# Evaluation of textual knowledge acquisition tools: a Challenging Task

Haïfa Zargayouna, Adeline Nazarenko
Yue Ma

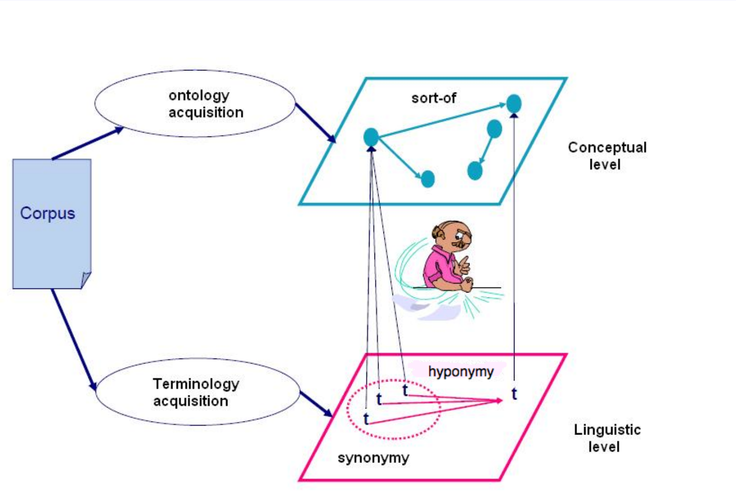
LIPN, Université Paris 13
*firstname.lastname*@lipn.univ-paris13.fr
http://www.lipn.univ-paris13.fr/∼*lastname*/

LREC 2010

# Overview of the talk

1. Context and related work
2. Difficulties
3. Propositions : similar methodology for evaluating term extraction and ontology acquisition tools
   - Task decomposition
   - Specific measures
4. Meta-evaluation

# (Terminologies – Ontologies) acquisition



# Evaluation Challenges

## Terminology acquisition

- CESART [El Hadi et al., 2006], CoRRect, NTCIR-TEMREC

## Ontology acquisition

- EON (Evaluation of Ontologies for the Web) : 2002, 2006

–> Unclear subtask definition, limited number of participants
–> No standard available benchmark, no stable quality criteria

# Why are KA tools difficult to evaluate ?

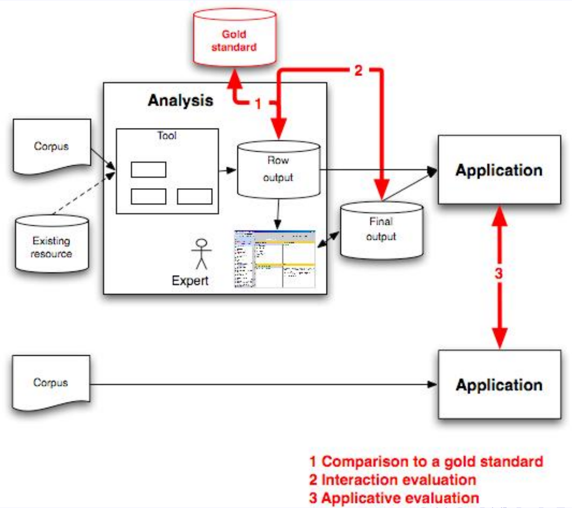
- The resulting artefacts are complex

  e.g. term extraction v.s. ontology acquistion
- Methods and goals are heterogeneous

  e.g. term numbers ? biword terms or complex terms ? size of classes ? the depth of class hierary ?
- Binary measures of relevance are inadequate

  e.g. a term candidate can be different but close to a standard term
- There exist a large variety of gold standards
- Acquisition tools are often designed to be used interactively

# Diversity of protocols



# Functional breakdown

Acquisition tasks must be decomposed into well-defined sub-tasks

- Go beyond a black-box evaluation
- Enable the comparison of heterogeneous tools
- Improve tool modularity and standardization

