***Antonio Toral**
^Rafael Muñoz
*Monica Monachini

*Istituto di Linguistica Computazionale (Pisa, Italy)
^University of Alicante (Spain)

Named Entity WordNet

LREC 2008
O12 - Named Entity Recognition

Marrakech, 2008-05-28

Outline

- Intro
  - Named Entities (NEs)
  - Language Resources (LRs)
  - Why NEs in LRs?
  - How to enrich LRs with NEs?
- Named Entity WordNet
  - Mapping & Disambiguation
  - Article extraction
  - NE identification
  - NE repository
- Conclusions & Future

NEs

- Usually refer to
  - Proper nouns: names of people, locations, organizations, ...
  - Numerical expressions: time, amounts, ...
- Important for NLP tasks
  - NEs: 10% of text + carry important semantic info
- Different sets of NE categories
  - ConLL -> flat, 4 types (per, org, loc, misc)
  - Sekine -> hierarchy, +100 subtypes

LRs

- Manually created by expert lexicographers
- Broad-coverage resources
  - Common nouns, adjectives, verbs, adverbs
- Rich Semantic Info (relations, roles, etc)
- WordNet
  - +100k word senses

LRs

- Manually created by expert lexicographers
- Broad-coverage resources
  - Common nouns, adjectives, verbs, adverbs
- Rich Semantic Info (relations, roles, etc)
- WordNet
  - +100k word senses
- LRs lack info about NEs
  - “building a proper noun ontology is more difficult than building a common noun ontology as **the set of proper nouns grows more rapidly**” (Mann, 2002)

Why NEs in LRs?

- Stored Knowledge can be applied to NLP tasks
- E.g. Question Answering
  - Question (CLEF 2006)
    - Who is Vigdis Finnbogadottir?
  - QA system
    - Linguistic analysis of text [S. Ferrandez et al. 06]
      - “[...] presidents: Vigdis Finnbogadottir (Iceland), [...]”
    - Solution (wrong): Iceland

Why NEs in LRs?

- Stored Knowledge can be applied to NLP tasks
- E.g. Question Answering
  - Question (CLEF 2006)
    - Who is Vigdis Finnbogadottir?
  - QA system
    - Linguistic analysis of text [S. Ferrandez et al. 06]
      - "[...] presidents: Vigdis Finnbogadottir (Iceland), [...]"
    - Solution (wrong): Iceland
  - Possible related knowledge in LR
    - "Vigdis Finnbogadottir" instance_of: "president", "icelandic", "female head of state"
  - LR can be useful within QA, for example to:
    - Find answers
    - Validate answers

How to enrich LRs with NEs?

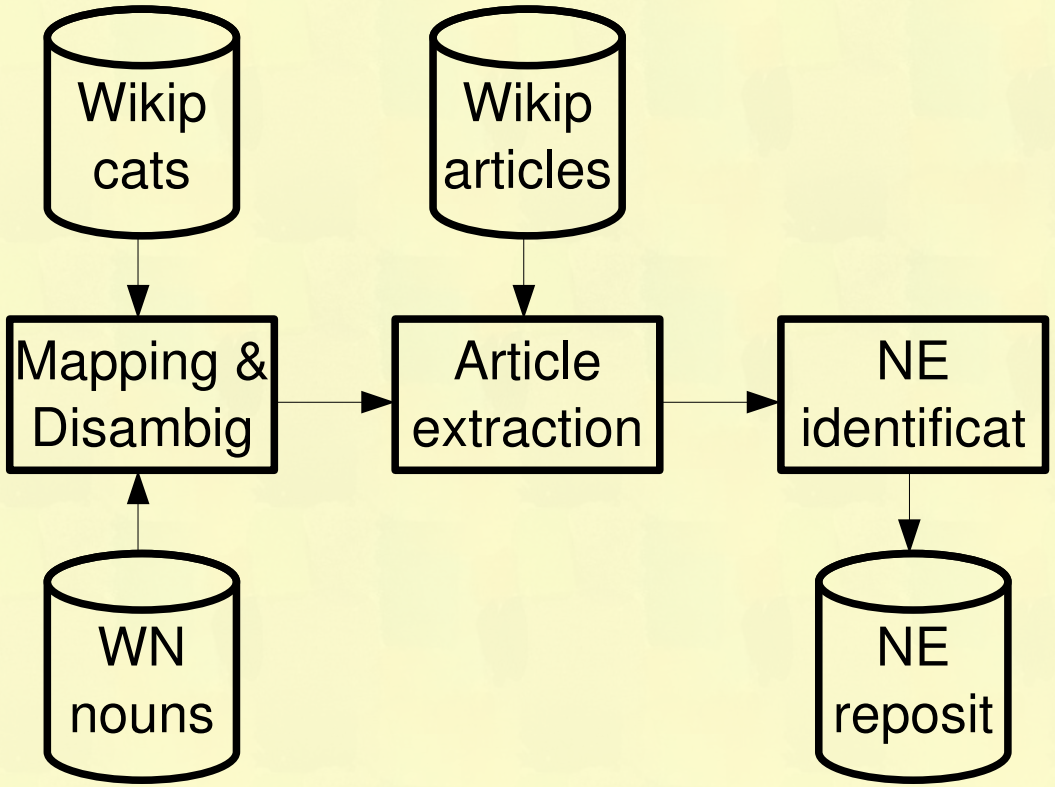
- NEs should be acquired & introduced automatically
- Ideal Source
  - Up-to-date
  - High Coverage
  - Allow a Good Quality Extraction

