DSIC

Arabic-English Question Answering

Yassine BENAJIBA

Dpto. Sistemas Informáticos y Computación

Universidad Politécnica Valencia, Spain - **ybenajiba@dsic.upv.es**

Information Retrieval and Question Answering

Information Retrieval
Retrieval of documents from a collection in response to a query.

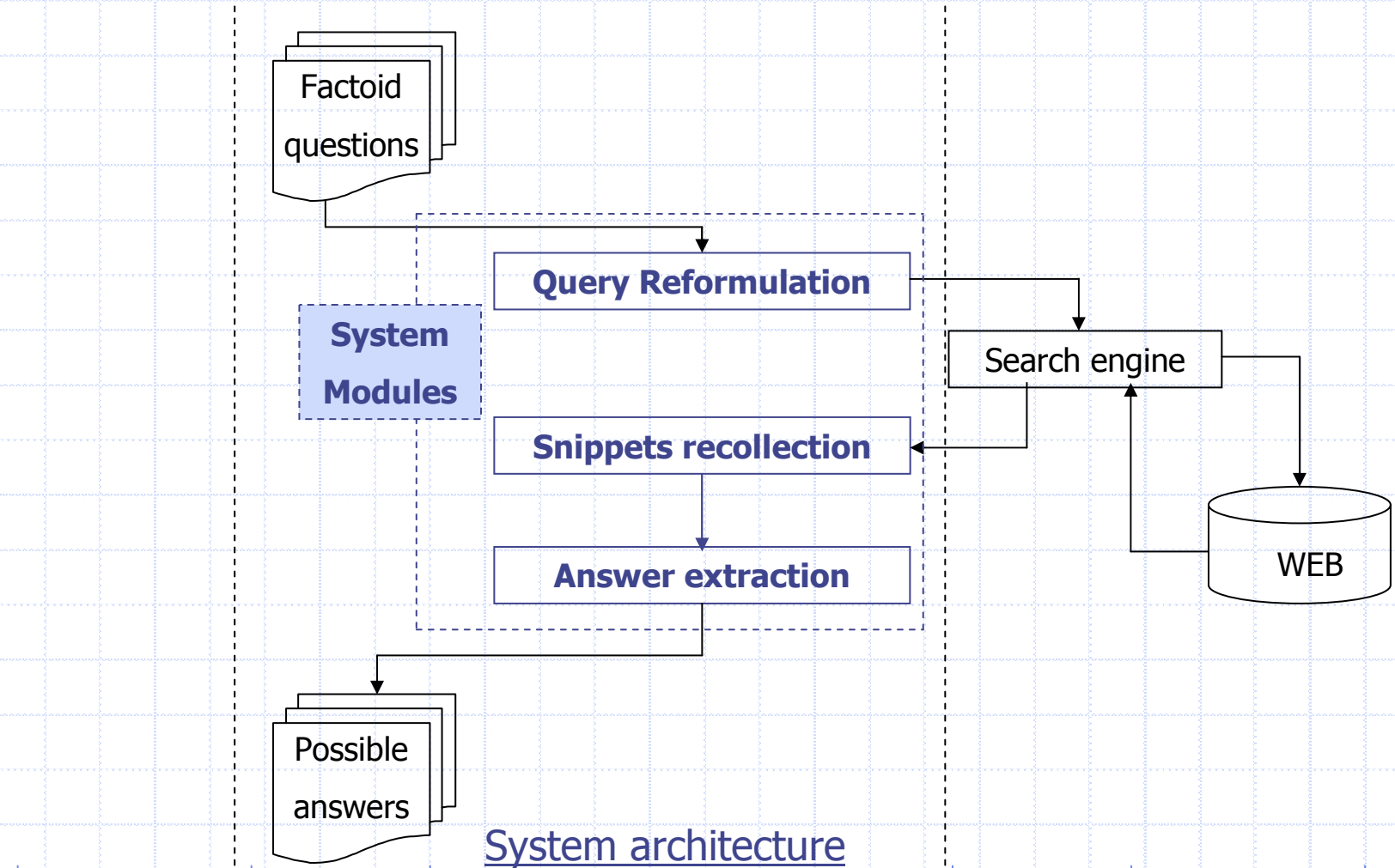
Cross-Language Information Retrieval
Where queries and documents are in different languages.

