JIRS
The mother of all the passage retrieval systems for multilingual question answering?

Universidad Politécnica de Valencia

UNIVERSITAT POLITÈCNICA DE VALÈNCIA

José Manuel Gómez Soriano

Emilio Sanchis Arnal

Paolo Rosso

Instituto Nacional de Astrofísica, Óptica y Electrónica

Manuel Montes y Gómez

Index

- Introduction
- N-Gram Models
- Results
- Conclusions and Future Works

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Introduction

Information is increasing every day

It is necesary to dedicate more time to search the relevant information for the user needs

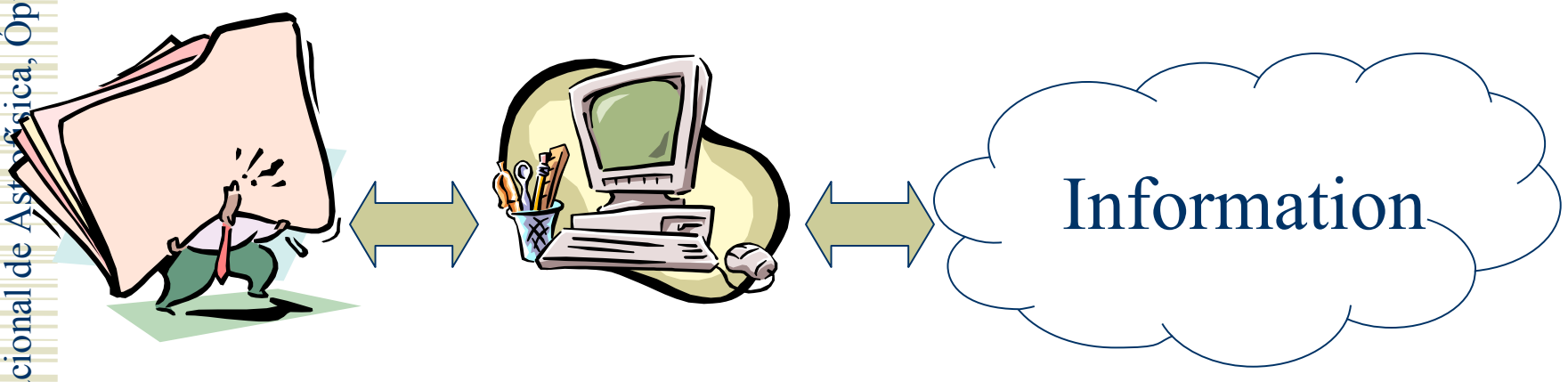
or...

¡¡USE BETTER INFORMATION RETRIEVAL SYSTEMS!!

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

What is an IR system?

It is an application that helps users searching for relevant information.



Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Question Answering Systems

??? Who is the President of Mexico?

