An Automatically Built Named Entity Lexicon for Arabic

M. Attia*, **A. Toral***, L. Tounsi*, M. Monachini^, J.v. Genabith*

*Dublin City University (Ireland)
^Istituto di Linguistica Computazionale - CNR, Pisa (Italy)

Contents

- Introduction
  - NLP acquisition bottleneck, MINELex
- Methodology
  - Mapping, Extraction, Identification, Diacritisation, ...
- Results
- Conclusions

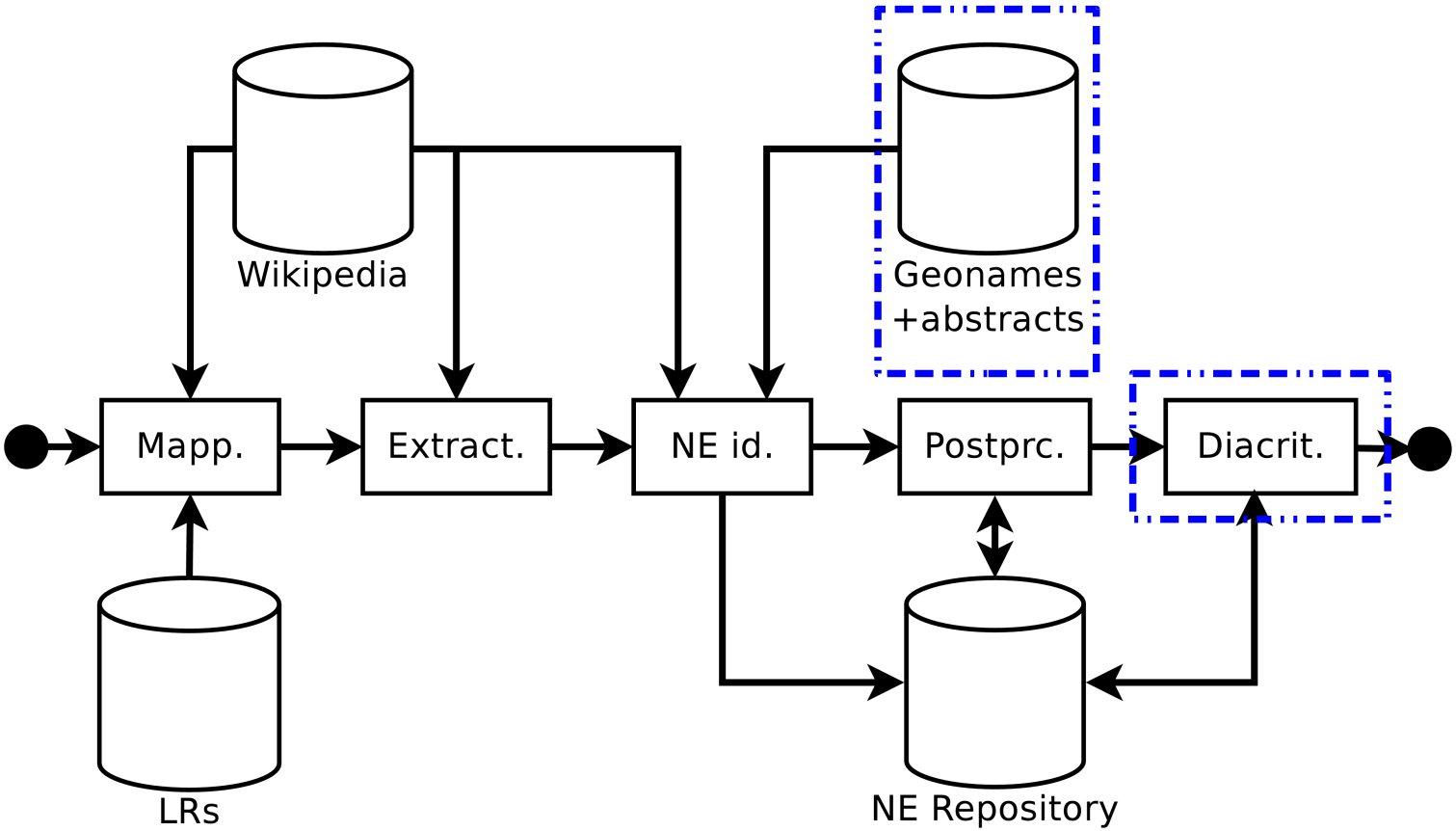
Intro

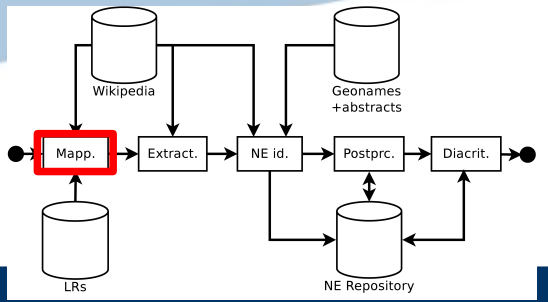
- NLP apps make extensive use of LRs
- Big effort during last 15 years to build resources
  - e.g. lexica: WordNet, EuroWordNet, SIMPLE, etc.
- Enough coverage?
  - ~OK → verbs, adjs, advs, common nouns
  - ¬OK → NEs, domain terms, multiwords
- "humans cannot manually structure the available knowledge at the same pace as it becomes available" (Philpot 05)
  - Automatic procedures needed!