How to enrich LRs with NEs?

- NEs should be acquired & introduced automatically
- Ideal Source
  - Up-to-date
  - High Coverage
  - Allow a Good Quality Extraction
- Wikipedia
  - Dynamic source
  - Huge amount of NEs
  - Some degree of structure

Named Entity WordNet

- Automatically Extend WordNet with NEs extracted from Wikipedia



Mapping

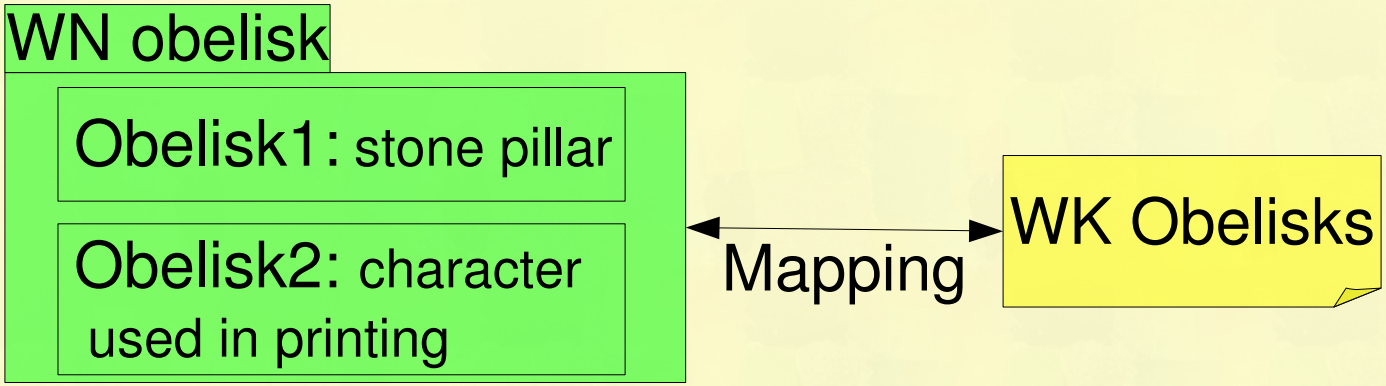
- Map lemmas
  - WordNet: noun classes (instantiated)
  - Wikipedia: categories
- Results

		Wikipedia dump date		
		200704	**200711**	**200801**
Synsets	**Total**	893		
	Mapped	513	536	541
	%	57.44%	60.02%	60.58%

- Analysis (non mapped)
  - 75% no matching category but matching article
  - 13% no matching category nor matching article
  - 10% matching category but PoS error