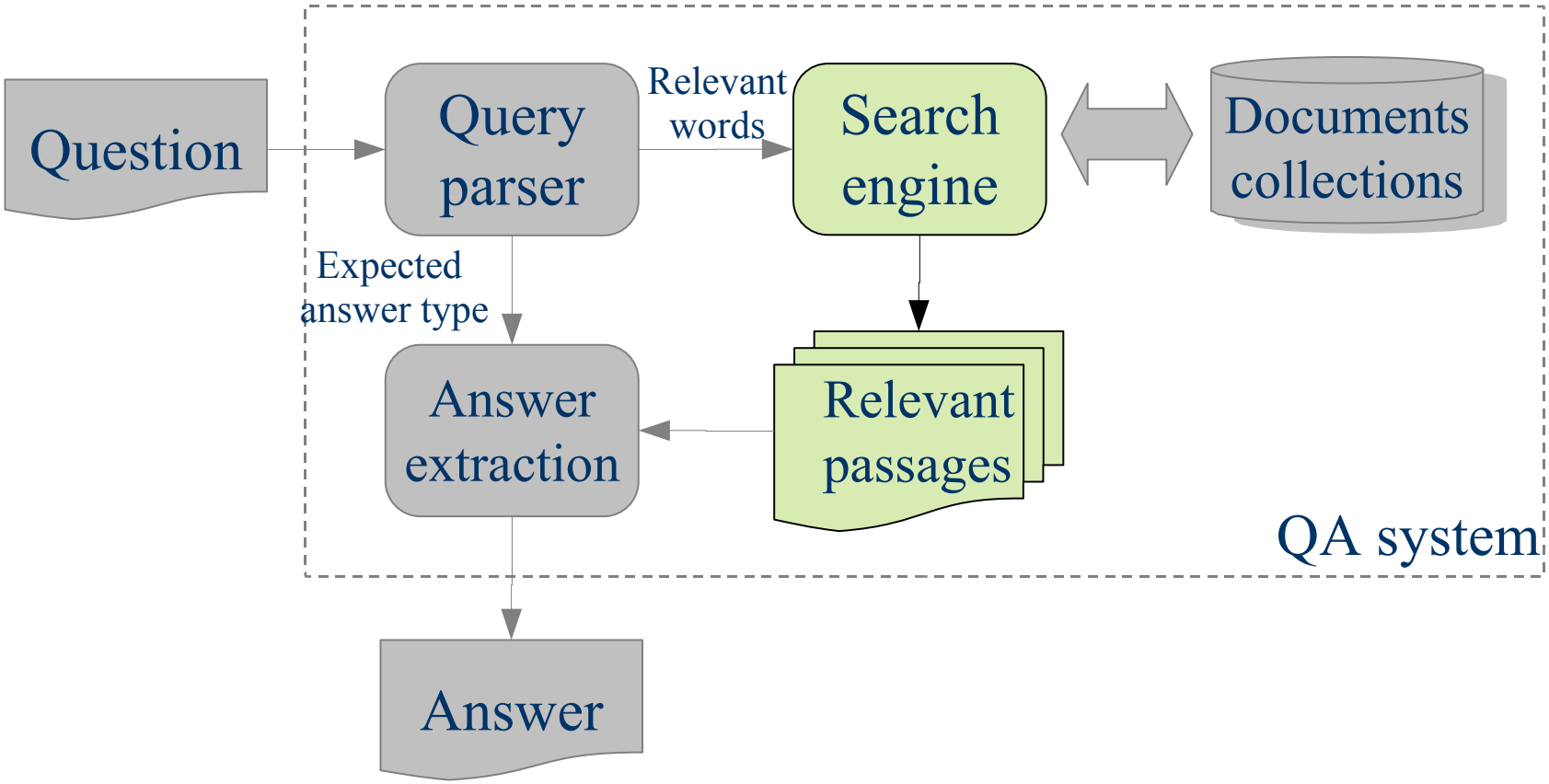
question

Vicente Fox

answer

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Architecture of a typically QA System



Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Passage Retreival systems

- The traditional PR systems only search for key words

  What is the capital of Croatia?

Passage 1

Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are traveling to Belgrade.

Passage 2

Yeltsin invited Tudjman and Milosevic to the capital of Russian to find a political solution to the Croatia and Bosnia conflicts.

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Traditional PR systems II

- Only **54%** of answers can be obtained by using the top 20 passages of the Okapi Passage Retrieval System.

(Gaizauskas et al. 2003)

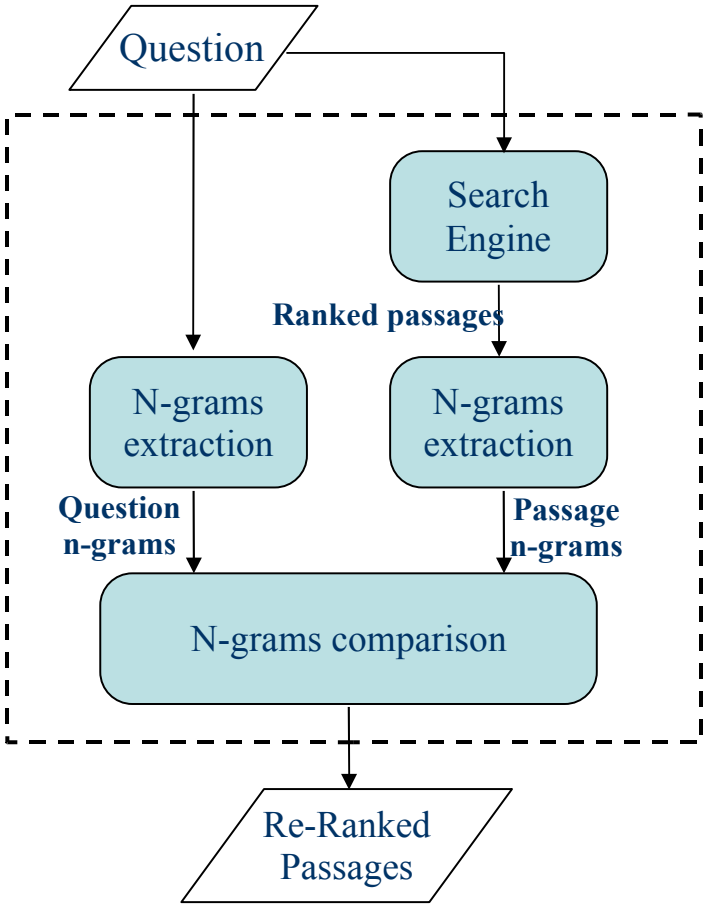
Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

