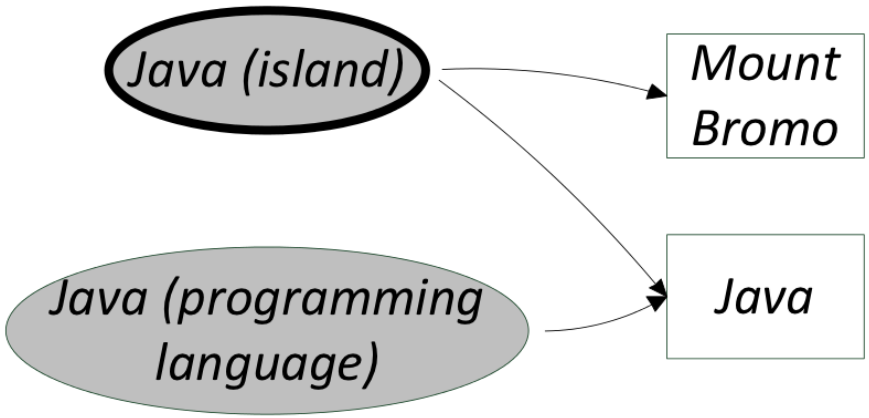
- "**Mount Bromo** is on of **Java**'s most popular tourist attractions."

Out-Degree

Nodes:

- $N = \{$ names of entities mentioned in the context $\}$
- $C = \{$ articles of candidates from the Knowledge Base $\}$

Search for each context name $n \in N$ in the article of each candidate $c \in C$, if it's found an edge is made from c to n

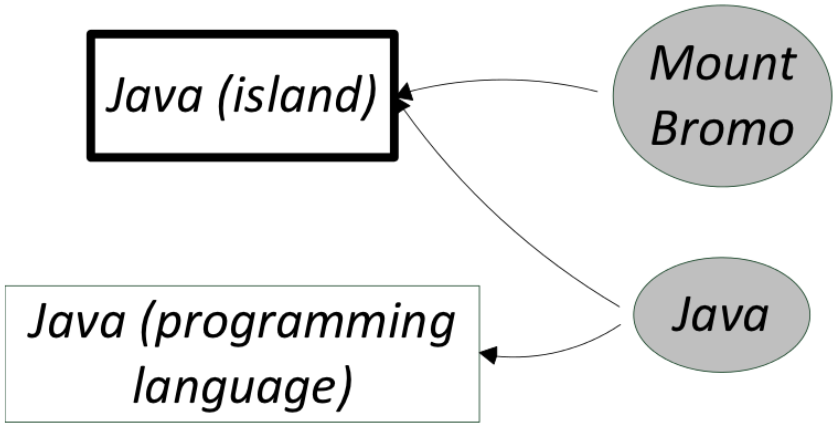


In-Degree

Nodes:

- $N = \{$ names of candidates from the Knowledge-Base $\}$
- $C = \{$ articles of entities mentioned in the context $\}$

Search for each candidate name string $n \in N$ in the article of each context name $c \in C$, if it's found an edge is made from c to n



Estimate importance

out-degree measure

$$deg_{out}(u) = |(u, v) \in E : v \in V)|$$

in-degree measure

$$deg_{in}(u) = |(v, u) \in E : v \in V)|$$

Estimate importance

out-degree measure

$$deg_{out}(u) = |(u, v) \in E : v \in V)|$$

in-degree measure

$$deg_{in}(u) = |(v, u) \in E : v \in V)|$$

combined-measure

$$(1 - \lambda)ndeg_{out}(u) + \lambda ndeg_{in}(u)$$

where,

$$ndeg_{out}(u) = \frac{deg_{out}(u)}{\sum_{u \in V} deg_{out}(u)} \quad ndeg_{in}(u) = \frac{deg_{in}(u)}{\sum_{u \in V} deg_{in}(u)}$$

Entity Disambiguation Process

- Candidate Generation: using acronyms, alternate spellings
- Candidate Ranking: using in- and out-degree