These sub-tasks must be evaluated independently of each other

# Simple independent functionalities

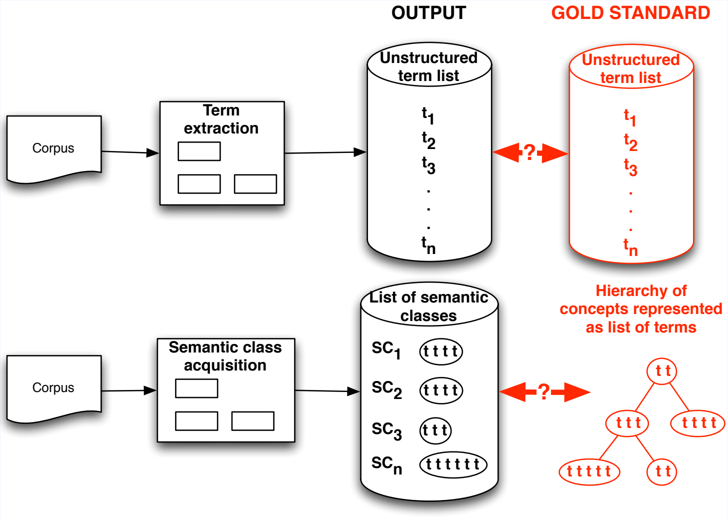
## Terminological tools

- Term extraction
- Terminological variation calculus
- Terminology structuring
- ...

## Ontology acquisition tools

- Semantic class acquisition
- Ontology structuring
- Role extraction
- ...

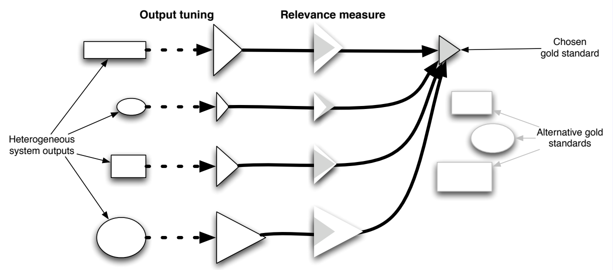
# Evaluation of term extraction and semantic class acquisition



# Specific precision and recall

- $precision = \frac{\sum_{i \in T(O)} rel_i(T(O), GS)}{|T(O)|}$
- $recall = \frac{\sum_{i \in T(O)} rel_i(T(O), GS)}{|GS|}$

$T(O)$ : Tuned output w.r.t. the chosen gold standard $GS$



# Specific precision and recall

- $precision = \frac{\sum_{i \in T(O)} rel_i(T(O),GS)}{|T(O)|}$
- $recall = \frac{\sum_{i \in T(O)} rel_i(T(O),GS)}{|GS|}$

$rel_i(T(O), GS)$ : gradual relavance between tuned output and gold standard

Limits of classic measures :

$$precision = \frac{|O \cap GS|}{|O|} \qquad recall = \frac{|O \cap GS|}{|GS|}$$

$\leadsto$These measures rely on a binary judgement, but the outputs of the systems can be close to the gold standard although not exactly alike

# Specific precision and recall : technique details

- $precision = \frac{\sum_{i \in T(O)} rel_i(T(O), GS)}{|T(O)|}$
- $recall = \frac{\sum_{i \in T(O)} rel_i(T(O), GS)}{|GS|}$

**Specificity :**

- Matching elements
- Output tuning
- Gradual relevance

# Matching elements

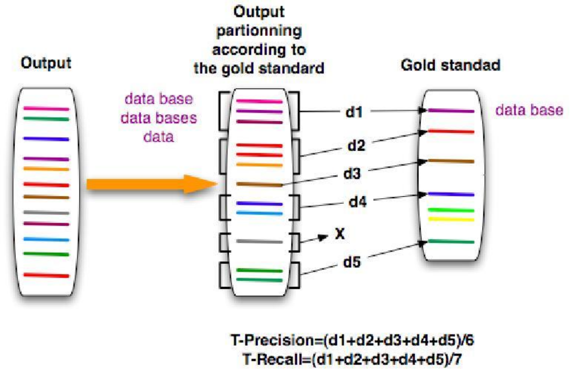
## Term matching

- Terminological distance ($d_t$) : the mean of string and complex term distances [Nazarenko & Zargayouna, 2009]
  - Simple terms : $d_s$(base, bases)=1/5=0.2
  - Complex terms : $d_c$(relational data base, data base)=0.33
- $match(e_o, e_{cgs})$ iff $e_{cgs} = \arg\min_{e_{gs} \in GS} d_t(e_o, e_{gs})$ and $d_t(e_o, e_{gs}) < \tau$