Intro

- Step forward → 3 ingredients
  - Web 2.0, LRs, interoperability
- MINELex: Multilingual, Interoperable NE Lexicon
  - Derived automatically from Wikipedia and LRs
  - General approach, applied to:
    - English WN: 975k NEs
    - Spanish WN: 137k NEs
    - Italian SIMPLE-CLIPS: 125k NEs
  - NEs linked to LRs and ontologies
  - Extrinsic eval, QA → 28% increment accuracy
- Is the approach applicable to other lang families?
  - Arabic (arWN, arWK)

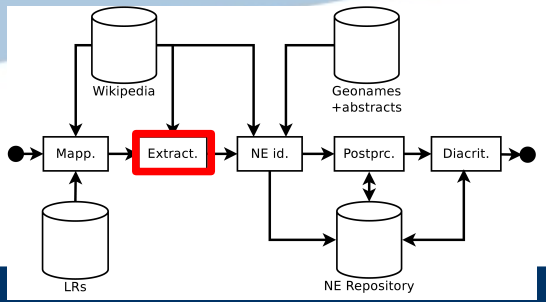
Methodology





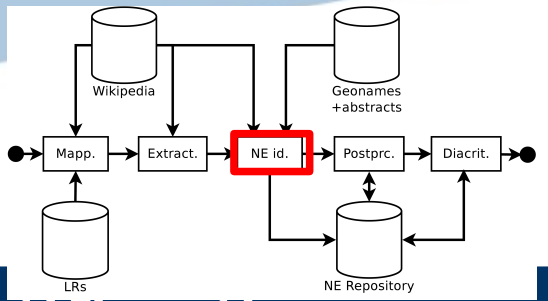
Methodology: Mapping

- Identify senses of arWN that can be extended with NEs, i.e. instantiable nouns
- arWN (and enWN) do not have this info but have instance_of relations, i.e. instantiated nouns
  - country1 has_instance Malta
- Union of instantiated nouns from both resources
  - A: arWN i.n. + recursive hyponyms → 384
  - B: enWN i.n. + recursive hyponyms → mapping arWN + recursive hyponyms → 1,475
  - Final set: A U B → 1,572 senses, 1,187 nouns
- Lemma matching: i.n. ↔ arWK cats
  - 40.6%



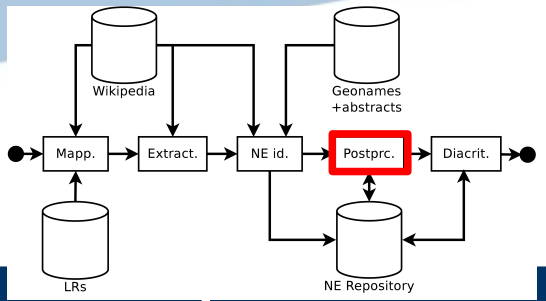
Methodology: Extraction

- Extract articles from mapped categories
- ...and hyponym subcategories → pattern:
  - ^category_
    - From "سياسيون" (politicians)
      - "سياسيون_حسب_الحزب" (politicians by nationality)
      - "سياسيون_بريطانيون" (British politicians)
- Discard administrative categories



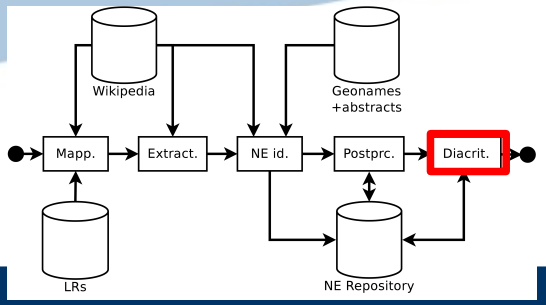
Methodology: NE Identification

- Original approach relied on capitalisation norms
  - Look for occurrences of title in body, check percentage it occurs with lowercase vs. uppercase
- … but Arabic does not follow them → exploit inter-lingual links to obtain equivalent article in 10 langs that follow cap. norms (en, es, fr, it, …)
  - Drawback: covers only 62.5% of articles
- Further heuristics to improve recall
  - Keywords from abstracts
    - LOC (16): abstract begins with “city”, “country”, etc
    - PER (60 + exclusion list 160): abstract contains “born in”, “studied in”, etc
  - Geonames: lexicon of geographic NEs



