Re-ranking of Yahoo Snippets with the JIRS Passage Retrieval System
Index

- Introduction
- Description of JIRS PR System
- Preliminary results
- Conclusions
Introduction

Huge repository of information

Internet

Relevant information

Irrelevant information

Incorrect information

http://jirs.dsic.upv.es
Redundancy

“Certain repetition of the information contained in the collection of documents or Web, which allows, in spite of the loss of a part of the information, to reconstruct its content”

- The redundancy can help to retrieve the correct information but also increases the irrelevant information
Question Answering Systems

Who is the President of Mexico?

Vicente Fox

Document collection or Internet
Architecture of a typically QA System

1. User question
2. Question analysis
3. Passage retrieval
4. Answer extraction
5. Ranked passages
6. Ranked answers
7. Document collection
Passage Retrieval in QA

User question → Passage Retrieval → Ranked Passages

- Relevant information but WITHOUT answer
- Relevant information but INCORRECT answer
- Relevant information WITH answer
Traditional PR 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 travelling to Belgrade.

Passage 2

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

- PR system specially suited for QA
- Takes advantage of the collection redundancy
- It searches question structure instead of question key words
- Language-Independent:
  - English, Spanish, French, Italian, Arabic, Urdu, Oromo
- Open source
- Download port: http://jirs.dsic.upv.es
Features of JIRS PR System

- Modular
- Hierarchical Structure
- Scalability
- Adaptability
- Configurable

- Client/Server
- Distributed
- Language-Independent
- Multiplataform
Modified Architecture to use Yahoo Search Engine

User question → Yahoo Search Engine → Document Collection

Ranked Snippets

N-gram Extraction

N-gram Similarity

Question N-grams

Snippets N-grams

N-gram based snippets ranking subsystem

Re-ranked Snippets

http://jirs.dsic.upv.es
Question N-gram Extraction

What is the capital of Croatia?

- is the capital of Croatia (1-gram)
- is the capital of (2-grams)
- the capital of Croatia (3-grams)
- is the capital (4-grams)
- the capital of (5-grams)

- capital of Croatia
- is, the, capital, of, Croatia
Snippet 1
Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are travelling to Belgrade.

Snippet 2
Yeltsin invited Tudjman and Milosevic to the capital of Russia to find a political solution to the Croatia and Bosnia conflicts.
Yesterday, the delegation visited Zagreb, the capital of Croatia, and after their stay in Sarajevo they are travelling to Belgrade.

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

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

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

Example

Snippet 1
What is the capital of Croatia?

Snippet 2
Yeltsin invited Tudjman and Milosevic to the capital of Russia to find a political solution to the Croatia and Bosnia conflicts.

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

Distance factor

\[ L = 7 \]
Coverage Comparative

![Coverage Comparative Graph](http://jirs.dsic.upv.es)
Searching in the CLEF Document Collection

![Graph showing coverage as a function of number of passages for different numbers of sentences.]

- Coverage increases with the number of passages.
- The graph includes lines for 1 sentence, 3 sentences, 5 sentences, 7 sentences, and 9 sentences.
- Coverage values range from 0.4 to 0.9.

http://jirs.dsic.upv.es
Conclusions

- The preliminary results are not good
- Likely due to the following reasons:
  - The answer set are not up-to-date answers
  - The snippets are not complete sentences
- Solution
  - Update the answer set
  - Do not use snippets → use entire documents
Conclusions

- JIRS PR System is specially suited for QA
- It does not try to obtain relevant passages but tries to retrieve passages with the correct answer
- JIRS has already been adapted to English, Spanish, French, Italian, Arabic, Urdu and Oromo languages.
- JIRS is better with a high redundancy document collection
- Language-Independent
Thank you very much

- JIRS © Copyright 2005
- GNU General Public License
  - [http://opensource.org/licenses/gpl-license.php](http://opensource.org/licenses/gpl-license.php)
- Can be downloaded at
  - [http://jirs.dsic.upv.es](http://jirs.dsic.upv.es)