ugr | Universidad de Granada

CLEF 2011 Amsterdam

Multilingual Question-Answering System in biomedical domain on the Web: an evaluation

María-Dolores Olvera-Lobo & Juncal Gutiérrez-Artacho

QA Systems: Beyond Information Retrieval

- **Medical specialists** invest an average of more than **two minutes** searching for information related to questions that arise and, despite the time taken up, **adequate answers are often not found**.

- **QAS** are designed to offer **understandable responses** to factual questions of specialized content **rapidly** and **precisely** in such a way that the user does not have to read the complete documents to satisfy a particular query.

- QAS try to **overcome the limitations** of the traditional tools of information retrieval, such as the consultations being **monolingual**.

- Although the Cross-lingual QAS of restricted domain are not yet avalaible for the final users, on the Web it begins to find some on the sphere of **multilingual** (such as HONqa).

ugr | Universidad de Granada

Method

About the questions

- The **English** questions were obtained from the website **WebMD**
- They were formulated as consultations of the type **"What is"** in the search engine of the website
- The questions were translated by a team of professional translators to **French** and **Italian**
- The 120 questions that elicited responses in the **three languages** in the QA system were selected
- Finally, we used **a set of 360 definitional biomedical questions in three** languages for the evaluation
- The set of questions used passed the **validity test** with a Chronbach's alpha of 0.936

http://www.webmd.com/

Method

About de QA system

- **HONqa** was developed by the *Health On the Net Foundation*.
- It is a **multilingual system** that retrieves information in **English**, **French**, and **Italian**.

HONQA

www.hon.ch/QA/

Method

About the measures

The **responses** offered by the system were evaluated by a group of **experts** from different medical areas as:

- **Correct:** Questions that were answered properly and did not add irrelevant information
- **Incorrect:** Answers that contained irrelevant information with regard to the question
- **Inexact**: All the answers that resolved the question but added irrelevant information

Method

About the measures

For the analysis of the answers retrieved, the applied **evaluation measures** were:

- ***Mean Reciprocal Rank*** **(MRR),** which assigns the inverse value of the **position** in which the **correct answer** is found, or 0 if there is no correct response

$$MRR = \frac{1}{q}\sum_{i=1}^{q}\frac{1}{far_i}$$

- ***Total Reciprocal Rank*** **(TRR),** useful for evaluating the existence of **several correct responses** offered by a system to the same query;

- ***First Hit Success*** **(FHS),** which assigns a value of 1 if the **first answer** offered is correct, and a value of 0 if it is not

ugr | Universidad de Granada

Method

About the measures

Measures related to precision

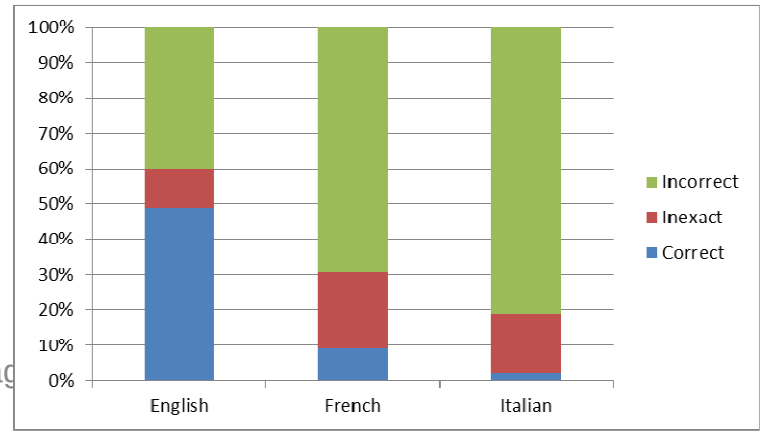
- ***Precision*** $$\text{Precision} = \frac{|\{\text{relevant answers}\} \cap \{\text{retrieved answers}\}|}{\{\text{retrieved answers}\}}$$
- ***Precision considering also the inexact answers (P*)***
- ***Precision of the 3 first results (P@3)***
- ***Precision of the 3 first results including inexact answers (P@3*)***
- ***Mean Average Precision*** **(MAP),** which measures the average precision of a set of queries for which the answers are arranged by relevance. $$\text{MAP} = \frac{\sum_{q=1}^{Q} \text{AveP}(q)}{Q}$$

ugr | Universidad de Granada

Results

	Total answers	Average of Answers	Answers analyzed	Correct answers	Inexact answers	Incorrect answers
English	5695	47.46	589	287 (48.73%)	67 (11.4%)	235 (39.9%)
French	3283	27.36	573	52 (9.07%)	124 (21.6%)	397 (69.3%)
Italian	3123	25.03	585	32 (5.47%)	95 (11.6%)	458 (82.9%)

- In **English** the *volume of answers* retrieved (5695 and 47.46 of average) was substantially higher than in **French** (3283 answers and average of 27.36), and for **Italian,** (3123 and 25.03) it has been registered similar values.
- We analyze the first five answers for each posed question
- The *correct answers* were present in greater measure in the **English** version of the system, which properly responded to more than **48**% of the cases, whereas **French** offered a low rate of **9.07**% and **Italian** provided only **5.47**%.
- The *incorrect answers* was **very high** in all three languages, **exceeding 50%** of the total in **French** (69.3%) and **Italian** (82.9%).
- The *inexact answers* was higher in **French** (21.64%), followed by **Italian** (**11.6%**).



