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

Universidad Politécnica
de Valencia

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
Introduction

Information is increasing every day

It is necessary to dedicate more time to search the relevant information for the user needs or...

¡¡USE BETTER INFORMATION RETRIEVAL SYSTEMS!!
What is an IR system?

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

Who is the President of Mexico?

Vicente Fox

Information

question

answer
Architecture of a typically QA System

Question -> Query parser -> Search engine

Expected answer type

Documents collections

Relevant passages

Answer extraction

QA system

Answer
Passage Retrieval systems

- The traditional PR systems only search for key words

What is the capital of Croatia?

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

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

Question

Search Engine

Ranked passages

N-grams extraction

N-grams extraction

Question n-grams

Passage n-grams

N-grams comparison

Re-Ranked Passages
Example: Question N-Grams

What is the capital of Croatia?  

1 x 5-gram  

is the capital of  

2 x 4-gram  

the capital of Croatia  

3 x 3-gram  

is the capital  

capital of Croatia  

4 x 2-gram  

is the  

capital of  

5 x 1-gram  

of Croatia  

is the capital of Croatia
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.

Passage 2

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

N-Grams

The capital of Croatia 1 x 4-gram

The capital of 2 x 3-gram

Capital of Croatia 2 x 3-gram

The capital 3 x 2-gram

Capital of 3 x 2-gram

Of Croatia 4 x 1-gram

The capital of Croatia 4 x 1-gram
Simple Model

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

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

where:

\[
h(x,D_j) = \begin{cases} 
1 & \text{if } x \in D_j \\
0 & \text{otherwise}
\end{cases}
\]
## Simple Model Example

**What is the capital of Croatia?**
- is the capital of Croatia
- is the capital of
- the capital of Croatia
- is the capital
- the capital of
- capital of Croatia
- is the
- the capital
- capital of
- of Croatia
- is
- the capital
- capital of
- of Croatia
- the, capital, of
- Croatia

**Passage 1**
- the capital of Croatia
- the capital of
- capital of Croatia
- the capital
- capital of
- of Croatia
- the, capital, of, Croatia

**Passage 2**
- the capital of
- the capital
- capital of
- the, capital, of, Croatia

### INTRODUCTION

**N-GRAM MODELS**

**RESULTS**

**CONCLUSIONS**
Term Weight Model

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

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

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

Passage 2
the capital of 0.5
the capital 0.4
capital of 0.4
the, capital, of, Croatia 0.9

2.2

0.65

0.32
The problem of previous models is that the weight n-grams is very higher.

Long n-grams are more relevant 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.
Distance Model Example

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

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

Distance factor:

\[ d(x, x_{\text{max}}) = \frac{1}{1 + k \cdot \ln(1 + D)} \]
Distance Model Example

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 Russian capital to find a political solution to the Croatia and Bosnia conflicts. \( D = 7 \)
Coverage

Model comparison

Number of passages

Coverage
N-Gram Distance Model

![Graph showing the coverage of N-Gram Distance Model with different number of passages.](image)
Multilingual results

![Graph showing coverage vs. number of passages for Spanish, Italian, and French]
## CLEF results

<table>
<thead>
<tr>
<th>Español -&gt; Español</th>
<th>Inglés -&gt; Español</th>
<th>Francés -&gt; Francés</th>
</tr>
</thead>
<tbody>
<tr>
<td>inao051eses</td>
<td>42,00%</td>
<td>upv051enes</td>
</tr>
<tr>
<td>tova051eses</td>
<td>41,00%</td>
<td>mira052enes</td>
</tr>
<tr>
<td>inao052eses</td>
<td>39,50%</td>
<td>mira051enes</td>
</tr>
<tr>
<td>tova052eses</td>
<td>38,50%</td>
<td></td>
</tr>
<tr>
<td>upv051eses</td>
<td>33,50%</td>
<td></td>
</tr>
<tr>
<td>alia051eses</td>
<td>33,00%</td>
<td></td>
</tr>
<tr>
<td>aliv051eses</td>
<td>32,50%</td>
<td></td>
</tr>
<tr>
<td>alia052eses</td>
<td>30,00%</td>
<td></td>
</tr>
<tr>
<td>talp051eses</td>
<td>29,00%</td>
<td></td>
</tr>
<tr>
<td>tallp052eses</td>
<td>27,00%</td>
<td></td>
</tr>
<tr>
<td>mira051eses</td>
<td>25,50%</td>
<td></td>
</tr>
<tr>
<td>mira052eses</td>
<td>23,00%</td>
<td></td>
</tr>
<tr>
<td>upv052eses</td>
<td>18,00%</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Italiano -&gt; Italiano</td>
<td></td>
<td></td>
</tr>
<tr>
<td>tova052itit</td>
<td>27,50%</td>
<td></td>
</tr>
<tr>
<td>tova051itit</td>
<td>26,50%</td>
<td></td>
</tr>
<tr>
<td>upv051itit</td>
<td>25,50%</td>
<td></td>
</tr>
<tr>
<td>upv052itit</td>
<td>24,00%</td>
<td></td>
</tr>
<tr>
<td>ist051itit</td>
<td>22,00%</td>
<td></td>
</tr>
<tr>
<td>ist052itit</td>
<td>19,00%</td>
<td></td>
</tr>
</tbody>
</table>
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
Thank you

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