Question Answering
Find an answer to an open domain question in a large collection of documents.

Cross-Language Question Answering
Where questions and answers are in different languages.

CL-QA web – based = Web + MT system + QA System web-based
QA System based on web data-redundancy.

The QA web-based approach



System architecture

The QA web-based approach

I. Query reformulation

E.g. Question: Where is the ICT EU-India Conference in 2005?

Bag of words : **Set of non stop-words:**

e.g. : "is ICT EU-India Conference 2005"

Verb movement : **Eliminate verb or move it to the end.**

e.g. : "the ICT EU-India Conference in 2005 is"

Components : **Divide the original query into Components:**

e.g. : "is the ICT EU-India Conference "
"in 2005"

+ Make new reformulations combing the Components:

e.g. : "in 2005 is the ICT EU-India Conference "

The QA web-based approach

I. Query reformulation

Components without the first word :

Eliminate the first word:

e.g. : "the ICT EU-India Conference in 2005"

+ Make "Components" reformulation:

e.g. : "the ICT EU-India Conference" "in 2005"

"in 2005 the ICT EU-India Conference"

Components without the first and second words :

Eliminate the first wordand second words:

e.g. : "ICT EU-India Conference in 2005"

+ Make "Components" reformulation:

e.g. : "ICT EU-India Conference" "in 2005"

"in 2005 ICT EU-India Conference"

The QA web-based approach

II. Snippet recollection

Example of a snippet retrieved with the reformulation:

