Improving Entity Linking using Surface Form Refinement

Eric Charton[1], Marie-Jean Meurs[2], Ludovic Jean-Louis[1]
Michel Gagnon[1]

[1]Polytechnique Montreal, [2]Concordia University,

Montréal, QC, Canada

LREC 2014 May 26-31, 2014, Reykjavik, Iceland

Task & Context

- **Entity Linking**

Linking name mentions of named entities (NEs) found in a document to their corresponding entities in a reference Knowledge Base (KB).

- **TAC-KBP Entity Linking evaluation campaign**

Given a name (of a Person, Organization, or Geopolitical Entity) and a document containing that name, determine the KB node for the named entity, adding a new node for the entity if it is not already in the KB. The reference KB is derived from English Wikipedia.

Example

```
<query id="EL_000101">
<name>Reykjavik</name>
<docid>LREC_ENG_20140530</docid>
<beg>565</beg>
<end>574</end>
</query>
```

...

...

We attended a great conference in Reykjavik.

...

http://en.wikipedia.org/wiki/Reykjavik

Dealing with Ambiguity

Ambiguity is a key difficulty of the task.

- mentions of NEs often polysemous
- potentially related to several KB entries

make use of **surface forms** (extracted from Wikipedia)

- word or group of words e.g. *Paris*, *New York City*
- matching sequences
  - → to locate candidate entries in KB
  - → to disambiguate candidate entries

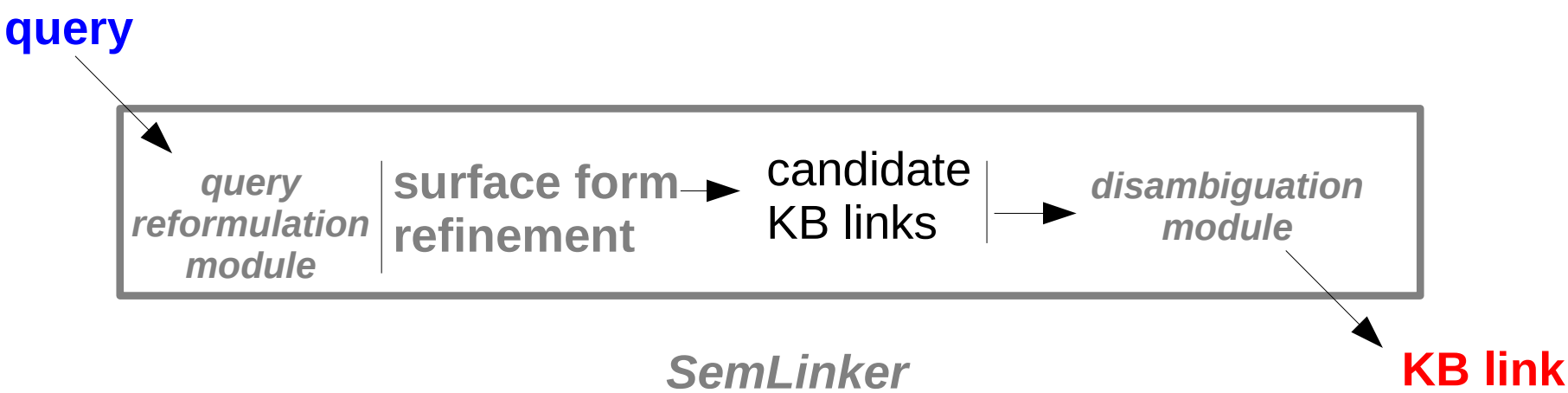
Spelling mistakes ⇒ missing or wrong identification of KB entries

Proposed Method & System Workflow

Algorithm for **surface form refinement** based on Wikipedia resources

→ correct name mentions of NEs (PER, ORG, GPE)

according to their possible sources of variations and errors.



