An Arabic Question-Answering system for factoid questions

Wissal BRINI¹, Omar TRIGUI¹, Mariem ELLOUZE¹, Slim MESFAR², Lamia BELGUIITH¹, Paolo ROSSO³

¹ MIRACL Lab., University of Sfax, Tunisia
² ISI, University of El-Manar, Tunisia
³ NLEL, Universidad Politécnica, Valencia, Spain

Tozeur - Tunisia, June 2009

This work was funded by the AECID-PCI B/017961/08 research project.
Outlines

- Context and Problematic
- Question-Answering Systems for Arabic Language
- QASAL system using NooJ platform
- Conclusion & Future Work
Where was Ibn Khaldoun born?

List of documents

Human Intervention

Loss of time

Tunisia

(Tunisia)
Question/Answering Systems for Arabic Language

**Question / Answering System**

- User
- Question
- Q/A process
- Answer
- Documents
- Question Analysis Module
- Passage Retrieval Module
- Answer Extraction Module

**Existing Arabic Q/A Systems**

- **QARAB** [Hammo, 2002] (based on Al-Raya corpus, uses IR and NLP techniques).
- **AQAS** [Mohammed, 1993] (extracts answers from only structured texts).
- **ArabiQA** [Benajiba, 2007] (deals with factoid questions, embeds NER module and adapts JIRS to extract passages from Arabic texts).
Question/Answering Systems for Arabic Language

Problems

• Arabic is highly inflectional, derivational and agglutinative.
• Linguistic resources (Corpuses and NLP tools) are very few.
• The lack of vowels and the absence of capitalization.

We propose to use linguistic resources included into the NooJ’s Arabic module.
**Question analysis module**

- The linguistic analysis of: 
  متي استقلت تونس؟ (“When Tunisia became independent?”).

Text Annotation Structure:

```
متي استقلت تونس؟

N+u+LOC+Ville
N+a+LOC+Ville
N+i+LOC+Ville
N+LOC+PAYS
N+LOC+Ville+PAYS
```

```
V+A+I+3+f+s+Tr
V+K+I+3+f+s+Tr
```

```
ADV+INTERR
```

**Text Annotation Structure:**

متي استقلت تونس؟
• Construction of sub-graphs which process, respectively, questions starting with: When "متى", How much "كم", Who "من" and Where "أين".

• The main graph contains the four corresponding sub-graphs : Time, Quantity, Person and Localization.
QASAL system using NooJ platform (2/5)

- Construction of the local grammar for the question: متى استقلت تونس؟ (When Tunisia became independent?)

NooJ syntactic grammar (Time sub-graph)
$V$: the value of the verbal word form استقللت (became independent),
$N$: the value of the nominal word تونس (Tunisia),
(<PREP>+<E>): an eventual presence of a preposition,
(<PRON>+<E>): an eventual presence of a pronoun,
<TIMEX>: any temporal expression already annotated in the answer text using the named entity recognition tool in NooJ.
• Application of NooJ’s local grammars to the question: متى استقلت تونس؟ (“When Tunisia became independent ?”).

Concordance table for the factoid question: متى استقلت تونس؟ (Where Tunisia became independent ?)
Passage Retrieval module

• Application of Regular Expressions (RegEx) for texts containing answers for the question: متى استقلت تونس؟ (“When Tunisia became independent?”).

• The detected occurrences are displayed into the concordance table.
**Answer Extraction module**

• Using the displayed concordance table to automatically extract the answer of the question.

Answer Extraction for the factoid question: متى استقلت تونس؟

استقلت تونس في 20 مارس 1956
تحصلت تونس على استقلالها في 20 مارس 1956
Conclusion & Future Work

- NooJ platform could be used to implement Q-A systems.
- Including Linguistic resources using NooJ’s Arabic module could improve the Q-A task.

- Considering, for further experiments, a semantic query expansion (using Arabic WordNet).
Thank you for your attention

>> Questions/Answers