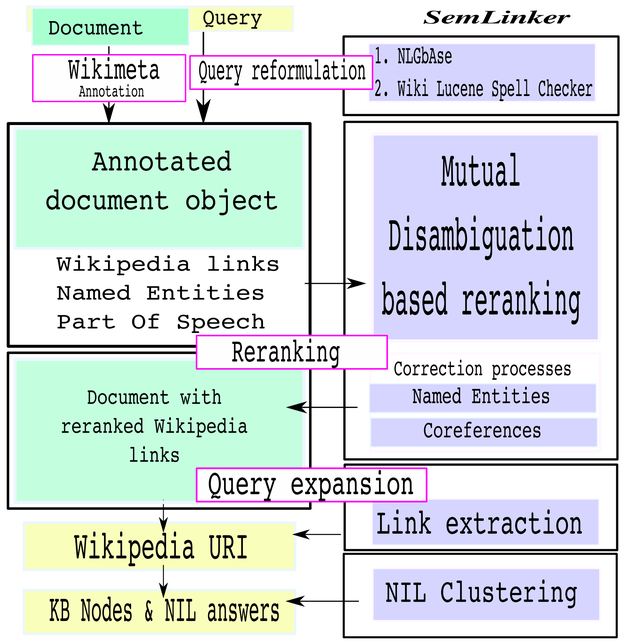
SemLinker system for KBP2013

Eric Charton[1], Marie-Jean Meurs[2], Ludovic Jean-Louis[1], Michel Gagnon[1]
[1] Polytechnique Montréal [2] Concordia University

The *SemLinker* system

SemLinker is the system presented by the Polymtl team in the English entity linking track of TAC-KBP 2013. *SemLinker* re-uses and enriches the entity links provided by a generic annotation engine.
The linking is done through a re-ranking process on the candidate links associated with a given entity. This process relies on the mutual relations between all the entities mentioned in the document.

System architecture



System Components

1. *Query Reformulation module*
2. *Mutual Disambiguation module*
3. *Link Extraction module*
4. *Clustering module*

System resources

SemLinker makes use of complementary software:

- **Wikimeta** as annotator,
- **Lucene-Search for Wiki** as spell checker,
- corpus-based resources like **NLGbAse** and a **Wikipedia dump**.

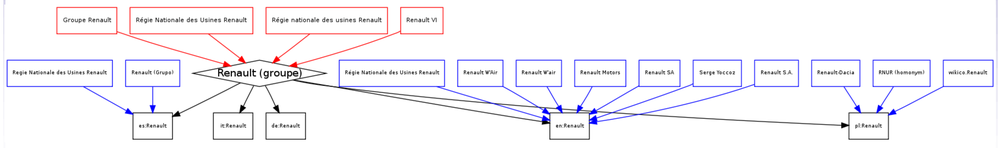
These resources are involved in various aspects of the system like:

- mention correction and reformulation process,
- entity linking annotation process.

Wikimeta output

Word	POS	NE	Semantic Link
Laura	NNP	PERS.HUM	*NORDF*
Colby	NNP	PERS.HUM	
in	IN	UNK	
Milan	NNP	LOC.ADMI	http://en.wikipedia.org/Milan

NLGbAse surface forms



Related Publications

Eric Charton, Michel Gagnon, **Disambiguation resource extracted from Wikipedia for semantic annotation**, LREC2012, Istambul, May 2012
Eric Charton, Juan-Manuel Torres-Moreno, **NLGbAse: a free linguistic resource for Natural Language Processing systems.** LREC2010, Malta, May 2010.

A disambiguation algorithm based on mutual relations of semantic annotations inside a document

Query Reformulation Module

A *Mention Correction Algorithm* to improve surface form coverage of EL systems. First strategy: automatically adding additional surface forms generated by heuristics to an existing resource of surface forms. Second strategy: adding a lexical correction step in the surface form detection process.

Algorithm

- Step 1: *Improved Surface Form Detection algorithm* tries to match surface form with NLGbAse. If match, Step 3.
- Step 2: *Surface Form Correction algorithm* tries to reformulate query using spell checker. If match, Step 1 again.
- Step 3: Rewrites query in the document if needed, and proceeds to annotation.

Improvement using mention correction

KBP2013 Results	Original Syst. B^3+F_1	Improved Syst. B^3+F_1
Overall	0.554	**0.583**
KB (in KB)	0.484	**0.543**
NIL (not in KB)	0.620	0.617
NW (news doc)	0.625	0.636
WEB (web doc)	0.574	0.586
DF (forum doc)	0.426	**0.492**
PER (person)	0.666	0.689
ORG (organization)	0.608	0.607
GPE (geopolitical entity)	0.405	0.467

Mutual Disambiguation Module

Document:

IBM has 12 research laboratories worldwide.
In 1952, Thomas J. Watson, Jr., became president of the company.

Annotation object:

NE Mentions	Candidates annotations
IBM	International Brotherhood of Magicians
	International Business Machines
Thomas J. Watson	Thomas Watson, Jr.

Direct Semantic relations:

International Business Machines ⟷ Thomas Watson, Jr (10 / 0)
International Brotherhood of Magicians ⟷ Thomas Watson, Jr (10 / 0)

Common Semantic relations:

International Business Machines ⟷ Thomas Watson, Jr {IBM 7070, Software, History of IBM, ...}
International Brotherhood of Magicians ⟷ Thomas Watson, Jr {∅}

Principle

The *SemLinker* Mutual Disambiguation module consists in 3 main steps:

1. build a set of ranked candidate annotations,
2. apply correction processes (NE and co-reference normalizations),
3. apply Mutual Disambiguation Process.

Idea:
Use all the semantic content of an annotated document to locally improve the precision of each annotation in this document.

KBP2013 results (no-wiki)

Category	Median B^3+F_1	*SemLinker* B^3+F_1
Overall	0.568	**0.583**
KB (in KB)	0.505	**0.543**
NIL (not in KB)	0.622	0.617
NW (news doc)	0.603	0.636
WEB (web doc)	0.486	0.586
DF (forum doc)	0.453	**0.492**
PER (person)	0.550	0.689
ORG (organization)	0.510	0.607
GPE (geopolitical entity)	0.488	0.467

KBP2012 development results

Category	B^3+P	B^3+R	B^3+F1
Overall	0.695	0.696	0.695
NIL	0.786	0.759	0.772
KB	0.635	0.639	0.637

Official results of the 3 best systems.

Rang B^3+F1	1	2	3
Overall	0,730	0,699	0,689
NIL	0,789	0,781	0,765
KB	0,685	0,653	0,620

Acknowledgments

This research was supported as part of Dr Eric Charton **Mitacs Elevate** Grant, sponsored by **3CE** (www.3ce.com). Participation of Dr Marie-Jean Meurs was supported by the Genozymes Project, a project funded by Genome Canada and Génome Québec. Authors would like to thank Wikimeta Technologies Inc. (www.wikimeta.com) for providing support and computing resources.