Best Practices in Language Resources for Multi-Lingual Information Access (MLIA)

Khalid Choukri, Nicolas Moreau

choukri@elda.org

ELDA – Paris, France

http://www.elda.org

Outline

- Presentation of ELRA / ELDA
- Definition of MLIA Resources
- Inventory of MLIA Resources
- Priority Requirements
- Recommendations

European Language Resource Association (ELRA)
An Improved infrastructure for Data sharing

- Centralized non-for-profit organization for the collection, distribution and validation of speech, text, and terminology resources and tools,
  extension to multimodal/multimedia resources
- Evaluation Activities (since 1998)
- Evaluation campaigns as well as evaluation methodologies & research

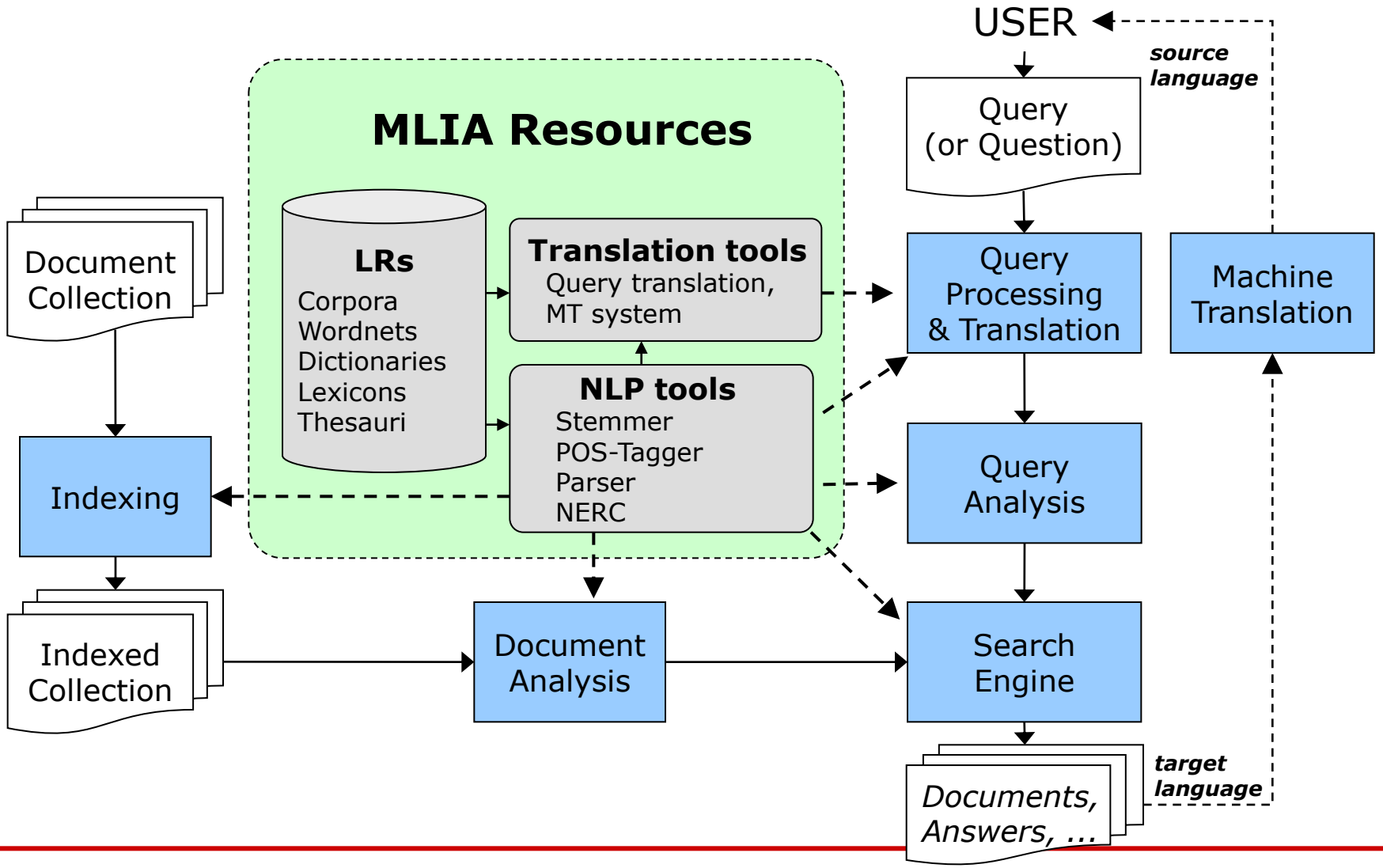
⇨A Repository Center:

