2009/10/01
Centre for Next Generation Localisation
Dublin City University

Web 2.0, Language Resources and standards to automatically build a multilingual Named Entity Lexicon

Antonio Toral
{fistname.lastname}[at]ilc.cnr.it
Istituto di Linguistica Computazionale - CNR, Pisa (Italy)
Dep. Llenguatges i Sistemes Informàtics - Universitat d'Alacant (Spain)

Outline

- Intro
- Knowledge Acquisition bottleneck
- NEs, why NEs?
- Building MINELex
- Applying this to the real world
- Conclusions

Intro

- NLP: necessary to deal with huge amount of digital information
  - IR, IE, QA, MT, ...
- World knowledge: required for the semantic level
  - Conceptualisations of reality: from definition of ontology in Ancient Greece to KBs in AI, LRs in CL
  - LRs extensively applied in NLP but...

Intro

- … but LRs are expensive to build
- Much effort devoted during last 15 years
  - WordNet, EuroWordNet, SIMPLE, …
  - Rich semantic info (relations, roles,..)
- Enough coverage?
  - ~OK -> verbs, adjs, advs, common nouns
  - ¬OK -> Named Entities, domain terms, multiwords, …

Outline

- Intro
- **Knowledge Acquisition bottleneck**
- NEs, why NEs?
- Building MINELex
- Applying this to the real world
- Conclusions

Knowledge Acq. bottleneck

- “humans cannot manually structure the available knowledge at the same pace as it becomes available” (Philpot 05)
  - Automatic procedures needed!
- Step forward -> 3 “ingredients”
  - Web2.0 (vs MRDs and corpora)
  - LRs
  - Standards / Interoperability

MRDs

- Explicit structure
  - Facilitates extraction
  - ACQUILEX (89-92)
    - Extraction syntactic, semantic, taxonomies, ...
- Small and fixed size
  - Research moved to corpora in the 90s (Hearst 92)...

Corpora

- Does not suffer from size problem
- No structure
- Subjectivity
  - Relations in corpora more subjective than those in dicts/encyclopedias (Hearst 98)
  - 44% sentences subjective in non opinion pieces of WSJ (Wiebe 04)
- Detect lexical variability?
  - Bill Clinton, William Clinton (Fleischman 03)

New Text

- Emerged with Web 2.0
  - Wikis, blogs, folksonomies
- Some structure -> facilitates extraction
- Dynamic -> up-to-date knowledge
- Collaboratively built -> lang variety

Wikipedia

- Multilingual, collaborative encyclopedia
  - Structure: pages, categories, inter e intralingual links, redirects, infoboxes
- Quality comparable to traditional ones (Britannica, Brockhaus)
- Interests in CL community
  - WiQA, GikiCLEF, *New Text*, WikiAI08, ...
  - APIs: JWPL, JWKTL
  - Applied to NLP: QA, IE, MT, ...

MRDs vs Corpora vs Web2.0

	MRDs	corpora	wikis
size	small	~unlimited	big
subjectivity	~none	high	low
structure	high	none	medium
dynamic	no	no	yes

Language Resources

- Result of many man-years expert work
- Pay off effort <-> applicability?
  - Political, distribution, technical
- Technical
  - Formal validation
  - Automatic access methods (APIs)
  - Formalisation -> support reasoning

Interoperability

- Historical differences in formats, structures, semantic of categories in LRs
  - Barrier for sharing / interchanging
  - Lack coordination, competing practices, ...
- Inititatives
  - EAGLES, ISLE, PAROLE, ...
  - LMF, MAF, SynAF, SemAF, ...

Outline

- Intro
- Knowledge Acquisition bottleneck
- **NEs, why NEs?**
- Building MINELex
- Applying this to the real world
- Conclusions

NEs

- Usually refer to
  - Proper nouns: names of people, locations, organizations, ...
  - Numerical expressions: time, amounts, ...
- Important for NLP tasks
  - NE Recognition, subtask of Information Extraction

Why NEs?