the ICT EU-India Conference in 2005

ICT for **EU-INDIA** CCD **2005**

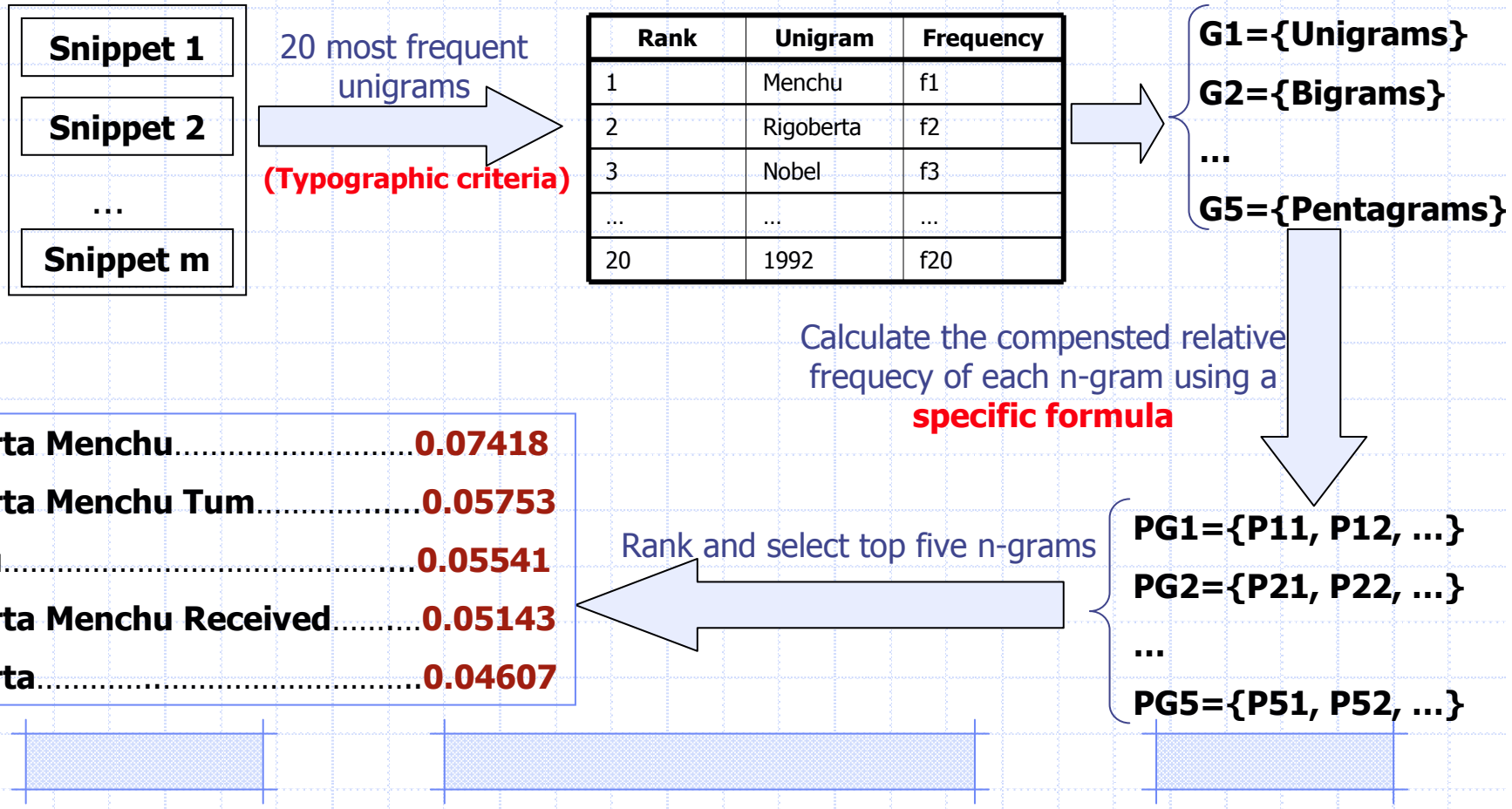
ICT for **EU-INDIA** CCD **2005**. **ICT** for **EU-INDIA** Cross Cultural Dissemination 2nd Annual **Conference Valencia, Spain**, November 14-15, **2005 …**

www.dsic.upv.es/workshops/**euindia**05/ - 9k – 7 Nov 2005 - Cached – Similar pages

The QA web-based approach

III. Answer extraction

E.g. Question: Who received the Nobel Peace Prize in 1992? (Rigoberta Menchu)