Results

	MRR	TRR	FHS	P	P*	P@3	P@3*	MAP
English	0.76	1.55	0.575	0.55	0.65	0.57	0.67	0.25
French	0.19	0.27	0.12	0.10	0.31	0.11	0.32	0.05
Italian	0.13	0.15	0.06	0.05	0.16	0.06	0.15	0.03

- **MRR** indicates that while the results of the **English** option were quite plausible, at 0.76, the other two languages offered very poor results (0.19 for French and 0.13 for Italian), indicating the low reliability of the first answers for these languages.

- In **TRR measure**, which considers all the answers correct among the first 5 results analysed, it was found that, except for **English,** the results did not substantially improve.

•**FHS** is an important measure, as the users often tend to focus on the **first** response retrieved, skipping the rest. It was found that more than 50% of the answers offered in **English** (0.575) provided an initial correct answer while the other cases were not encouraging (0.12 in French and 0.06 in Italian).

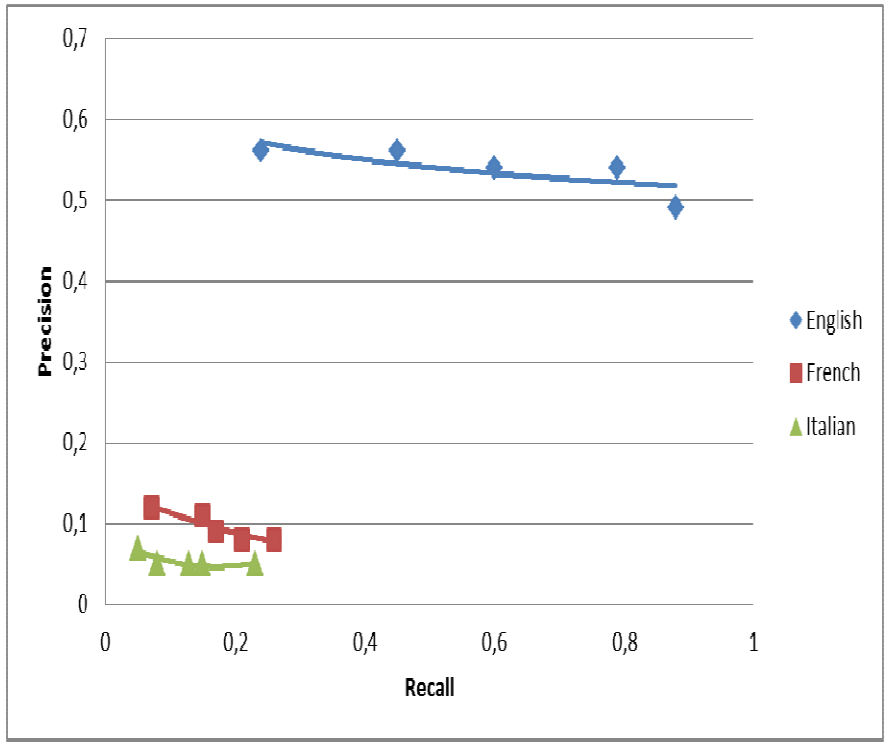
•The results indicate that the **arrangement of the answers** retrieved according to their relevance to the question was not the best.

ugr | Universidad de Granada

Results

	MRR	TRR	FHS	P	P*	P@3	P@3*	MAP
English	0.76	1.55	0.575	0.55	0.65	0.57	0.67	0.25
French	0.19	0.27	0.12	0.10	0.31	0.11	0.32	0.05
Italian	0.13	0.15	0.06	0.05	0.16	0.06	0.15	0.03

- **Precision** was measured considering as relevant only the responses scored as **correct** (measures **P** and **P@3**) and consider also the **imprecise** answers (measures **P*** and **P@3***) as relevant –that is, being more flexible to evaluate an answer as adequate.
- In this latter case, clearly, the precision values significantly **increased** in some cases.
- As with the rest of the measures, **there was a marked different between English and the other languages**



Conclusions

- The analysis of the results from posing **360 questions** in the QA system of the **biomedical domain** HONqa has enabled the evaluation of its performance in the retrieval of **multilingual information** by applying **specific measures**

- Most of **information sources** used by the QA system to **extract the answers** are **portals** or **websites specializing in medical topics** though the type of portal differed from one language to the other:
  - In **English**, most of the portals presented their content in the form of **questions**, posed by the developers themselves or the users of the system, and their corresponding **answers**.
  - In the portals used to extract anwers for the **other languages**, the information offered was of quality but not all showed definitions or information relevant to diseases, treatments, etc.
- Therefore, this is probably one of the causes of the great differences in the results for the different languages.

ugr Universidad de Granada

Conclusions

- Despite the **restrictions** that these systems show, the study indicates that this QA system is **valid** and **useful** for the retrieval of **definitional medical information**, mainly in the **English language**, although it is not yet the most advisable resource to gather multilingual information in a quick and precise way.

- The search for **multilingual answers** in the context of the **Web** still **needs to progress** a long way to reach the effectiveness levels of general retrieval systems, and especially in monolingual ones.

- Nevertheless, the **results are promising** as they show this type of tool to be a new possibility within the sphere of precise, reliable, and specific information retrieval in a brief period of time.

ugr Universidad de Granada