- 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 02)
- NEs provide salient clues and have a special role in translation
- Stored Knowledge can be applied to NLP tasks

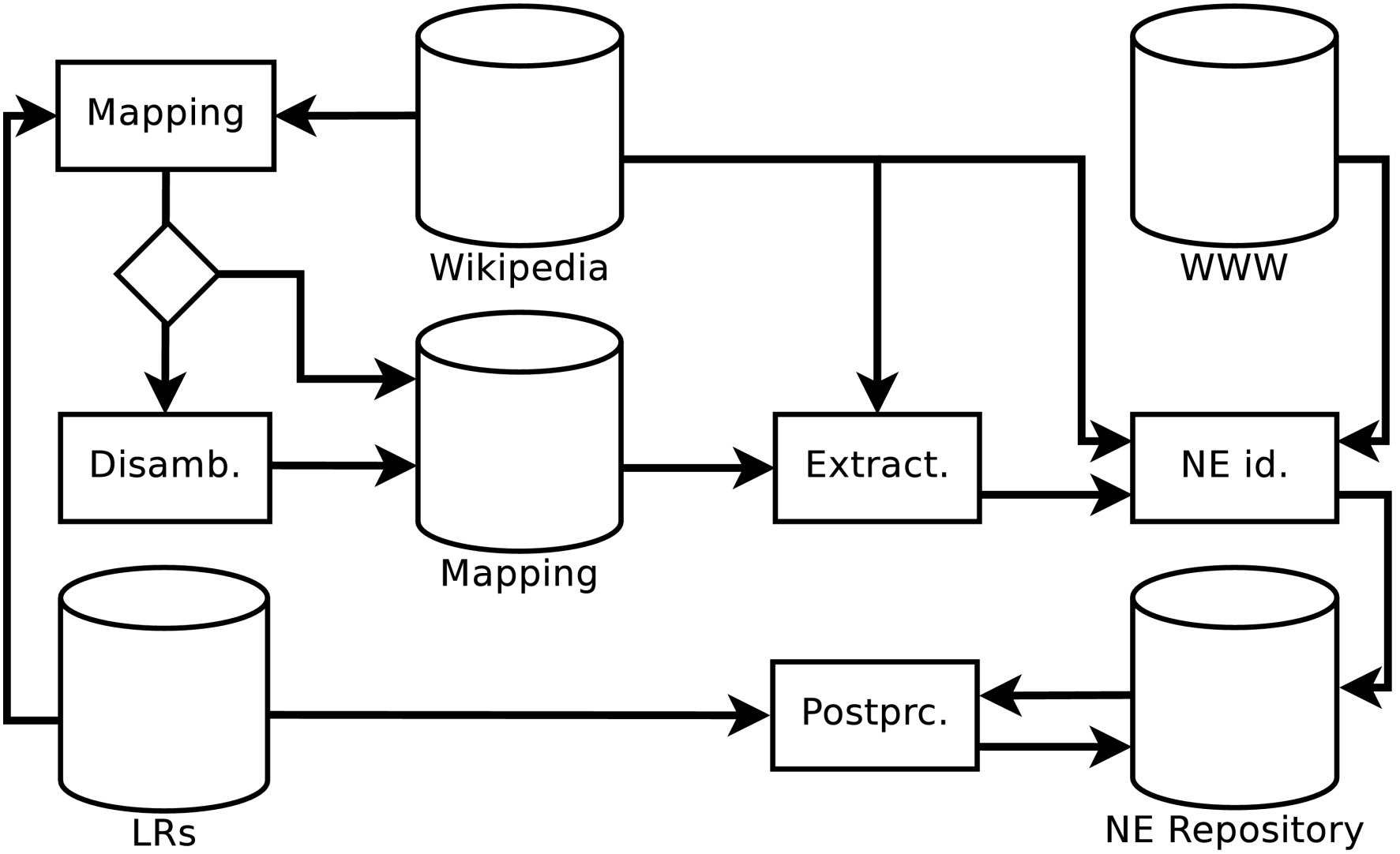
Why NEs?