The QA web-based approach

III. Answer extraction

$$P_{g(n)} = \sum_{i=1}^{n} \sum_{j=1}^{n-i} \frac{f_{j(i)}}{\sum_{\forall x \in G_i} f_{x(i)}}$$

The compensated relative frequency of a n-gram g(n) = (w1, …, wn)

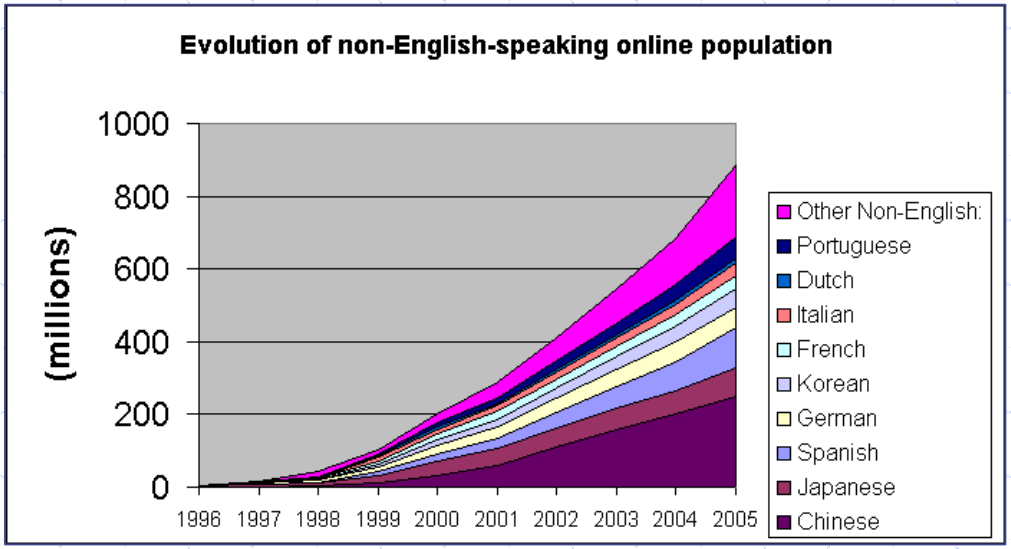
Gi : the set of n-grams of size *i*
|Gi| : the cardinality of this set
j(i) : an n–gram *j* of size *i* contained in *g*
fj(i) : the frequency of occurrence of this n-gram

