

An Experimental Study of a Hybrid Entity Recognition and Linking System

Julien Plu, Giuseppe Rizzo, Raphaël Troncy

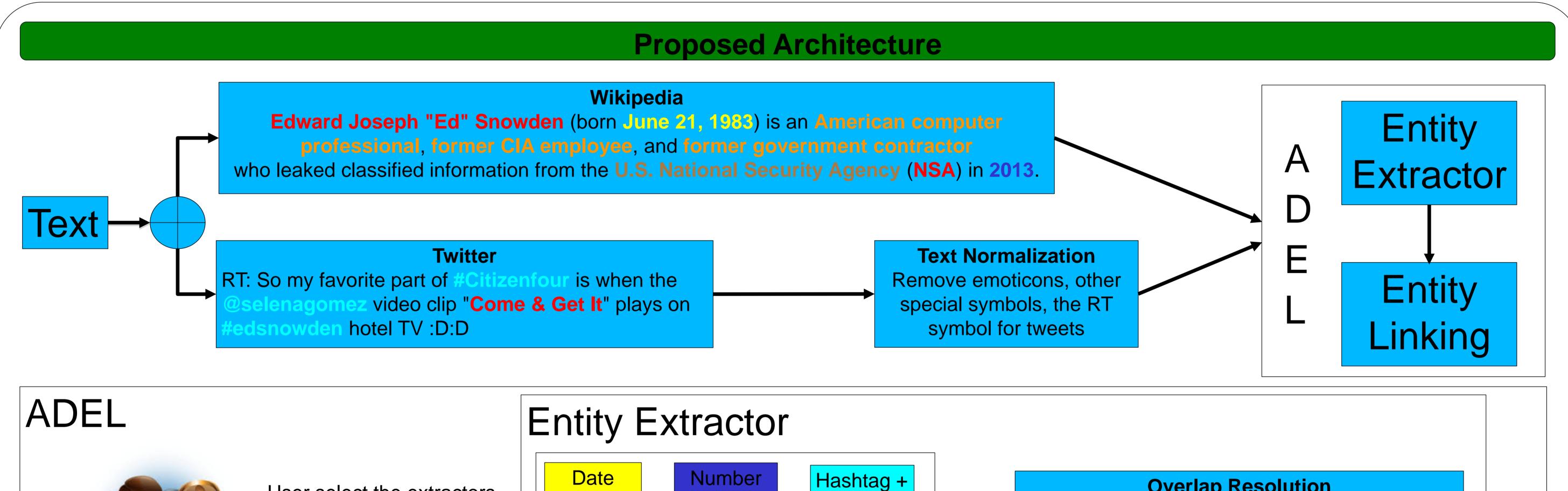
EURECOM, Sophia Antipolis, France {plu, rizzo, troncy}@eurecom.fr