Surface Form Matching

Surface Form Matching Problem

Surface Form Matching $\sim$ Edit Distance

- string-to-string matching

Matching of NE name mentions:

$\rightarrow$ more difficult than matching of common words

- spelling variations (e.g. transcription from an alphabet to another)
- phonetic variations (e.g. shortening)
- reduction of double names
- alternate names

Surface Form Matching for Entity Linking

String-to-string matching:

→ does not solve the surface form identification problem.

Entity Linking context:

- correct/reformulate mentions + associate them to candidate entities
  - → make use of a resource of valid surface forms for KB entities
  - → Wikipedia-based corpora (benefiting from their internal structure)

Surface Form Matching for Entity Linking

Cases for which there is **no surface form in Wikipedia-based corpora**:

- abbreviation (that refers to a NE) not existing in Wikipedia

  e.g. *JGL* for *Joseph Gordon-Levi* [EL13_ENG_0319 KBP2013]

- abbreviation existing in Wikipedia but not redirected to its entity

  e.g. *IPI* for *Intellectual Property Institute* [EL13_ENG_1604 KBP2013]

- uncommon surface form, different lexical description in Wikipedia

  e.g. *Bagdahd* for *Bagdad* [EL13_ENG_1872 KBP2013]

→ cannot be handled by approaches based on Wikipedia content only

Surface Form Refinement

Improved Surface Form Detection Module

- relies on an **enriched set of surface forms**
- surface forms from **all Wikipedia internal links** to encyclopedic docs
  → links: redirections, interwikis, disambiguation pages
- surface forms from **6 language editions** of Wikipedia
  → English, German, Italian, Spanish, Polish, French
  $\sim$ 10 million surface forms
- automatic **generation of additional surface forms** + 4 millions
  → abbreviations (e.g. JGL), alternative forms (e.g. plural), re-ordering n-grams (Barack Obama, Obama Barack)

$\Rightarrow$ **14 millions of additional surface forms**
related to at least one Wikipedia document.

Surface Form Correction Module

- **database of potential spelling errors** built from a Wikipedia dump

  $\rightarrow$ generation: Lucene-Wiki (Lucene-search extension)

- set of **rules for validating suggested corrections**

  Rule A m common words between original and suggestion

  Rule B maximum lexical distance of n letters

  Rule C Levenshtein distance under given threshold

Surface Form Refinement Algorithm

S1. submit mention to *Improved Surface Form Detection Module*. If matching surface form candidates are returned, proceed to step 3; else to step 2.

S2. submit mention to *Surface Form Correction Module*.

- If suggestions of alternative surface forms are returned, repeat step 1 using them to collect candidates.
- Else return no suggestion,and exit.

$\rightarrow$ disambiguate candidates, and select entity link.

Implementation in SemLinker

- the **Surface Form Refinement Algorithm** is integrated in the **SemLinker** system presented in TAC-KBP 2013 evaluation campaign
  - SemLinker is based on four modules:
    - → **Query Reformulation** module
    - → Mutual Disambiguation module
    - → Link Extraction module
    - → Clustering module
- our original surface form resource is **NLGbAse**, a Wikipedia-based multilingual resource. `http://www.nlgbase.org`

Experiments and Results

TAC-KBP Entity Linking task 2013

Category	All	PER	ORG	GPE	News	Web	Forum
# queries	2190	686	701	803	1134	343	713

	refSF	**QR**
Category	$B^3 + F_1$	$B^3 + F_1$
Overall	0.574	**0.596**
KB (in KB)	0.494	**0.535**
NIL (not in KB)	0.665	0.662
NW (news doc)	0.645	0.649
WEB (web doc)	0.579	0.592
DF (forum doc)	0.454	**0.508**
PER (person)	0.695	0.708
ORG (organization)	0.604	0.607
GPE (geopolitical entity)	0.440	0.486

Conclusion

Our **Surface Form Refinement Algorithm**:

- → improves the performance of our EL system
- → is generic and reusable in other (EL) systems
- → is publicly released in the SemLinker open source software

http://code.google.com/p/semlinker

All the presented experiments are fully reproducible on NIST KBP data using the SemLinker software.