Candidate Selection

Entity can have three kinds of name variations

- Acronyms (ex: ABC vs. American Broadcasting Company)
  - Look in support document
- Aliases (ex: Robert Gates vs. Bob Gates)
  - Redirect Pages
- Alternate Spellings (ex: Air Air Macao vs. Air Macau)
  - "Air Macao site:en.wikipedia.org"

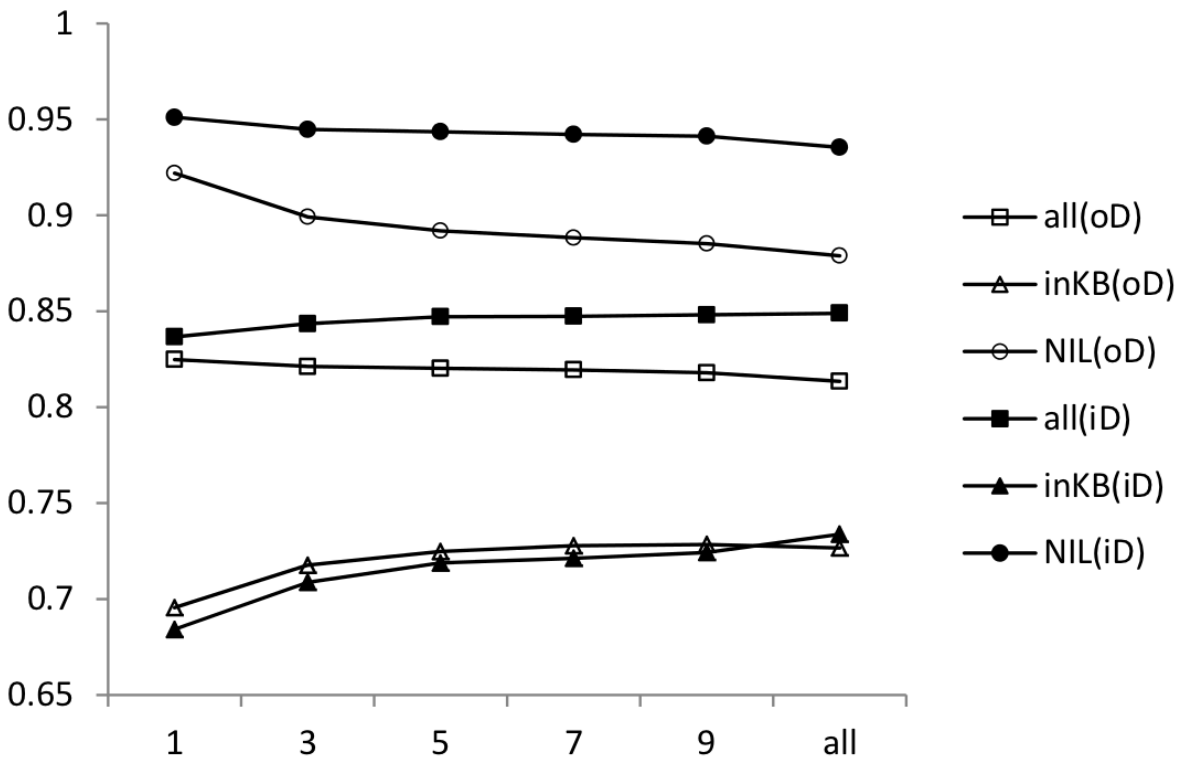
Disambiguation

- Extract names from the context
- Built graph: out- and in-degree algorithm
- Select candidate with maximum:
  - out-degree
  - in-degree
  - combined measure
- If out- and in-degree is zero, system returns NIL