- ⇨Technical & Logistic issues
- ⇨Commercial issues (prices, fees, royalties)
- ⇨Legal issues (Licensing, IPR)
- ⇨ Information Dissemination

An operational company:

Evaluations and Language Resources Distribution Agency (ELDA)

MLIA System / MLIA Resources



BLARK View

BLARK: Basic LAnguage Resource Kit (www.blark.org)

- Goal: identify minimal set of modules and LRs necessary to address a given technology
- 3 axes: Technologies / Components (Modules) / Resources

Application 1 (Technology, System, ...)

- **Module 1.1** (+++)
  - LR 1.1.a (+)
  - LR 1.1.b (+++)
  - LR 1.1.c (+)
- **Module 1.2** (+)
  - LR 1.2.a (+++)
  - LR 1.2.b (++)
- **Module 1.3** (+)
  - LR 1.3.a (++)
  - LR 1.3.b (++)
  - LR 1.3.c (+)

•••

Application *N*

- **Module *N*.1** (+++)
  - LR *N*.1.a (++)
  - LR *N*.1.b (+++)
  - LR *N*.1.c (++)
- **Module 1.1** (+++)
  - LR 1.1.a (+)
  - LR 1.1.b (+++)
  - LR 1.1.c (+)
- **Module *N*.3** (++)
  - LR *N*.3.a (++)
  - LR *N*.3.b (++)

Relevant (+) / Important (++) / Essential (+++)

Principle of BLARK Matrices

Application vs. Modules:

Modules \ Applications	Tech 1	Tech 2	Tech 3	(...)
M 1	+++	+	+	
M 2	∅	++	+++	
M 3	++	++	++	
(...)				

Modules vs. Resources:

Resources \ Modules	M 1	M 2	M 3	(...)
LR 1	+++	+	∅	
LR 2	++	++	+++	
LR 3	∅	++	+	
(...)				

Relevant (+) / Important (++) / Essential (+++)

BLARK Matrices for MLIA (1)

Application vs. Modules :

Modules \ Applications	CL-IE	CL-IR	CL-QA
Sentence Boundary Detection	+	+	+
Tokenizer	++	++	+++
Morphological Analyzer (deriv., stemm., diacritic, …)	++	++	+++
POS Tagger	+++	+++	+++
Chunker (Shallow Parser)	++	++	++
Named Entity Recognizer	+++	++	++
Word Sense Disambiguation	++	++	++
Syntactic Analyzer	++	++	+++
Semantic Analyzer (incl. coreference resolution)	+++	++	+++
Language Identifier	++	++	++
Translation (MT, query translation…)	+++	+++	+++

BLARK Matrices for MLIA (2)

Resources vs. Modules :

Modules \ Resources	Stop word list	Un-annotated Corpora	Annotated Corpora	Parallel Multiling Corpora	Monoling. Lexicons	Multiling Lexicons	Grammars	Monoling Thesauri, Ontologies	Multiling. Thesauri, Ontologies
Sentence Boundary Detection		+++	+				+		
Tokenizer	+++	+++	+						
Morphological Analyzer	+	+	+++		+++	+			
POS Tagger			+++		+++				
Chunker (Shallow Parser)			+++				+++		
Named Entity Recognizer			+++		+++	+++		+++	++
Word Sense Disambiguation					+++		+++		
Syntactic Analyzer			+++				+++		
Semantic Analyzer (incl. coreference resolution)			+++					+++	
Language Identifier				+++		+++	+++		
Translation (MT, query translation…)				+++		+++	++		+++

