Multilingual Question-Answering System in Biomedical Domain on the Web: An Evaluation

María-Dolores Olvera-Lobo[1, 2] and Juncal Gutiérrez-Artacho[2]

[1] CSIC, Unidad Asociada Grupo SCImago, Madrid, Spain
[2] University of Granada, Spain
{molvera,juncalgutierrez}@ugr.es

Abstract. Question-answering systems (QAS) are presented as an alternative to traditional systems of information retrieval, intended to offer precise responses to factual questions. An analysis has been made of the results offered by the QA multilingual biomedical system HONqa, available on the Web. The study has used a set of 120 biomedical definitional questions (*What is…?*), taken from the medical website WebMD, which were formulated in English, French, and Italian. The answers have been analysed using a serie of specific measures (MRR, TRR, FHS, precision, MAP).

The study confirms that for all the languages analysed the functioning effectiveness needs to be improved, although in the multilingual context analysed the questions in the English language achieve better results for retrieving definitional information than in French and Italian.

Keywords: Multilingual information, Multilingual Question Answering Systems, Restricted-domain Question Answering Systems, HONqa, Biomedical information, Evaluation measures.

1 Introduction

The advent of the Web and its subsequent expansion has provided the general public with access to enormous volumes of information, offering unquestionable benefits. Nevertheless, this has also brought disadvantages such as overloads of information — which in this environment is even more acute— or the fact that much of the information is incorrect, incomplete, or inaccurate, whether intentionally so or not. Consequently, it becomes indispensable to develop tools and procedures that enable the user to acquire reliable information that is relevant for a particular consultation. This is the challenge that faces Information Retrieval (hereafter IR). Some of the efforts to improve IR in the Web have focused on the design and development of question-answering systems (hereafter QAS).

This work evaluates the multilingual search for definitional responses in the context of restricted-domain QAS. For this, the results offered by the multilingual biomedical QAS HONqa, available on the Web, were analysed. A set of 120 responses (in English, French, and Italian) related to this thematic field were assessed by a series of measures applicable to such systems.

P. Forner et al. (Eds.): CLEF 2011, LNCS 6941, pp. 83–88, 2011.

2 Question Answering Systems: Beyond Information Retrieval

According to a study by Ely [1], 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. In this sense, several works have demonstrated the confidence of medical specialists in the use of QAS as a method of searching and retrieving specialized information [2-3]. Patients have also increasingly consulted these systems, before and after seeing the doctor, to gather information on the nature of the illness, treatment recommendations, contraindications, etc. [4]

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. These systems begin with the user's question in order to construct coherent answers in everyday language [5]. The functioning of the QAS is based on short-answer models [6], since the potentially correct answer does not go beyond a number, a noun, a short phrase, or a brief fragment of text. Then the QAS locates and extracts one or several responses from different sources, according to the subject of the consultation [7]. Finally an evaluation is made and information that is redundant or that does not answer the question properly is eliminated in order to present specific responses designed to satisfy the information needed by the user [8-10].

For the environment of the Web, some of QAS have been developed only as prototypes or demos, and are very rare in systems available to the final user. Nonetheless, some QAS of a general domain can currently be consulted, as they are capable of addressing questions on very diverse subjects (such as START) and from very specific domains (such as EAGLi), which focus on a given context. In addition, the QA systems try to overcome the limitations of the traditional tools of information retrieval, such as the consultations being monolingual.

The appropriate retrieval tools that enable the procedure known as cross-lingual information in retrieval (CLIR) would enable consultations in several languages with information retrieval in all the languages accepted by the system [11]. 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 QAS (such as HONqa).