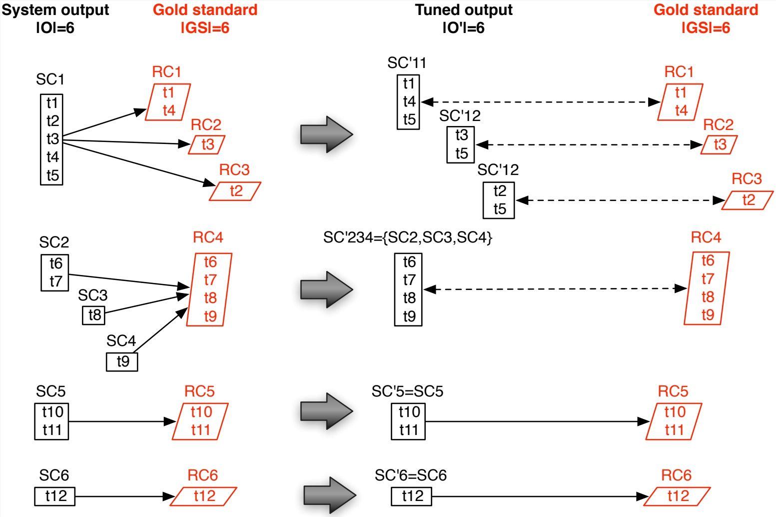
## Class-Concept matching

- Every class matches every concept
- Relevance : F-measure between extracted classes and GS concepts
  - SC={bicycle, bike} ; C={ordinary bicycle, bike, safety bicycle}
  - precision = 1/2, recall= 1/3, f-mesure = 0.39

# Output tuning : term extraction



# Output tuning : semantic class acquisition



# Gradual relevance

$$|O \cap GS| \leq \sum_{i \in T(O)} rel_i(T(O), GS) \leq min(|O|, |GS|)$$

**For term extraction**

- $rel_i(T(O))$ : the maximal value of the distances of elements of the partition

**For semantic class acquisition** (depends on the transformation step) :

- the mean relevance of merged classes
- a weighted semantic similarity measure in case of splitting

# Meta-evaluation

**Verification**

- Verify the behavior of proposed measures comparing to specifications
- Robustness of proposed measures and protocols

# 1st experiment (Term extraction)

- English corpus (Genomics, 405,000 words)
- Outputs of three term extractors
- Gold standard (*GS*) of 514 terms

| | $P$ | $R$ | $FM$ | $AP$ | $AR$ | $FM$ |
|---|---|---|---|---|---|---|
| $GS$ | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| $O_1$ | 0.71 | 0.42 | 0.52 | 0.95 | 0.48 | 0.63 |
| $O_2$ | 0.77 | 0.68 | 0.72 | 0.94 | 0.70 | 0.80 |
| $O_3$ | 0.76 | 0.28 | 0.40 | 0.95 | 0.34 | 0.50 |

Results of the output of three term extractors, $\tau = 0.4$ for terminological measures ($TP$, $TR$)

# 2nd experiment (Class acquisition)

- English corpus (volleyball, 5,078 words)
- 3 ontologies built from this corpus by master students
- Gold standard ($GS$) of 64 concepts

| | $P$ | $R$ | $FM$ | $AP$ | $AR$ | $FM$ |
|---|---|---|---|---|---|---|
| $GS$ | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 | 1.0 |
| $O_1$ | 0.4 | 0.4 | 0.4 | 0.83 | 0.47 | 0.60 |
| $O_2$ | 0.46 | 0.45 | 0.45 | 0.84 | 0.47 | 0.60 |
| $O_3$ | 0.34 | 0.36 | 0.34 | 0.81 | 0.37 | 0.51 |