- E.g. Question Answering
  - Who is Vigdis Finnbogadottir?
  - QA system
    - Linguistic analysis of text (Ferrandez 06)
      - “[...] presidents: Vigdis Finnbogadottir (Iceland), [...]”
  - Solution: Iceland
  - Possible related knowledge in LR
    - “Vigdis Finnbogadottir” instance_of: “president of Iceland”, “icelandic”, “female head of state”
  - LR can be useful within QA, for example to:
    - Find and validate answers

Outline

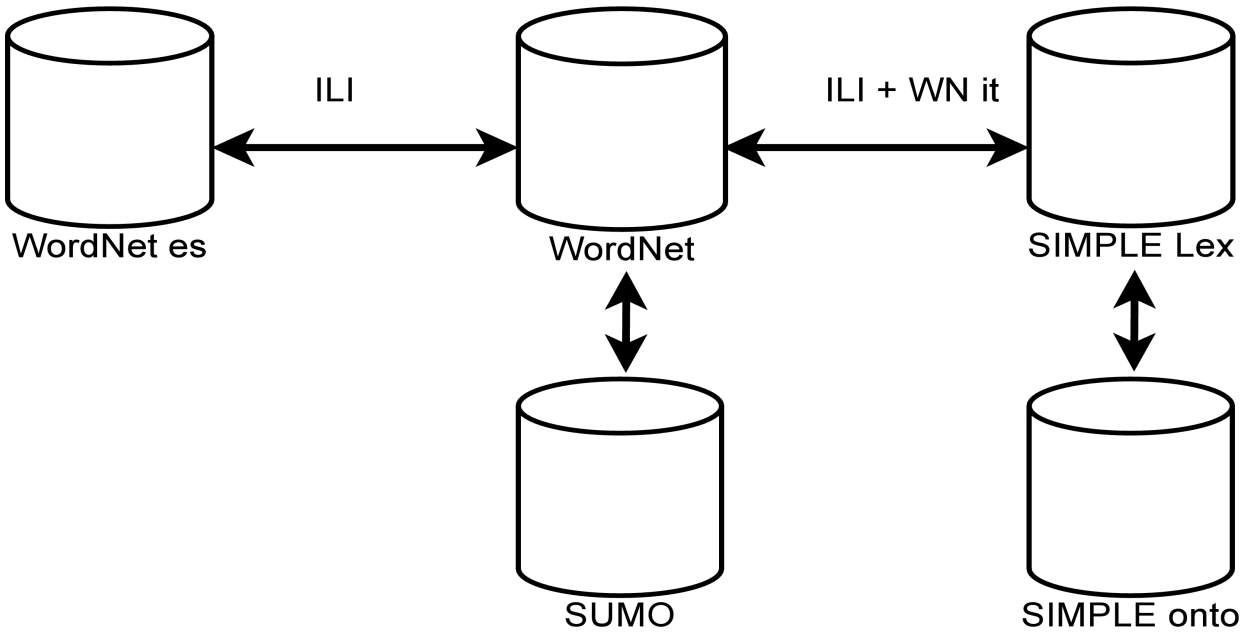
- Intro
- Knowledge Acquisition bottleneck
- NEs, why NEs?
- **Building MINELex**
- Applying this to the real world
- Conclusions

MINELex


Mapping
Wikipedia
WWW
Disamb.
Mapping
Extract.
NE id.
LRs
Postprc.
NE Repository

MINELex

- LRs used...
  - WordNet en&es, SIMPLE it



Mapping

- Establish an initial link LR<->WK
- Lemma (PoS tagger)
  - actor <-> Actors, aquarium <-> Aquaria
- 65.4% (57.4% older dump) of target synsets linked. The rest?
  - 75% no matching category but page
  - 13% nothing
  - 10% PoS errors
  - 2% B.E. vs A.E.

Disambiguation

- A mapping might be ambiguous
  - e.g. obelisk -> Obelisks
    1. Stone pillar
    2. Character used in printing
- Simply take the MFS: ~65% acc. (YAGO)
- or... Automatic disambiguation
  - Instance intersection
  - Text similarity

NE Intersection

- Look for common NEs in WN and WK hyponymy chains
  - WN obelisk, WK Obelisks
    - WN obelisk1 -> {Washington Monument, ...}
    - WN obelisk2 -> {Ø}
    - WK Obelisks -> {Washington Monument, ...}
  - Disambiguation: WN sense1 <-> WK cat
- Eval set: 260 polysemous words
  - 100% P, 39% R
    - Few instances in WN...