Online Survey … carried out within TrebleCLEF

- To Collect information (online questionnaire):
  - *Inventory of most used LRs and NLP tools (Key Resources)*
  - *Most needed MLIA Resources in the future*
  - *Less covered languages or language pairs*
- Answers from 69 respondents from 22 countries:
  - *Academic world (78.7%)*
  - *Independent R&D centres (13.1%)*
  - *Private companies (8.2%)*
- Respondents active in different MLIA domains:
  - *Cross-Language Information Retrieval*
  - *Cross-Language Question Answering*
  - *Information Extraction*
  - *Text Classification and Summarization*
  - *Domain specific and multimodal IR technologies*

Inventory of MLIA Resources

Language Resources

More than 160 key resources identified, covering more than 50 languages

Type	#
Multiling. Test Collections Sources	10
Monolingual Corpora	42
Multilingual, Parallel Corpora	26
Multimodal Corpora	6

Type	#
Monolingual Dic., Lexicons	10
Multilingual Dic, Lexicons	12
Monolingual Onto, Thesauri	45
Multilingual Onto, Thesauri	12

NLP Tools

More than 70 key tools identified

Type	#
Stemmers and Morphologic Analysis	19
POS Taggers	8
NER Classifiers	6

Type	#
Syntactic Parsers	10
Semantic Parsers	8
Toolkits	27

Identification of Needed Resources

- New corpora to cover **domain-specific** MLIA technologies:
  - Multimodal & Multilingual Corpora
  - Multilingual Web Data
  - Multilingual Patent Corpora
  - Other domain specific data (Medical, E-learning, ...)
- Large **parallel and aligned corpora** for more languages (not only English vs. another language)
- More (and improved) **WordNets** in many languages
- Easier Access to LRs produced by **Evaluation Campaigns**

Action Plan & Recommendations

To implement the identified and needed resources within the next few years:

1) Conduct **BLARK studies** for selected key technologies and languages
   See BLARK (*Basic LAnguage Resource Kit*): www.blark.org
2) Identify **real user needs** through domain specific surveys
3) Ensure a large **consensus on needed resources**
   Derive an **agenda** to make them available
   See ELARK (*Extended Language Resource Kit*)
4) Elaborate a **generic strategy** that can be applied to any specific languages
5) Foster **cross-disciplinary** collaborative LR creation efforts
   Constitute **common pools of resources**
6) Elaborate efficient **distribution strategies**

Conclusion 1/2

Main achievements of this study:

- Inventory of existing Language Resources for MLIA
  - Can be updated through the TrebleCLEF Best Practices website
  - Link: http://www.trebleclef.eu/jsbestpractices.php
- To help develop the most critical sets of MLIA language resources:
  - First attempt to design a BLARK for MLIA
  - Action plan and recommendations

Future tasks:

- Draw a more precise picture of developer's requirements through other consultations
- Revise and update more specific BLARK's on a regular basis
- Draw a roadmap that will outline areas and priorities for cross-disciplinary collaborations
- Set up close collaboration efforts:
  - Research communities (web search, speech recognition, OCR, text summarization, data mining, etc.)
  - User communities (medical, legal, humanities, etc.)
  - Focused on multimedia retrieval and application-oriented, domain specific tasks

Conclusion 2/2: CLEF Packages

- CLEF Evaluation Packages
  - Multilingual document collections and topics
  - Relevance assessments and results
  - Scoring tools
  - Step-by-step evaluation guidelines

- Available and upcoming packages
  - CLEF AdHoc (on news data)
  - CLEF Domain Specific (scientific documents)
  - CLEF Question-Answering (news & speech transcripts)
  - Many others: GeoCLEF, CLEF Patent Retrieval,...

- Distributed in ELRA/ELDA catalogue (http://catalog.elra.info/)