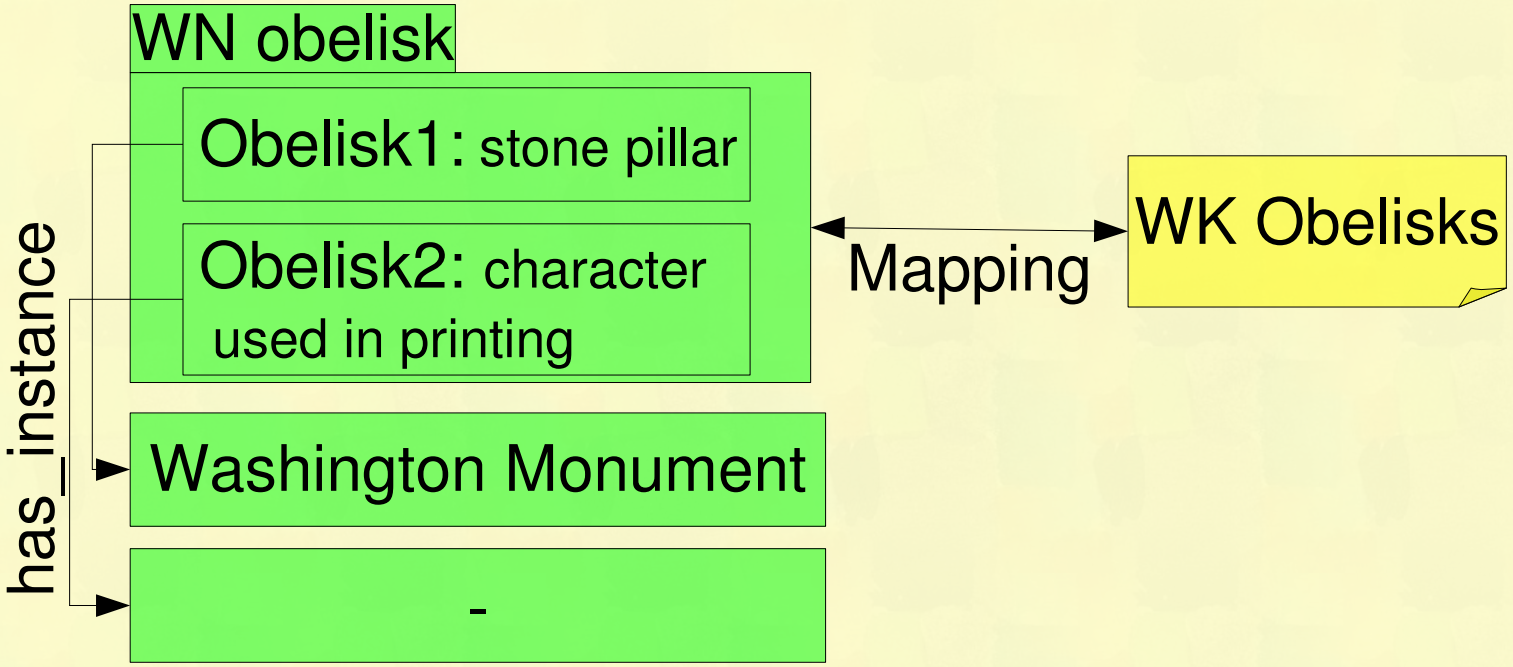
Disambiguation

- WordNet polysemous nouns to Wikipedia categories
  - Intersection of instances



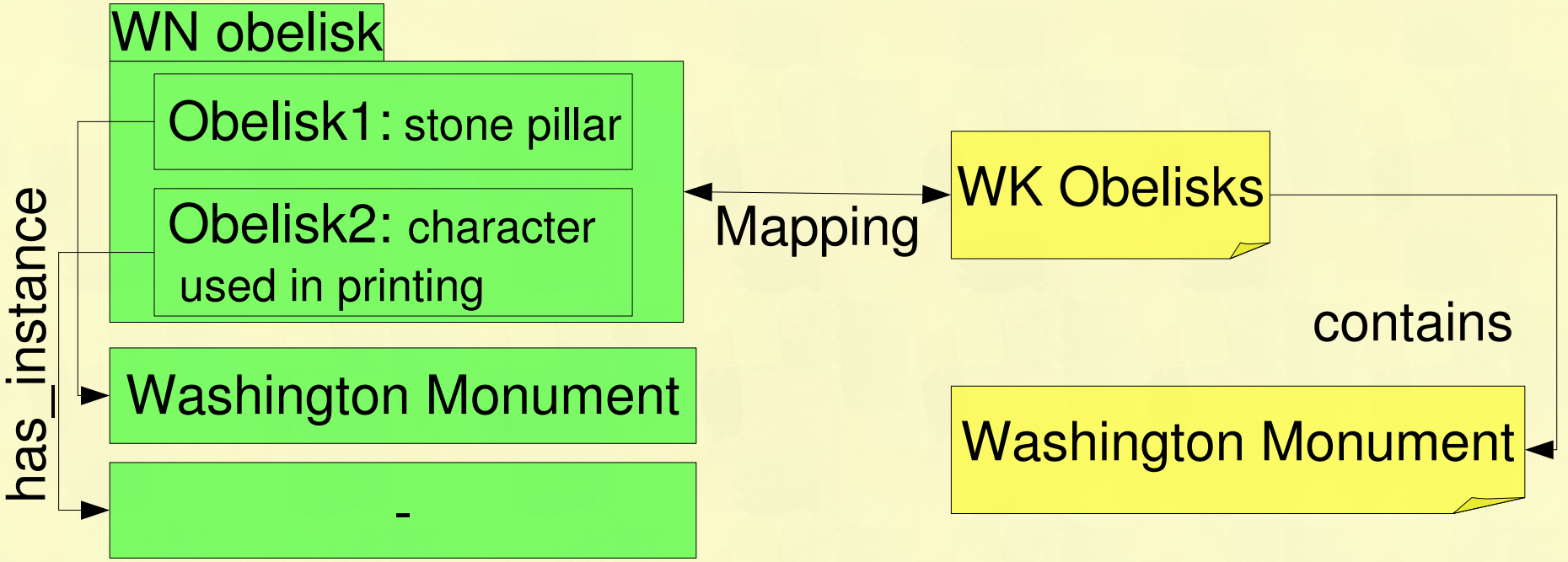
Disambiguation

- WordNet polysemous nouns to Wikipedia categories
  - Intersection of instances



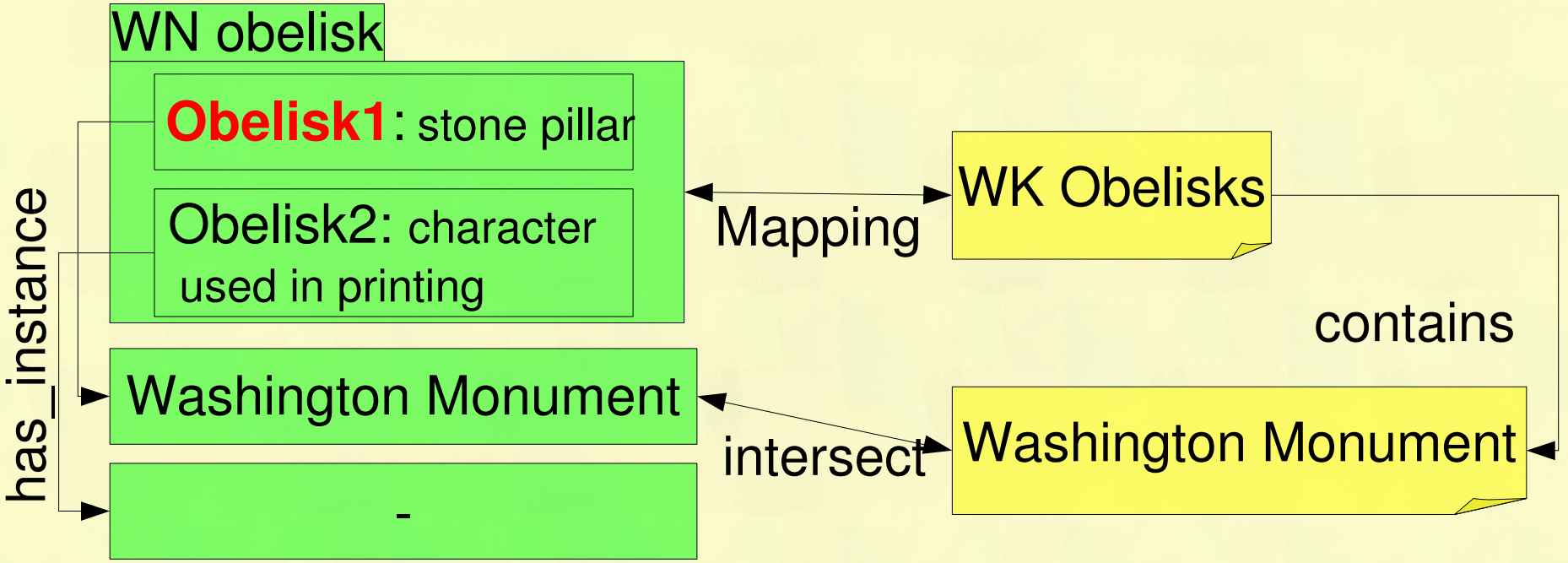
Disambiguation

- WordNet polysemous nouns to Wikipedia categories
  - Intersection of instances



Disambiguation

- WordNet polysemous nouns to Wikipedia categories
  - Intersection of instances



  - Results (262 words): 100% precision, 39% recall
  - Analysis non disambiguated words:
    - 78% no common instance found
    - 22% no sense corresponds to category