Results of the evaluation of three ontologies

# Conclusion

A common approach

- Evaluation of elementary functionalities
- Specific measures based on gradual relevance and output tuning

Measures closer to human intuition

- Same ranking than with classical Precision and Recall
- Higher values

Perspectives

- Challenges within the Quaero program
- Evaluation protocols for other acquisition tasks

# Distance between terms

$$d_t = (d_s + d_c)/2$$

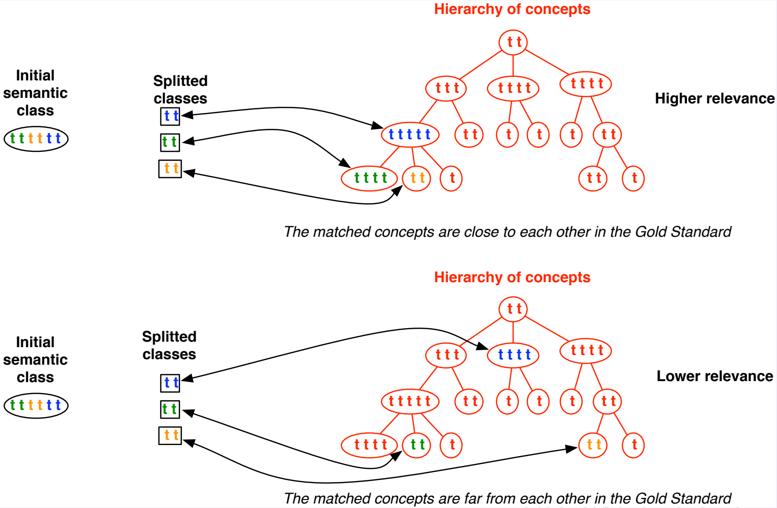
Simple terms : $d_s$(base, bases)=1/5=0.2

- String distance based on character comparison
- Edition distance between strings (character insertion & deletion)
- Normalisation on string length (# characters)

Complex (multi-word) terms : $d_c$(relational data base, data base)=0.33

- Best matching between the words of the terms
- String term distance between the matching pairs
- Edition distance between complex terms (word insertion & deletion, taking into account the string distance of the matching pairs)
- Normalisation on term length (# words)

# Splitting (1)


Hierarchy of concepts
Initial semantic class
Splitted classes
Higher relevance
The matched concepts are close to each other in the Gold Standard
Hierarchy of concepts
Initial semantic class
Splitted classes
Lower relevance
The matched concepts are far from each other in the Gold Standard

# Splitting (2)

- Selection of the central concept ($p$)

  $p = \arg\max_{c \in GS} fm(e_o, c)$ where $e_o$ is the element of the initial output from which $e'_o$ is derived by splitting.

- Similarity measure [Wu & Palmer, 1994] between two concepts where $C$ is the closest common ancestor of $p$ and $e_{gs}$, $depth(X)$ et $depth_Y(X)$ are resp. the distance from $X$ to the root of the ontology and the distance from $X$ to the root by way of $Y$

- Relevance of a splitted class ($e'_o$) wrt. GS

  $$rel_{GS}(e'_o) = fm(e'_o, e_{gs}) * Sim(p, e_{gs})$$

# References I


- Widad Mustafa El Hadi, Ismail Timimi, Marianne Dabbadie, Khalid Choukri, Olivier Hamon and Yun-Chuang Chiao
  *Terminological Resources Acquisition Tools : Toward a User-oriented Evaluation Model.*
  LREC, 2006.

- Adeline Nazarenko and Haïfa Zargayouna
  *Evaluating term extraction*
  RANLP, 2009

- Zhibiao Wu and Martha Palmer
  *Verb Semantics and Lexical Selection*
  ACL, 1994