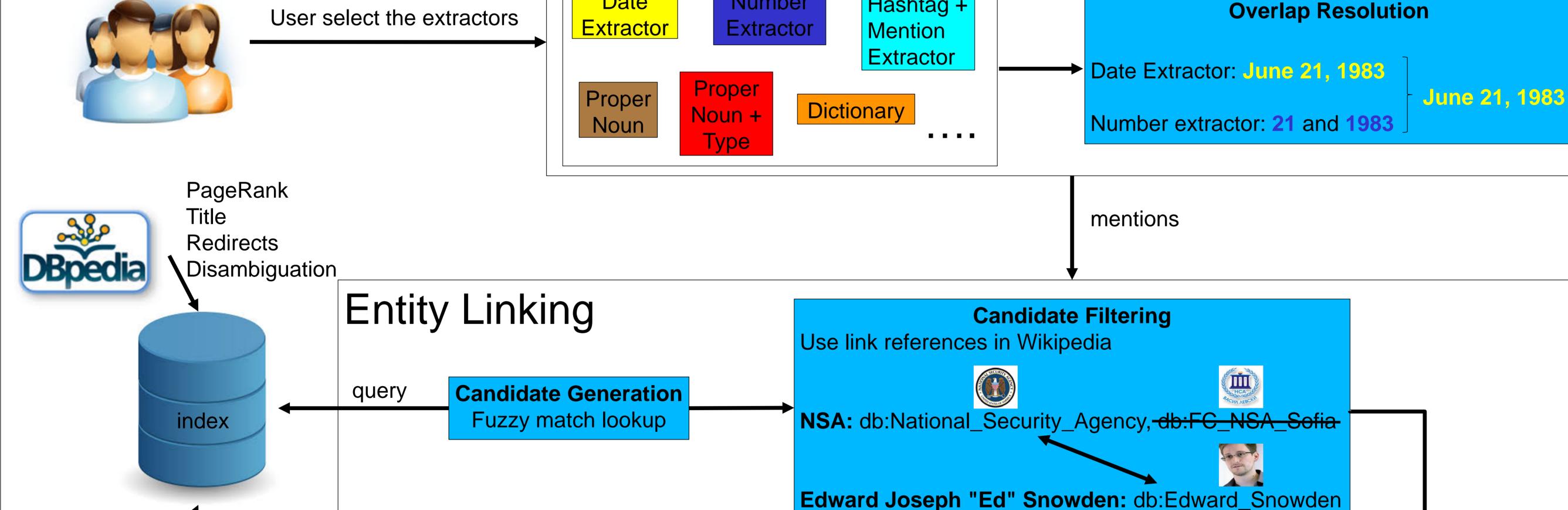


Problem

Numerous **entity linking systems**, that are generally tailored and optimized for specific data, scenario and applications This calls for a self-configurable **adaptive system**:

- Enable to process heterogeneous documents that vary:
 - in kind (e.g. tweet, news article, blog post, subtitles)
 - or language (e.g. English, French, Spanish, Chinese)
- Enable to personalize the information to extract and disambiguate:
 - common types (e.g. PER, LOC, ORG) versus specific types (Roles)
 - types interpretation (e.g. Facebook can be a Product or an Organization)
 - knowledge base (e.g. DBpedia, Wikidata, Geonames)





Scoring

Each candidate is scored using a linear formula. **L** for Levenshtein distance, **R** for set of redirects, **D** for set of disambiguation and **PR** for PageRank. **a**, **b** and **c** are weights set with **a** > **b** > **c** and **a** + **b** + **c** = 1

score(db:National_Security_Agency) = (a * L("NSA", "National Security Agency") + b * ("NSA", R("NSA")) + c * ("NSA", D("NSA"))) * PR(db:National_Security_Agency)

Evaluation

Type of extracted entities

					,		
Datasets	Co-reference	NIL entities	Type (LOC, PER,)	Date	Number	Kind of text	Nb of documents in test set
#Micropost2014	×	×	×	✓	✓	tweet	1165
#Micropost2015	X	✓	✓	×	×	tweet	2027
OKE2015	✓	✓	✓	×	×	Article (Wikipedia)	101

- 1. Similar results regardless of the kind of text
- 2. Performance at **extraction** stage similar to top state-of-the-art systems (or slightly better)
- 3. Big drop of performance at **linking** stage mainly due to an unsupervised approach

Features for each dataset

Task	Precision	Recall	F-measure		
Extraction	69.2	72.5	70.8		
Linking	47.4	45.2	46.3		
#Micropost2014					

Anchors

WIKIPEDIA

Link references

Task	Precision	Recall	F-measure
Extraction	68.4	75.2	71.6
Typing	62.8	45.5	52.8
Linking	48.8	47.1	47.9

#Micropost2015

Task	Precision	Recall	F-measure
Extraction	78.2	65.4	71.2
Typing	65.8	54.8	59.8
Linking	49.4	46.6	48

OKE2015