Article extraction

- For each category mapped (and its hyponyms*) fetch:
  - Titles
  - Abstracts
  - Variants
- *Hyponym identification (subcategories)
  - ^ category (“ by “ | “ of “ | “ in “ | “ stubs$”)
    - Obelisks in Argentina
  - ^ (JJ|JJR|NN|NP)+ (CC(JJ|JJR|NN|NP)+)* “ “ category$
    - Ancient obelisks

NE identification

- An extracted article might be a NE or a common noun
  - Look for occurrences of its title in its body text & check capitalisation (Bunescu & Pasca 2006)
  - Not only in the English Wikipedia, but in 10 Wikipedias for langs that follow these caps. norms
    - Text size to look for occurrences bigger -> results more representative
    - Language independent -> whatever the language we obtain the article equivalent in these languages

NE identification

- An extracted article might be a NE or a common noun
  - Look for occurrences of its title in its body text & check capitalisation (Bunescu & Pasca 2006)
  - Not only in the English Wikipedia, but in 10 Wikipedias for langs that follow these caps. norms
    - Text size to look for occurrences bigger -> results more representative
    - Language independent -> whatever the language we obtain the article equivalent in these languages
- Results
  - Only English -> F 78.06%, P 73.91%, R 87.93%
  - 10 languages -> F 82.26%, P 79.69%, R 87.93%