N-Gram Models

- Our method of Passage Retrieval is based on searching the question structures
- We find the passage with the expression more similar to the question
- The longer n-gram of question we find in the passage, the more similarity will have the passage

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

System Architecture



Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Example: Question N-Grams

~~What~~ is the capital of Croatia?	1 x 5-gram
is the capital of the capital of Croatia	2 x 4-gram
is the capital, the capital of, capital of Croatia,	3 x 3-gram
is the, the capital, capital of, of Croatia,	4 x 2-gram
is, the, capital, of, Croatia,	5 x 1-gram

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Example: Passage N-Grams

Passage 1

Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are traveling to Belgrade.

the capital of Croatia	1 x 4-gram
the capital of	2 x 3-gram
capital of Croatia	
the capital	3 x 2-gram
capital of	
of Croatia	
the capital of Croatia	4 x 1-gram

Passage 2

Yeltsin invited Tudjman and Milosevic to the capital of Russian to find a political solution to the Croatia and Bosnia conflicts.

the capital of	1 x 3-gram
the capital	2 x 2-gram
capital of	
the capital of Croatia	4 x 1-gram

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Simple Model

- Similarity between the question and the passage is measured by:

$$sim(d,q) = \frac{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x, D_j)}{\sum_{j=1}^{n} \sum_{\forall x \in Q_j} h(x, Q_j)}$$

where:

$$h(x, D_j) = \begin{cases} 1 & if \ x \in D_j \\ 0 & otherwise \end{cases}$$

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Simple Model Example

What is the capital of Croatia?	
is the capital of Croatia	1
is the capital of	1
the capital of Croatia	1
is the capital	1
the capital of	1
capital of Croatia	1
is the	1
the capital	1
capital of	1
of Croatia	1
is	1
the	1
capital	1
of	1
Croatia	1
	15

Passage 1	
the capital of Croatia	1
the capital of	1
capital of Croatia	1
the capital	1
capital of	1
of Croatia	1
the, capital, of, Croatia	4
	10

0.67

Passage 2	
the capital of	1
the capital	1
capital of	1
the, capital, of, Croatia	4
	7

0.47

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

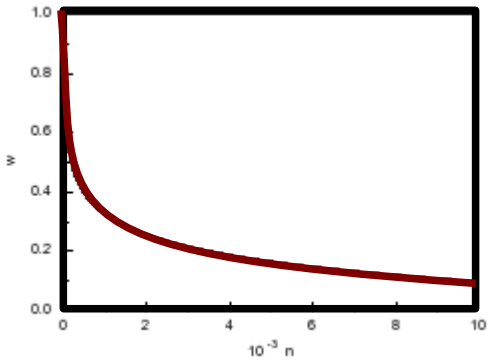
Term Weight Model

- In this model the function *h()* is changed by:

$$h(x, D_j) = \begin{cases} \sum_{k=1}^{j} w_k & if \ x \in D_j \\ 0 & otherwise \end{cases}$$

where:

$$w_k = 1 - \frac{log(n_k)}{1 + log(N)}$$



Term Weight Example

	0.1	0.1	0.3	0.1	0.4
What	**is**	**the**	**capital**	**of**	**Croatia?**

is the capital of Croatia	1
is the capital of	0.6
the capital of Croatia	0.9
is the capital	0.5
the capital of	0.5
capital of Croatia	0.8
is the	0.2
the capital	0.4
capital of	0.4
of Croatia	0.5
is	0.1
the	0.1
capital	0.3
of	0.1
Croatia	0.4
	6.8

Passage 1

the capital of Croatia	0.9
the capital of	0.5
capital of Croatia	0.8
the capital	0.4
capital of	0.4
of Croatia	0.5
the, capital, of, Croatia	0.9
	4.4