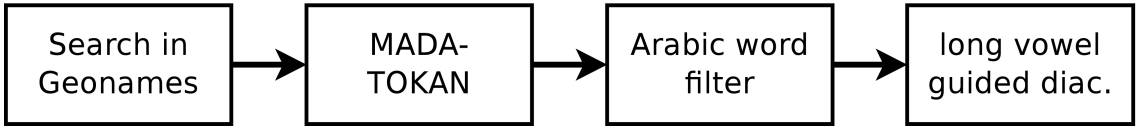
Methodology: Postprocessing

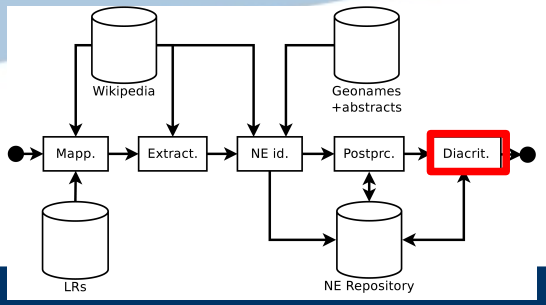
- Cross-fertilisation
  - Further ar NEs can be obtained by exploiting
    - Links between en, es, it NEs and their LRs
    - Interconnections among these LRs
  - E.g. NE extracted for es has equivalent in arWK but has not been extracted
    - Extract and connect to arWN following mapping esWN → enWN → arWN



Diacritisation

- Diacritics: Short marks above or under letters
  - الإِمَارَاتُ العَرَبِيَّةُ المُتَّحِدَةُ / al-imaratu al-arabiyyatu al-muttahidatu / “United Arab Emirates”
- Why needed? Speech, Syntactic disambiguation, WordNet
- Approach for restoring diacritics:
  - Checking available diacritised lists
  - Using a diacritisation tool
  - Using heuristics





Diacritisation

- Diacritised lists: geonames.de, geonames.org
  - 3,5k NEs matched (10%)
- Diacritisation tool: MADA
  - 29% coverage, mainly due to OOV (NEs)
- Using heuristics
  - Most unknown words are foreign names
  - Transliteration of foreign names usually employs long vowels
  - Native Arabic names do not follow this assumption and must be excluded
  - 59% coverage
- Combination: 73% coverage

Evaluation

- Data used
  - arWN (connected to enWN 2.0)
  - enWN 2.1
  - Automatic mapping enWN 2.1 ↔ enWN 2.0
  - arWK dump Feb 2010. 234k articles, 33k categories
- Test set
  - 1k arWK articles that belong to the categories mapped
  - Annotated as [NE, not NE]
- Measures: P, R, F1, F0.5

Evaluation: NE identification

Heur.	Threshold	P	R	F1	F0.5
no	0.91	**99.25**	42.39	59.40	78.25
	0.41	98.33	50.16	66.43	**82.49**
	0.01	94.70	**51.33**	**66.57**	81.01
yes	0.91	**99.28**	58.68	73.76	87.21
	0.41	98.55	65.07	**78.38**	**89.35**
	0.01	95.83	**66.13**	78.26	87.94

Evaluation: NE extraction

Heur.	Threshold	NEs	Relations	Variants
no	0.91	23,910	27,422	24,887
	0.41	28,048	32,287	29,451
	0.01	30,354	34,901	32,205
yes	0.91	31,284	36,271	32,386
	0.41	35,423	41,136	36,940
	0.01	37,729	43,750	39,693

- Postprocessing:
  - 11.7k en, 6.8k it, 6.9k es NEs have ar equivalent
  - Discard duplicates + NEs extracted for ar → 6.5k NEs
  - Added to MINELex → contains 44k ar NEs

Output Example

FormRepresentation

LE id	written form	v. type	script	orthog. n.
ar_le_الأمم المتّحدة	ar_الأمم المتّحدة	full	Arab	arabicUnpointed
ar_le_الأمم المتّحدة	ar_اَلأُمَم المتَّحِدَة	full	Arab	arabicPointed
ar_le_mnZm	ar_mnZm	full	Latin	
en_le_United_Nations	en_United_Nations	full	Latin	

SenseAxis

SA id	element
1	ar_s_الأمم المتّحدة
1	en_s_United_Nations

Sense

S id	LE id	res.	res. id
ar_s_الأمم المتّحدة	ar_le_الأمم المتّحدة	ar_WK	2270
ar_s_109710501	ar_le_mnZm	ar_WN	109710501
en_s_United_Nations	en_le_United_Nations	en_WK	31769

SenseAxisExternalRef

SA id	resource	resource id	relation
1	SUMO	PoliticalOrganization	at

SenseRelation

source id	target id	relation
ar_s_الأمم المتّحدة	ar_s_109710501	instanceOf

Confidence (NE id)

S id	mode	occurrences	confidence
ar_s_الأمم المتّحدة	wiki10	250	0.996

Conclusions

- Adapted and extended generic methodology to build a NE lexicon to Arabic: arWN and arWK
- Challenges: NE identification and diacritisation
- Result: 44k NE lex
  - Connected to
    - Intralingual: arWN synsets
    - Interlingual: equivalent NEs in en, es, it + ontologies
  - Can be used with different levels of granularity
  - Compliant with ISO LMF format
- Available at
  - www.ilc.cnr.it/ne-repository

End

Thank you very much!

Questions?