Text similarity

- Exploit definitions to disambiguate

<word id="obelisk">

<sense number="1">a stone pillar having a rectangular cross section tapering towards a pyramidal top</sense>

<sense number="2">a character used in printing to indicate a cross reference or footnote</sense>

<category id="Obelisks">An obelisk (Greek ὀβελίσκος , diminutive of ὀβελός , "needle") is a tall, narrow, four-sided, tapering monument which ends in a pyramidal top. Ancient obelisks were made of a single piece of stone (a monolith). </category>

</word>

Text similarity

- Set of representative methods
  - Textual Entailment system
  - Personalised PageRank
  - Semantic Vectors
  - Baselines
    - MFS, word overlap
  - Combinations
    - oracle, unsupervised, supervised, voting

Textual

- TE task: two texts (Hypothesis and Tesis) -> Decide if H entails T
- TE system used: several Inferences
  - Lexical distance measures: Euclidean, Smith-Waterman, ...
  - Semantic: WordNet similarity measures, verbs' similarities, ...
- Apply bidirectionally:
  - WordNet noun might imply Wikipedia category or the other way

Personalised Pagerank

- Graph-based algorithm over LR
  - Represents WordNet as graph
  - For each text computes PPR over graph, producing probability distribution over synsets
  - Compares how similar these two discrete probability distributions are by encoding them as vectors and computing the cosine between the vectors

Semantic Vectors

- LSA-like
- Tokenisation and indexing (term document matrix) by using Lucene
- Creates WORDSPACE model from matrix
- Uses Random Projection to perform dimension reduction
- Corpus for this task:
  - Glosses from WordNet (117,598)
  - Abstracts from WK 01/2008 (2,179,275)

Text similarity

Run	Accuracy
MFS	64.7%
Word Overlap	56.3%
Word Overlap (no stop words)	62.7%
Semantic Vectors	54.1%
PPR	61.8%
PPR (no stop words)	64.3%
TE (trained on TE corpus)	52.8%
TE (no training)	64.7%
TE (supervised)	77.74%

Text similarity

Run	Accuracy
Oracle (PPR+SV+TE+WO)	84.5%
Voting (PPR+SV+TE+WO)	66.5%
Voting (PPR+TE+WO)	68%
Unsupervised (PPR+SV+TE+WO)	65.2%
Unsupervised (PPR+TE+WO)	65.7%
Supervised (PPR+SV+TE+WO)	77.24%
Supervised (PPR+TE+WO)	77.11%

Extraction

- For each category mapped (and its subcategories*) extract articles and related data
  - Redirects: variants
  - Abstracts: definitions
  - Equivalents in other langs
- Subcategories not always hyponyms
  - Philosophers -> Philosophers by era
  - Philosophers -> Timelines of philosophers

Extraction

- Hyponym identification
  - Morphosyn. patterns
    - ^ category (“ by “ | “ of “ | “ stubs$”)
      - philosophers by nationality, philosophers of mind
    - ^ (JJ|JJR|NN|NP)+ (CC(JJ|JJR|NN|NP)+)* “ “ category$
      - Spanish philosophers
  - Similar patterns for es, it

NE id

- Identify which WK articles are NEs
  - Capitalisation norms (in some langs)
    - NEs -> proper nouns -> begin by uppercase
    - ¬NEs -> common nouns -> lowercase
  - Fetch article's title and redirects and look for occurrences
    1. In the WWW
    2. In its article body
    3. Combine

NE id

- Look for title in web search engine
  - Occurrences in first 50 hits
    - Begin with upper/lowercase + threshold
  - 76.7% P, 89.8% R
  - Drawback: noise due to sense variation
    - “Children's Machine” -> NE, laptop by OLPC
    - But in the WWW we might find: “... The children's machine ...”
      - Seymour Papert's book

NE id

- Look for title (and variants) in article's body (Bunescu & Pasca 06)
  - Exploit interlingual links
    - Whatever the lang compute occurrences in 10 langs that follow capitalisation norm
      - ca, en, es, fr, it, pt, sv, ...
    - More occurrences -> results more representat.
    - Almost language independent
  - Results
    - 76.5% P, 88.4% R

