Building A Modern Standard Arabic Corpus

Prepared by
Turki Bakodah
214065

Agenda

- Introduction.
- The Objective of this Presentation .
- Problem facing our example.
- Building the Corpus.
- Collection processing.
- Corpus Assessments.
- Conclusion
- References.

Introduction

- The report by Madar Research Journal which includes statistics and forecasts on Internet users in 17 Arab countries estimates the size of the Internet community in the Arab world in excess of 25 million by end of 2005.
- currently 1.9 million online websites in Arabic and number is expected to double every year.

Introduction

- so if we want to keep up the growth we need to do the following:

1. Providing users with quality web portals.
2. Efficient search engines.
3. Trying to come up with solutions to some obstacles that faced MSA

The Objective of this presentation .

- to learn the process of building an Arabic corpus.
- Building Arabic corpus that would help in compare the language used in different parts of the Arabic world.

Problem facing our example

- The same word different meanings.

English Word	El-khabar Algeria	Al-anbaa Morocco
Arrest	حجز	توقيف
Tend to fall	آيلة للسقوط	معرضة للسقوط

- names used in different regions for the same object for example, Ministry of Education".

Egypt ,Saudi Arabia	Qatar, Kuwait, Jordan	Mauritania
وزارة المعارف	وزارة التربية والتعليم	وزارة التهذيب

Building the Corpus.

- we collected text from newspapers and news services from different Arab speaking countries.
- We used a developed spider program to get the data from each site. The spider will traverse the links and save the pages linked to the main page.

Collection processing

- the next step was to convert the data to a common encoding We used URSA, Unicode Retrieval System Architecture, is a high-performance text retrieval system that can index and retrieve Unicode texts and provide word frequencies and other data.

Corpus Assessments.

- Using available tools we first experimented by applying some statistical and probability tests, such as Zipf's law.
- Zipf's law, if we count up how often each word occurs in a corpus and then list these words in the order of their frequency of occurrence, then the relationship between the frequency of a given word and its position in the list will be a constant k.

Corpus Assessments.

- Ideally a simple graph will show a straight line with a slope –1. So we checked the situation in our corpus by starting with one file and increasingly adding more files

  to a corpus and checking the behavior of the relation between the rank and the frequency.

Conclusion

- To summarize the process for building Arabic corpus.
- collect data.
- Collection processing.
- Corpus Assessments.

References

- **Building A Modern Standard Arabic Corpus**
- Ahmed Abdelali.
- Jim Cowie.
- Hamdy S. Soliman.