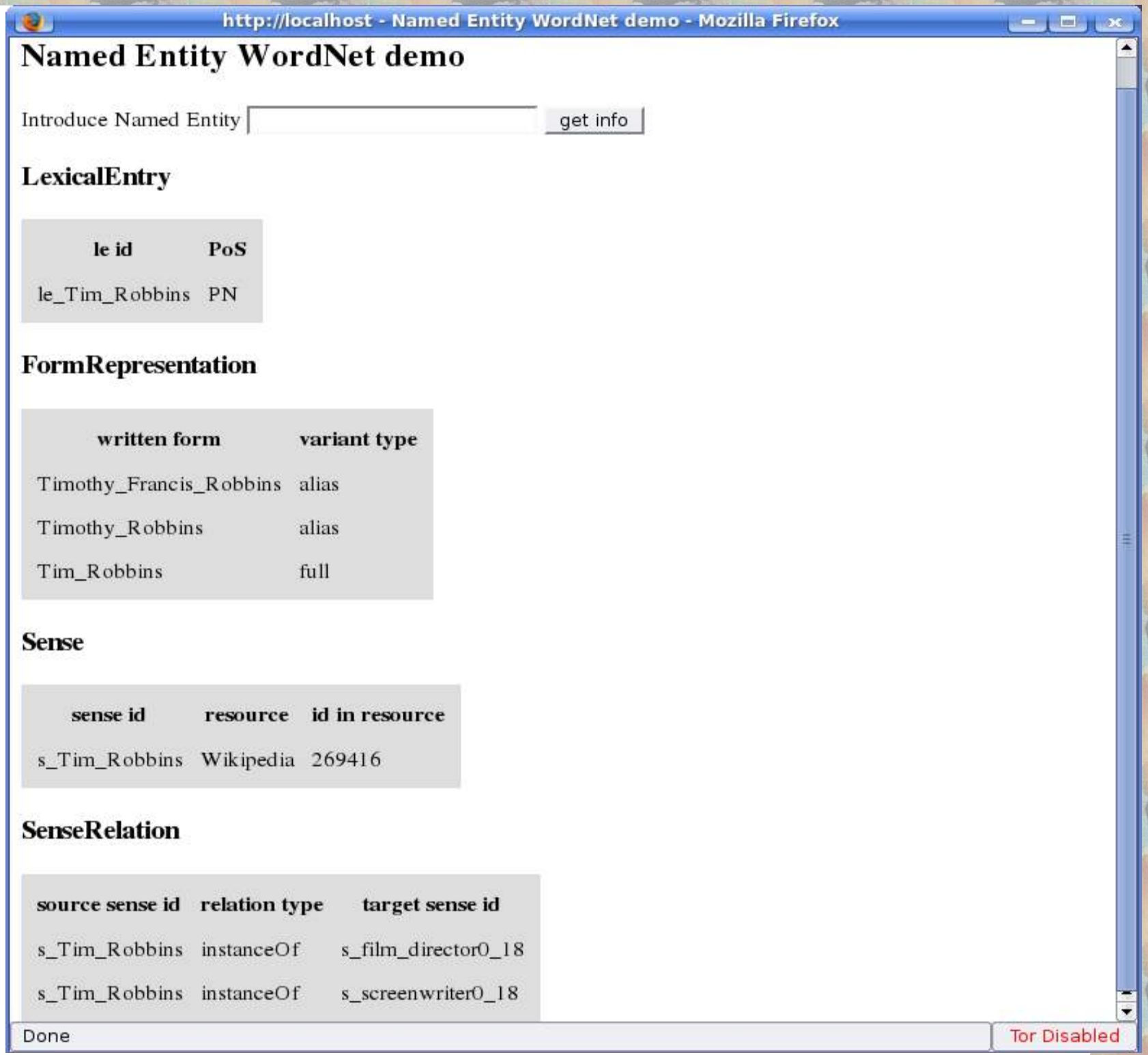
Extracted NEs

- General
  - 310,742 Nes, 452,017 variants, 381,043 instance rels
- Detailed (per lexicographic file)

Lex File	Nes	Example
act	4,214	Project_Pluto instanceOfproject0_4
artifact	23,878	Akinada_Bridge instanceOf suspension_bridge0_6
communication	1,973	Flower_of_Scotland instanceOf national_antherm0_10
event	58	Sino-Soviet_split instanceOf schism0_11
group	1,216	Medici instanceOf family0_14
location	43,582	Incense_Route instanceOf trade_route0_15
object	28,180	Pyxis instanceOf constellation=_17
person	277,941	Vladimir_Kotelnikov instanceOf electrical_engineer0_18

NE repository

- Elements: NEs, classes, relations, variants, definitions
- LMF compliant: ISO standard for lexicons
  - Independent from specific LRs
- Web test & download
  - dlsi.ua.es/~atoral/#Resources
  - www2.ilc.cnr.it/ne-repository


http://localhost - Named Entity WordNet demo - Mozilla Firefox
Named Entity WordNet demo
Introduce Named Entity
get info
LexicalEntry
le id
PoS
le_Tim_Robbins
PN
FormRepresentation
written form
variant type
Timothy_Francis_Robbins
alias
Timothy_Robbins
alias
Tim_Robbins
full
Sense
sense id
resource
id in resource
s_Tim_Robbins
Wikipedia
269416
SenseRelation
source sense id
relation type
target sense id
s_Tim_Robbins
instanceOf
s_film_director0_18
s_Tim_Robbins
instanceOf
s_screenwriter0_18
Done
Tor Disabled

Conclusions & Future

- High Quality & Large NE extension of WordNet
  - +310k Nes (it had 7k), +380k relations
  - Standard-compliant output
- Future
  - Apply to other LRs for different languages
    - Empirically demonstrate generality of the approach
    - Derive a Multilingual NE repository
  - Exploit Textual Entailment to disambiguate mapping

End

Thanks for your attention!

Questions?