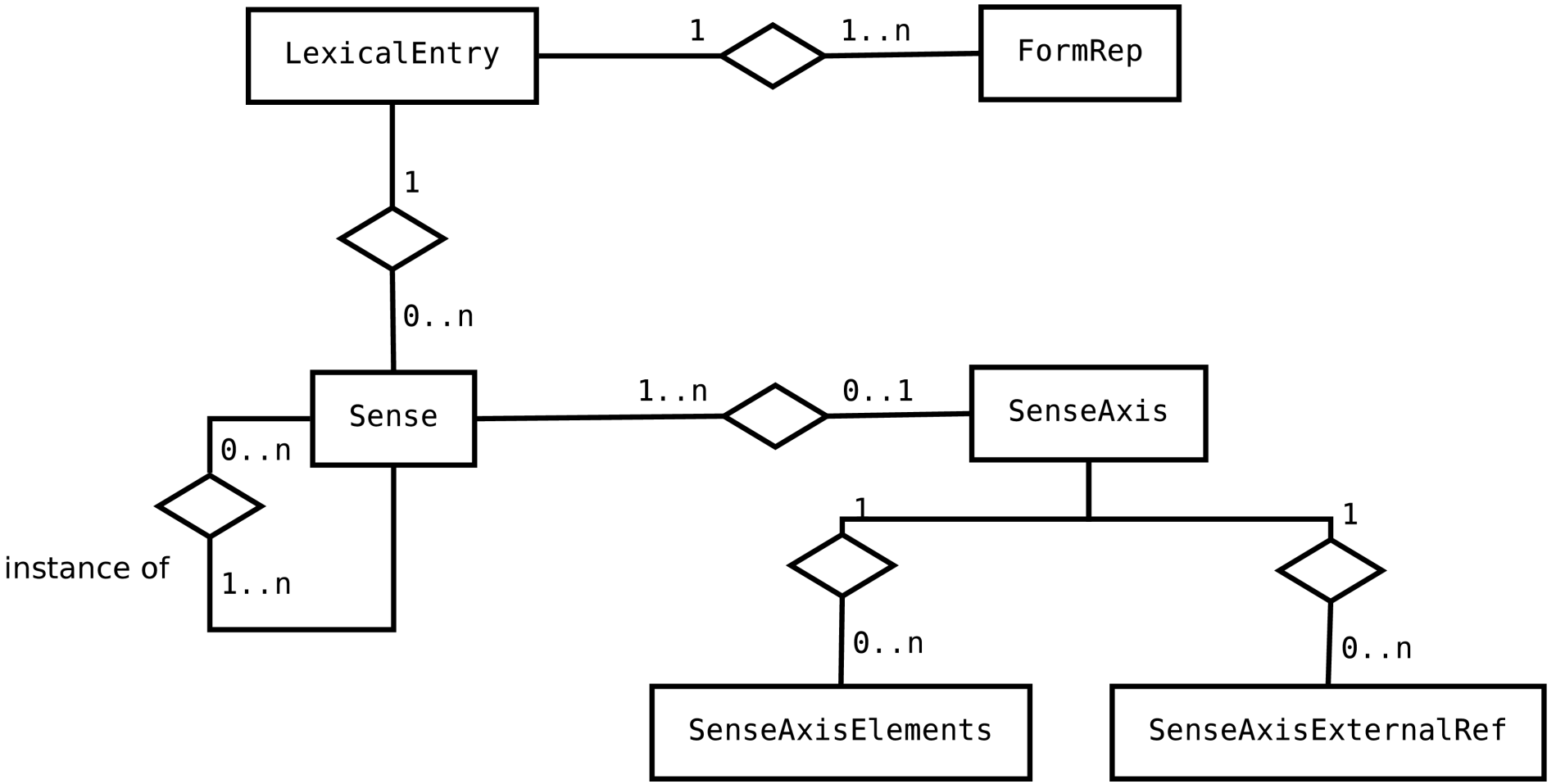
NE id

- Extract salient terms in article's body. Look for title + salient terms in WWW
  - TF-IDF corpus and stopword derived WK
  - e.g. “Children's Machine”
    - Salient terms: OLPC, 100$ laptop, ...
    - Search:
      - “Children's Machine”, OLPC, “100$ laptop”
        - Hopefully we won't get hits regarding the book!
  - 79.17% P, 90.48% R
    - WWW: 76.7% P, 89.8% R
    - WK: 76.5% P, 88.4% R