Arabic Language

- One of the six official languages of the United Nations.

- Mother tongue of 300 millions people.

Egyptian Demographic Center, 2000



http://www.glreach.com/globstats/evol.html

Arabic Language

- The orientation of writing is from right to left.

أشجار اللوز

- Arabic Alphabet consists of 28 characters.

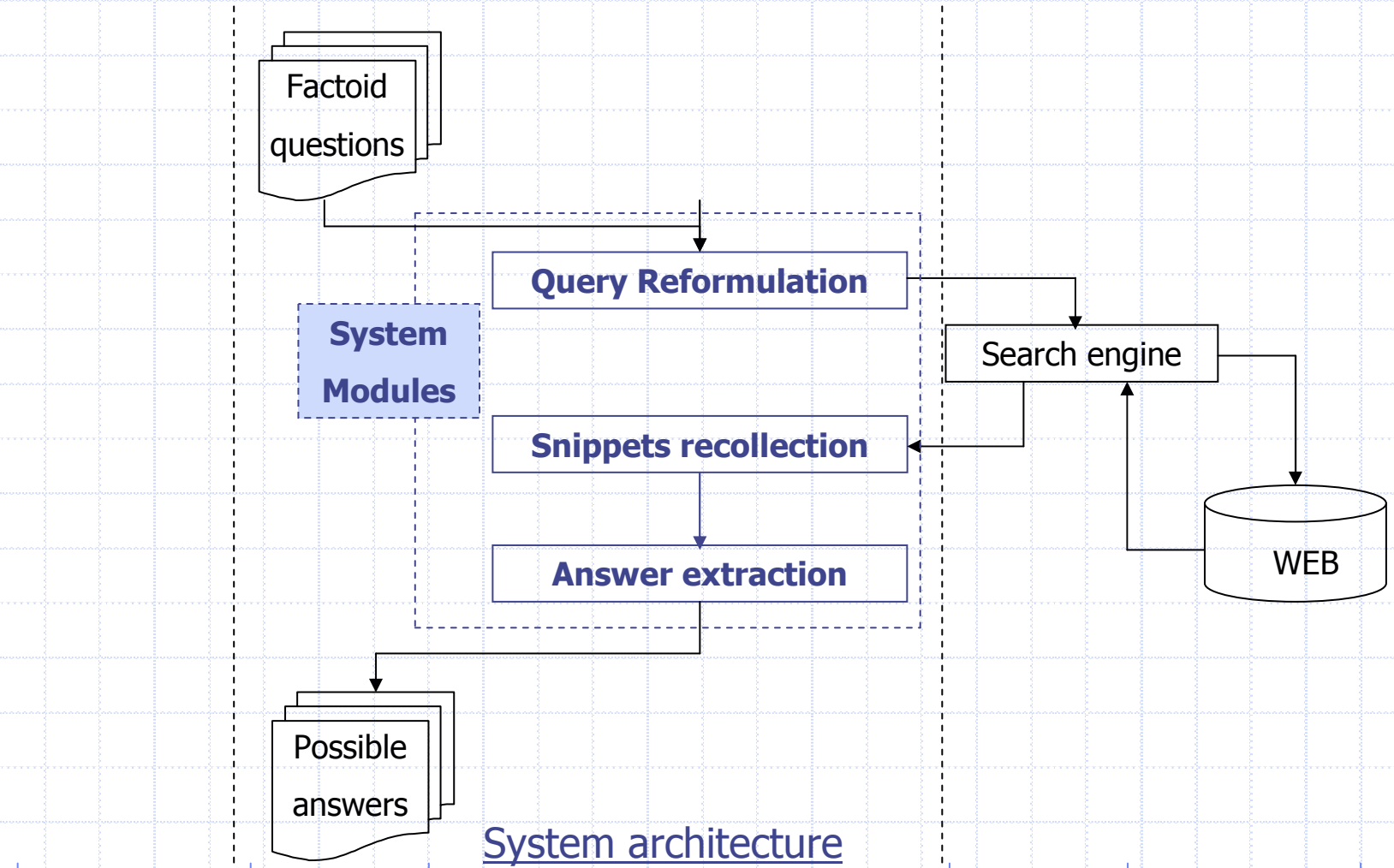
- Some characters exist in the Arabic language and are absent in English,