Data Set: TAC-KBP track 2009 and 2010

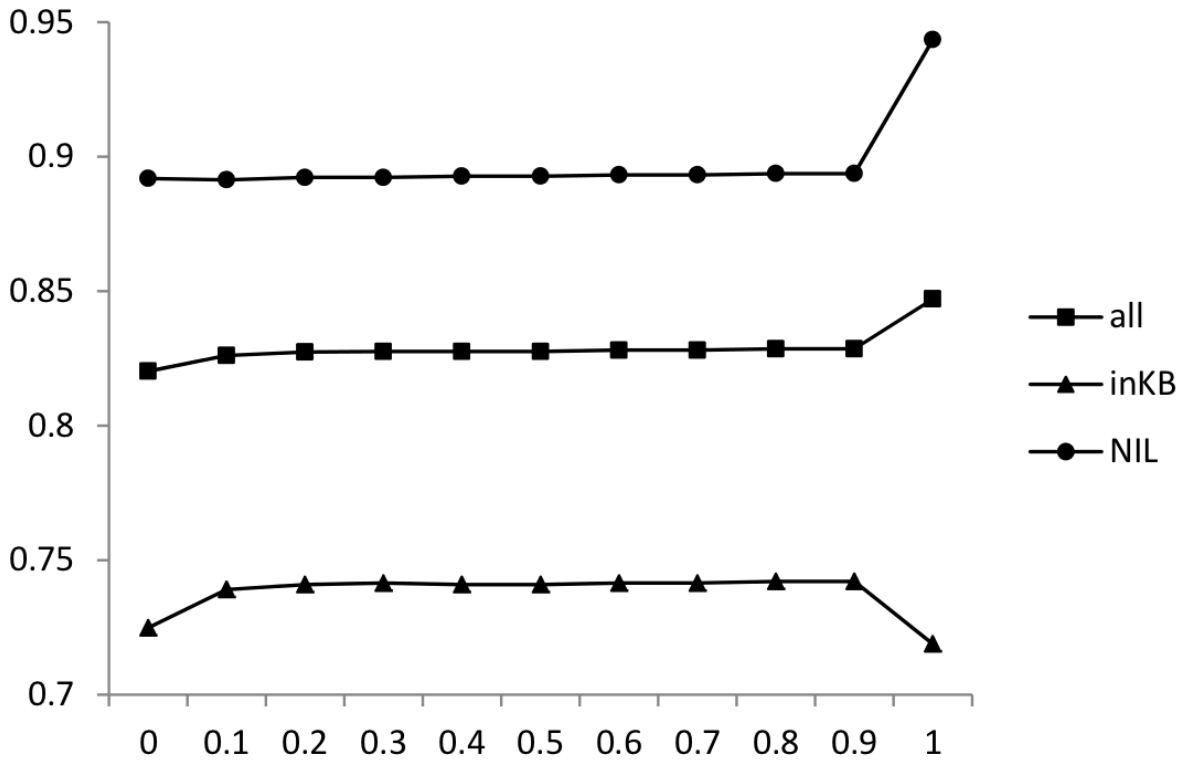
TAC-KBP Track	**2009**	**2010**
Candidates per Query	6.36	4.55
Coverage	80%	78%

Accuracy (TAC 2009): in-degree and out-degree measure



varying context window size in number of sentences

Accuracy (TAC 2009): combined measure



λ parameter variation

Results

	TAC-KBP 2009			TAC-KBP 2010		
Acc.	all	inKB	NIL	all	inKB	NIL
Rank 1	0.8217	0.7654	0.8641	0.8680	0.8059	0.9195
Rank 2	0.8033	0.7725	0.8264	0.8373	0.7520	0.9081
Rank 3	0.7984	0.7063	0.8677	0.8191	0.7373	0.8870
sLesk	0.8066	0.7075	0.8811	0.7938	0.7059	0.8667
oD	0.8248	0.6955	0.9219	0.8169	0.7059	0.9089
iD	0.8489	0.7337	0.9354	0.8240	0.7127	0.9163
Comb.	0.8276	0.7409	0.8928	0.8160	0.7402	0.8789

- accuracies of the in-degree measure are higher than the out-degree measure in all context ranges
- for non-NIL queries the combined measure performs better than the two methods
- good results, does not require training, has few parameters to tune

The End