Postprocessing

- Additional NEs
  - An extracted NE for lang *a* may extract a further NE in lang *b*
- Connect NEs to ontologies
  - Exploit mappings LRs to ontologies
    - SUMO -> English
    - SIMPLE -> Italian

The NE Lexicon



LexicalEntry

le id	PoS
en_le_Tim_Robbins	PN

FormRepresentation

written form	variant type
en_Timothy_Francis_Robbins	alias
en_Timothy_Robbins	alias
en_Tim_Robbins	full

Sense

sense id	resource	id in resource	definition
en_s_Tim_Robbins	en_Wikipedia	269416	| location = West Covina, California, United States

SenseRelation

source sense id	relation type	target sense id
en_s_Tim_Robbins	instanceOf	en_s_actor0_18
en_s_Tim_Robbins	instanceOf	en_s_film_director0_18

en_s_Tim_Robbins	instanceOf	en_s_militant0_18
en_s_Tim_Robbins	instanceOf	en_s_screenwriter0_18

SenseAxis

senseaxis id	senseaxis type
sa_853829	eq_syn

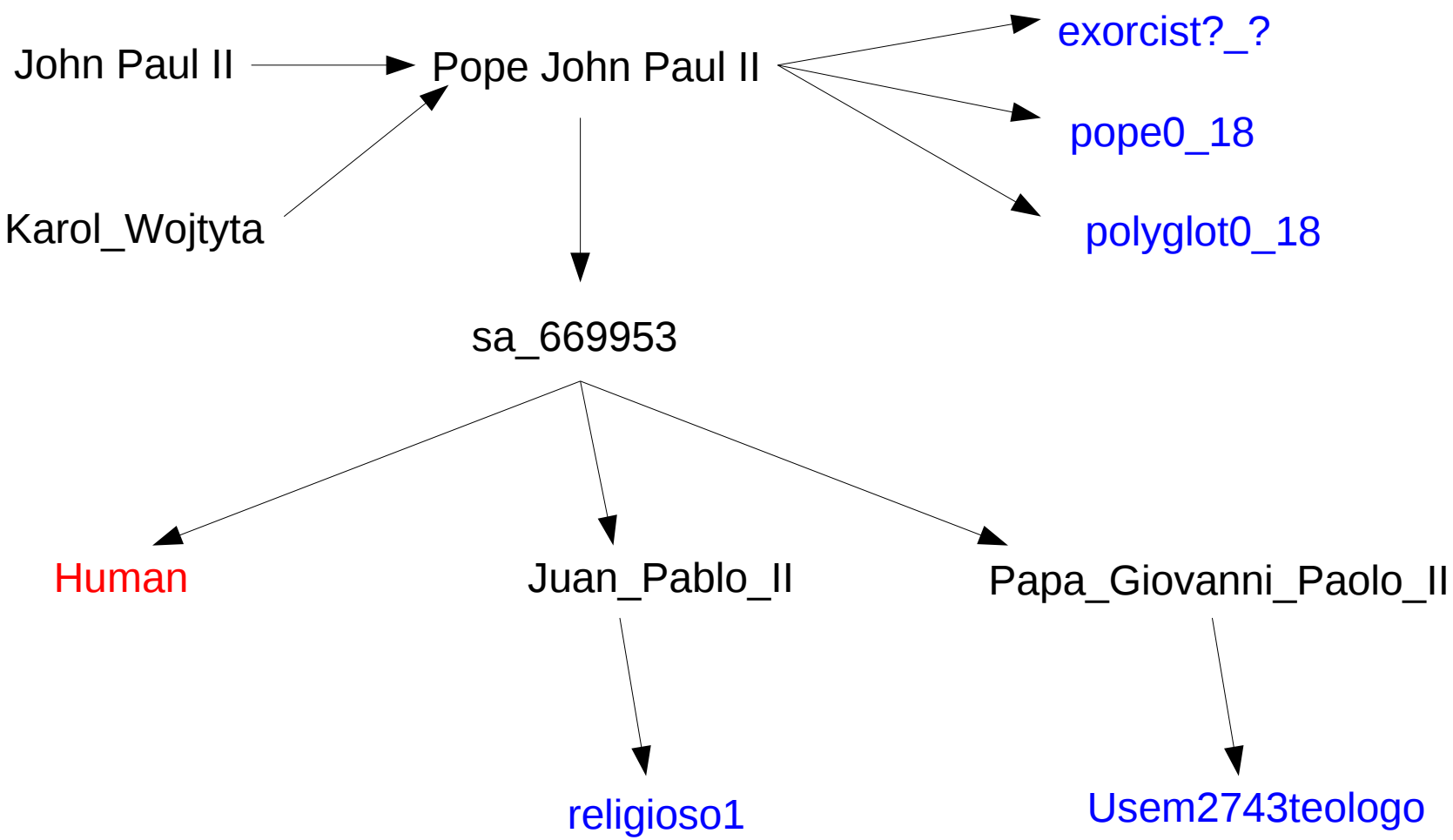
SenseAxisElements

senseaxiselements id	element	senseaxis id
sae_1479620	en_s_Tim_Robbins	sa_853829
sae_1479622	es_s_Tim_Robbins	sa_853829
sae_1479624	it_s_Tim_Robbins	sa_853829

SenseAxisExternalReference

senseaxisexternalref id	senseaxis id	resource	resource id	relation type
saer_737131	sa_853829	sumo	Position	+
saer_737132	sa_853829	sumo	believes	+
saer_854146	sa_853829	simple	Profession	

The NE Lexicon



The NE Lexicon

	EN	ES	IT
NEs	948,410	99,330	78,638
Written forms	1,541,993	128,796	104,745
Instance rels	1,366,899	128,796	139,190

- Postprocessing
  - NEs: 974,567 en, 137,583 es, 125,806 it
  - Ontolinks:
    - 814,251 SUMO, 42,824 SIMPLE
- http://www.ilc.cnr.it/ne-repository/

Outline

- Intro