0.65

Passage 2

the capital of	0.5
the capital	0.4
capital of	0.4
the, capital, of, Croatia	0.9
	2.2

0.32

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Distance Model

- The problem of previous models is that the weight n-grams is very higher
- Long n-grams are more relevants than short n-grams although these ones have more relevant words
- The Distance Model does not search the longest n-gram
- It searches the heaviest n-grams

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Distance Model Example

Pasaje 1

0.9

Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are traveling to Belgrade.

$$h(x, D_j) = \begin{cases} \sum_{k=1}^{n} w_k \, d(x, x_{max}) & if \;\; x \in D_j \\ 0 & otherwise \end{cases}$$

Pasaje 2

0.3 $D = 7$

Yeltsin invited Tudjman and Milosevic to the Russian capital to find a political solution to the Croatia and Bosnia conflicts.

0.5

Distance factor: $d(x, x_{max}) = \frac{1}{1 + k \cdot ln(1 + D)}$

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Distance Model Example

0.1 0.1 0.3 0.1 0.4
What is the capital of Croatia?
is the capital of Croatia 1

1

Passage 2

Yeltsin invited Tudjman and Milosevic to the Russian capital to find a political solution to the Croatia and Bosnia conflicts. $D = 7$

Passage 1

Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are traveling to Belgrade.