e.g. : ..., ق, ع

... and Vice Versa,

e.g.: V, P, G.

Preliminary experiments



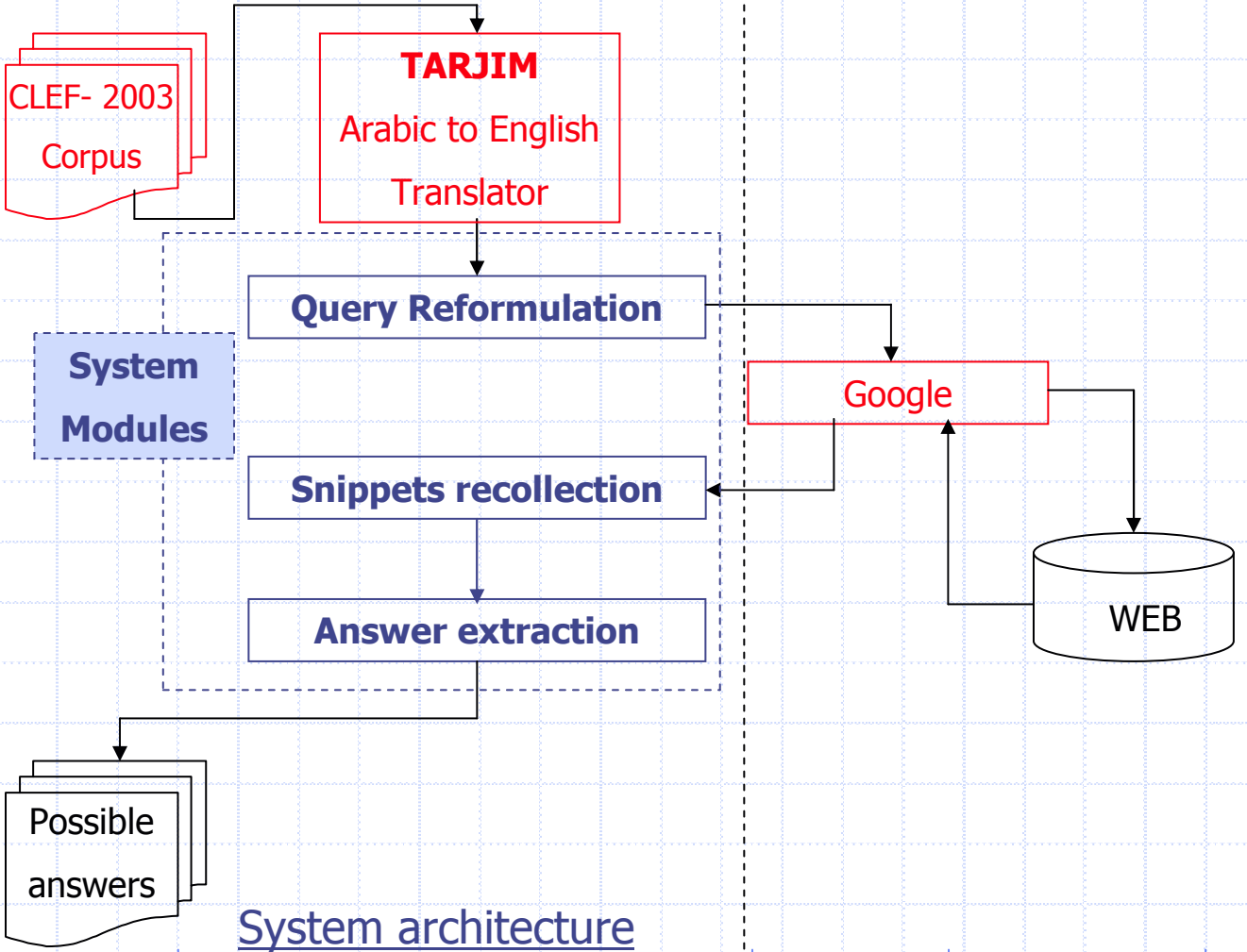
System architecture

Preliminary experiments

Cross Language Evaluation Forum:
European consortium for international competitions about IR and QA systems

Online machine translation system
http://tarjim.ajeeb.com/

Web Search Engine
www.google.com



System architecture

Experimental results (1)

Questions	Bag words	Comp.	Comp no 1st word	Comp no 1st and 2nd words	Verb mov.
English (original)	9.1% (41)	17.1% (77)	14.9% (67)	10.4% (47)	**24% (108)**
English (from Arabic)	3.8% (17)	1.6% (7)	4.9% (21)	4.9% (21)	**7.2% (31)**

Precision of correct answers (over 450)

Experimental results (2)

Questions	Bag words	Comp.	Comp no 1st word	Comp no 1st and 2nd words	Verb mov.
English (original)	17.1% 0.12	24.4% 0.19	26.7% 0.20	22.0% 0.16	**39.5%** **0.31**
English (from Arabic)	6.0% 0.04	2.4% 0.02	7.4% 0.06	8.4% 0.06	**10.7%** **0.08**

Precision and *Mean Reciprocal Rank* (MRR) measures%

$$MRR = \frac{1}{n}\sum_{i=1}^{n} r_i$$

The translation process causes a decreasing of even more than 30%

Experimental results (3)

Original	What was the name of the singer and head of Nirvana?
Arabic	ما اسم المغني و رئيس نرفانا ؟
Translation	What is the name of the main singer of Nirfana?

Example in which also a proper name was badly translated

Experimental results (4)

Original	How many European countries form part of the G7?
Arabic	كم عدد الدول الأوربية المكونة لمجموعة السبع؟
Translation	Quantity of an European country belongs to the group of seven?

Example of bad translation

Experimental results (5)

Original	Which American state has the strictest environmental laws?
Arabic	ما هي الولاية الأمريكية ذات القانون البيئي الأكثر صرامة؟
Translation	What she is the American state for which the en vironmental laws with more stricness?

Rare example of wrong translation and right answer

(California)

Experimental results (6)

Original	During what month do almond trees blossom?
Arabic	متى تزهر أشجار اللوز؟
Translation	During any month the almonds trees bloom ?

Rare example of wrong translation and right answer

(February)

Conclusions and further work

The performance of a cross-language Arabic-English QA system is very much affected by the translation process.

More machine translators should be used at the same time in order not to rely just on one translation and to choose the best one on a statistical basis.

Very interesting to use the query reformulation technique directly to the Arabic language.

Thank You

Gracias

Grazie

شكرا

धन्यवाद