- Knowledge Acquisition bottleneck
- NEs, why NEs?
- Building MINELex
- **Applying this to the real world**
- Conclusions

Application to QA

- BRILIW QA system
  - Cross-lingual en-es
  - Ranked 1st at CLEF 2006
  - Syntactic patterns
    - Detect expected answer type
    - Extract answer
      - After Passage Retrieval

Application to QA

- Validation module with NE knowledge
  - Two types of questions
    - Expect NE as answer type
      - Who is the General Secretary of Interpol?
    - Ask definitions of NEs
      - Who is Vigdis Finnbogadottir?
  - An answer is assessed as:
    - UNKNOWN -> NE not found in MINELex
    - CORRECT -> NE found, type matches expected
    - INCORRECT -> NE found, type does not match
  - Can reorder the answers provided
    - CORRECT >> UNKNOWN >> INCORRECT

Application to QA

- Evaluation
  - CLEF 2006 question set
  - Validation improves accuracy by 28.1%

- Example

Who is the General Secretary of Interpol?		
Answer	Validation	Reranking
Organización Internacional de Policía Criminal	UNKNOWN	2
Enrique Gómez	CORRECT	1
Jefe de la Policía Interna	UNKNOWN	3
Policía Internacional	UNKNOWN	4

To conclude…

- Combination to circumvent KABP
  - *Exploit* wiki community & lexicographers
  - Contribute yourself:
    - Standards -> interoperability
- Practical case of study
  - “ingredients” + NLP machinery -> NE lex.
  - MINELex
    - Usefulness? -> QA

End

Thank you for listening!

Questions, comments, ...?

Antonio Toral

{fistname.lastname}[at]ilc.cnr.it
Istituto di Linguistica Computazionale - CNR, Pisa (Italy)
Dep. Llenguatges i Sistemes Informàtics - Universitat d'Alacant (Spain)