Passage 1
the capital of Croatia **0.9**

0.9

0.9

Passage 2
capital
the Croatia

distance factor

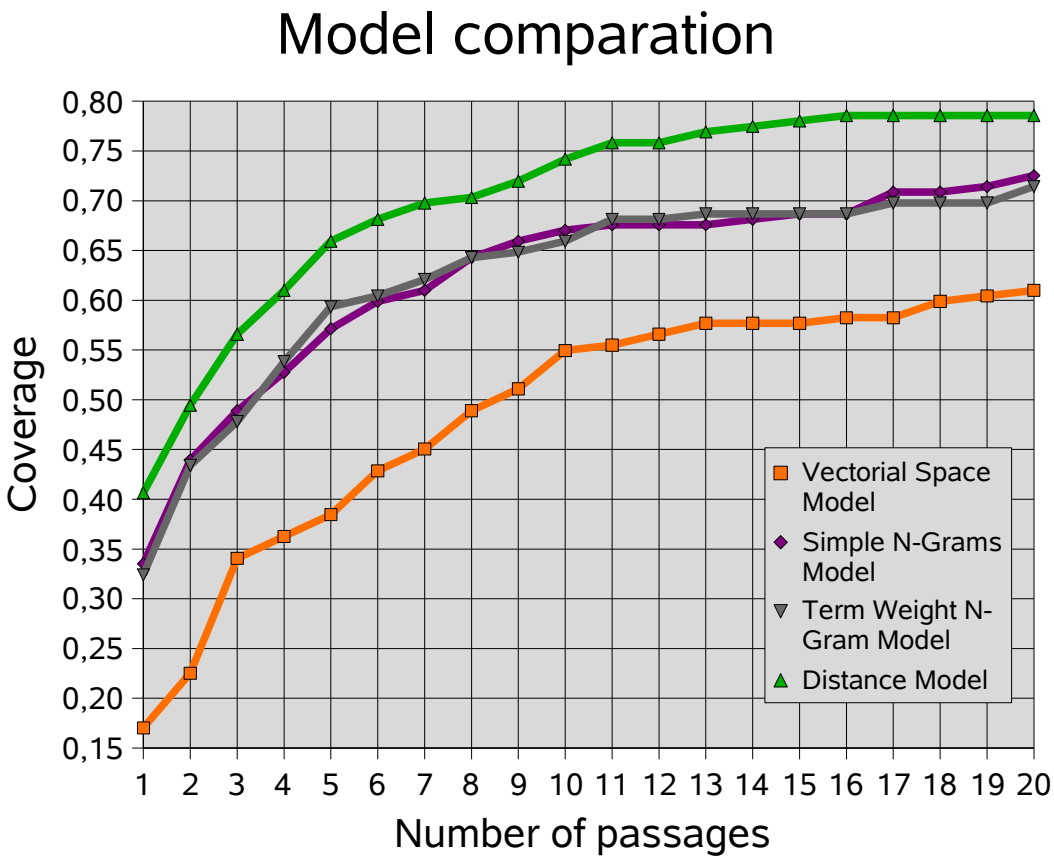
0.3 x 0.32
0.5

0.6

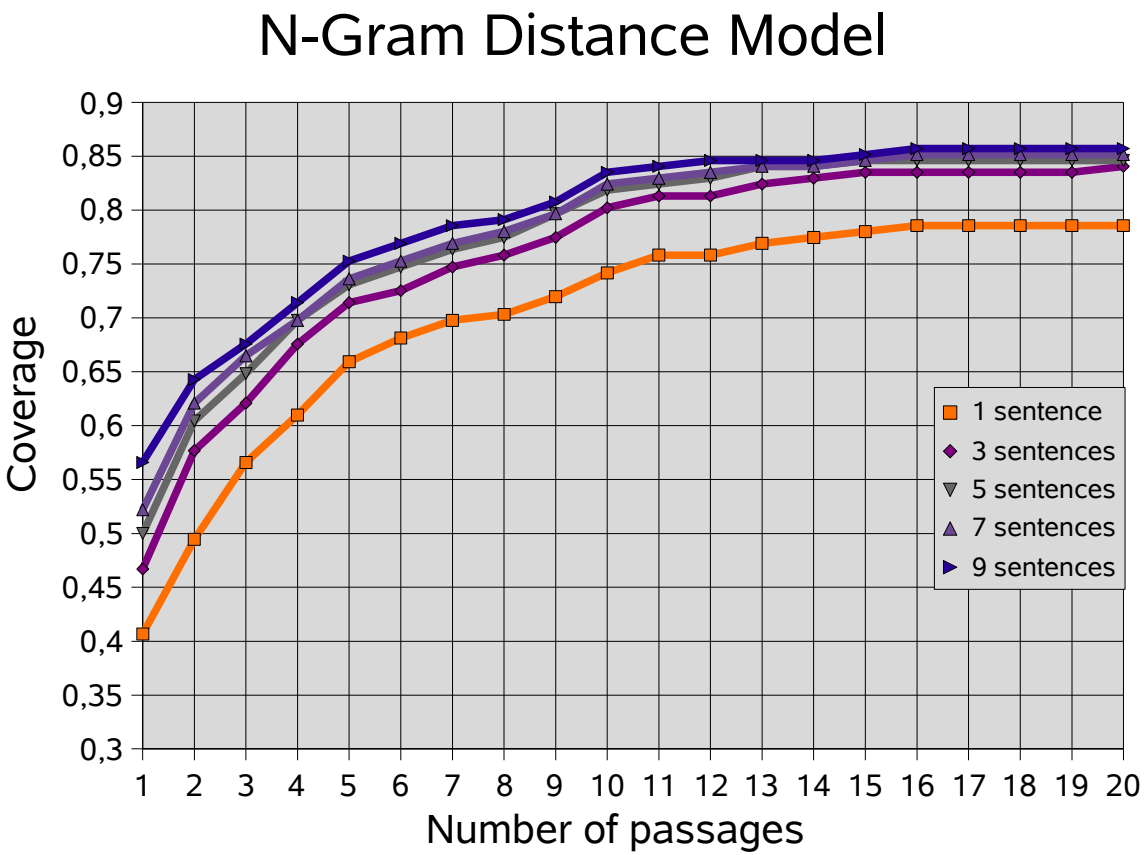
0.6

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Coverage

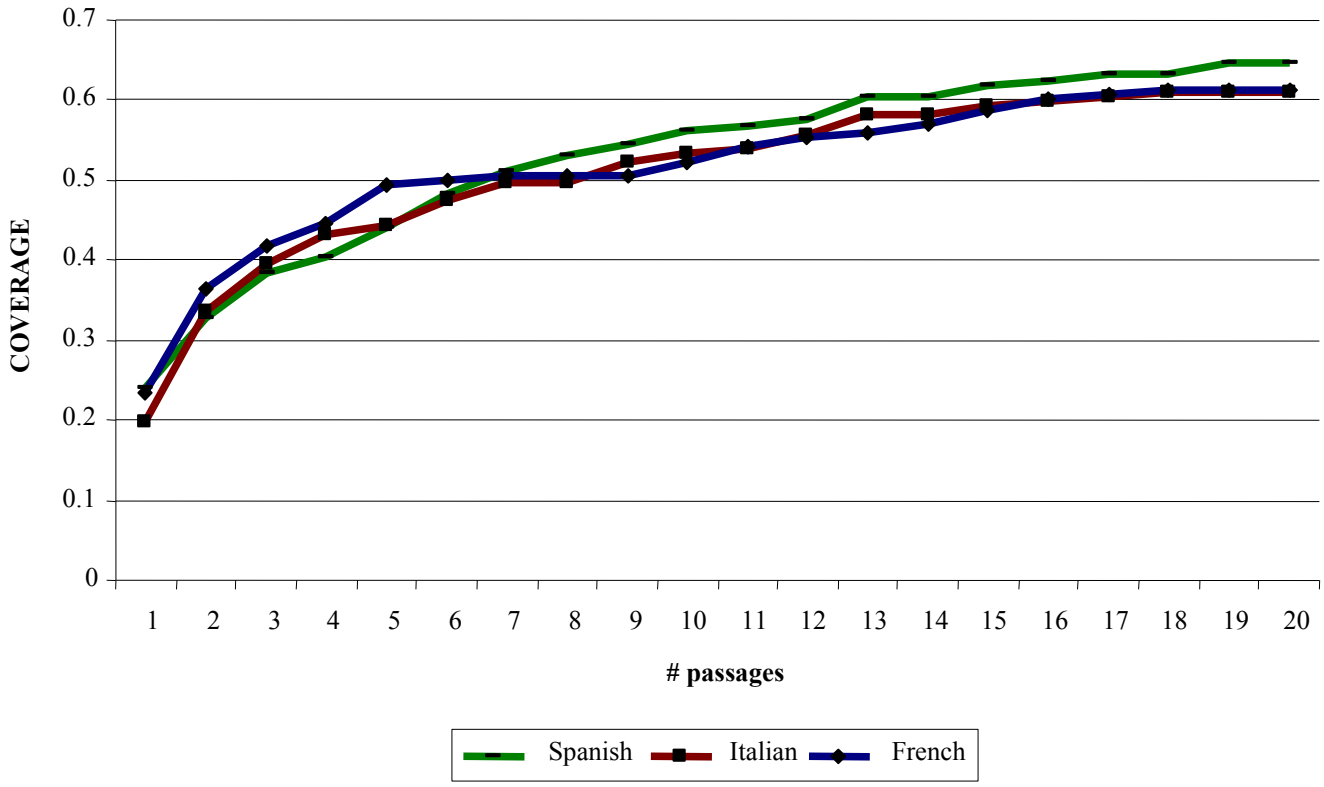

Model comparation
Coverage
0,80
0,75
0,70
0,65
0,60
0,55
0,50
0,45
0,40
0,35
0,30
0,25
0,20
0,15
1 2 3 4 5 6 7 8 9 10 11 12 13 14 15 16 17 18 19 20
Number of passages
Vectorial Space Model
Simple N-Grams Model
Term Weight N-Gram Model
Distance Model

N-Gram Distance Model



Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Multilingual results



Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

CLEF results

Español -> Español	
inao051eses	**42,00%**
tova051eses	**41,00%**
inao052eses	**39,50%**
tova052eses	**38,50%**
upv051eses	**33,50%**
alia051eses	33,00%
aliv051eses	32,50%
alia052eses	30,00%
talp051eses	29,00%
tallp052eses	27,00%
mira051eses	25,50%
mira052eses	23,00%
upv052eses	**18,00%**

Inglés -> Español	
upv051enes	**22,50%**
mira052enes	19,50%
mira051enes	19,50%

Italiano -> Italiano	
tova052itit	**27,50%**
tova051itit	**26,50%**
upv051itit	**25,50%**
upv052itit	**24,00%**
irst051itit	22,00%
irst052itit	19,00%

Francés -> Francés	
syna051frfr	64,00%
tova052frfr	**35,00%**
tova051frfr	**34,50%**
upv051frfr	**23,00%**
hels051frfr	17,50%
upv052frfr	**17,00%**
lire051frfr	16,50%
hels052frfr	16,50%
lina051frfr	14,50%
lcea051frfr	14,00%

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Conclusions

- The N-Gram Models notably improve the coverage of the PR systems
- High redundancy of the answers
- Low computational cost
- Language independent
- Our method fails when there are differences between the question and the answer

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)

Thank you

- © Copyright 2005
- JIRS is a GNU project
- Can be downloaded at
  - http://leto.dsic.upv.es:8080/jirs

Universidad Politécnica de Valencia (UPV)
Instituto Nacional de Astrofísica, Óptica y Electrónica (INAOE)