3 Method Section

The methodology applied used 120 biomedical questions concerning definitions on diverse medical subjects. The questions were formulated as consultations of the type "What is" in the search engine of the website WebMD, created in the USA by medical specialists to resolve doubts held by patients. From the questions that this portal provides, a set was selected to be translated by a team of professional translators of French and Italian, and from this initial set, 120 questions that elicited responses in the three languages in the system were selected. These constituted the body of the questions used. The main biomedical aspects related to the selected questions were diseases, operations, treatments, syndromes, and symptoms. The set of questions used passed the validity test with a Chronbach's alpha of 0.936. HONqa, the QAS evaluated in this work, was developed by the *Health On the Net*

Foundation. It is a multilingual system that retrieves information in English, French, and Italian [12].

The responses offered by the system were evaluated by a group of experts from different medical areas, as incorrect, inexact, or correct, according to the evaluation methodology proposed in CLEF [13]. Questions that were answered properly and did not add irrelevant information were considered correct. All the answers that resolved the question but added irrelevant information were considered inexact. Finally, answers that contained irrelevant information with regard to the question were considered incorrect. From the evaluation of the answers retrieved, the evaluation measures were applied are: *Mean Reciprocal Rank* (MRR), which assigns the inverse value of the position in which the correct answer is found, or zero if there is no correct response; *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; *Precision,* which measures the ratio of retrieval responses are relevant to the query*;* and *Mean Average Precision* (MAP), which measures the average precision of a set of queries for which the answers are arranged by relevance.

The users access to the QA systems for their quickness and precision, and they are not likely to read a long list of answers for each question. So their futility point [14-16] –the maximum number of responses they would be willing to begin browsing through–, will be probably more exigent than others Information Retrieval Systems. For this reason, only the first five answers in each of these systems were analysed – although the mean of the answers retrieved by the system in the three languages approached and in some cases exceeded.

4 Results Section

The average of the total answers retrieved by the system was 47.46 in the case of English, 27.36 for French, and 25.03 for Italian.

Table 1. Answers retrieved by HONqa in the three languages

	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%)

The volume of answers retrieved in English was substantially higher (5695 answers retrieved) than in the other cases, the other two languages registering similar values (3283 for French and 3123 for Italian).

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 2.05%. The number of imprecise answers was higher in French (21.64%), followed by Italian (16.24%). In relation to the incorrect answers, the number was very high in all three languages, exceeding 50% of the total in French (397) and Italian (458).

This behaviour directly influenced the results found when applying the evaluation measures proposed. The MRR value for the responses offered in the three languages reflect the above comments. 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 response by the system for these languages.

In relation to the 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). MAP is a widely used measure that offers an overall idea of the functioning of the system. The evaluation of the system did not register an adequate level for any of the languages analysed.

Table 2. Evaluation measures (P=Precision, P*=Precision considering also the inexact answers, P@3=Precision of the 3 first results, P@3*=Precision of the 3 first results including inexact answers)

	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

The precision value is closely related to the rest of the measures discussed above. The precision was measured, on the one hand, considering as relevant only the responses scored as correct (measures P and P@3) and, on the other hand, considering 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. Nevertheless, as with the rest of the measures, there was a marked different between English and the other languages. On considering P@3 or only the precision of the first three results, the values found for this new measure only weakly improved though not very different from the previous values. This indicates that the arrangement of the answers retrieved according to their relevance to the question was not the best. The small number of correct answers in some cases made the recall values of the QAS very low, except in the case of English.

5 Conclusions

The analysis of the results from posing 120 questions in the QA system of the biomedical domain HONqa has enabled the evaluation of its functioning in the retrieval of multilingual information by applying specific measures and analysing the

information sources used for each language. 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 responses 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.

References

1. Ely, J.W., Osheroff, P.N., Ebell, M., Bergus, G., Barcey, L., Chambliss, M., Evans, E.: Analysis of questions asked by family doctors regarding patient care. British Medical Journal 319, 358–361 (1999)
2. Lee, M., Cimino, J., Zhu, H.R., Sable, C., Shanker, V., Ely, J., Yu, H.: Beyond Information Retrieval –Medical Question Answering. AMIA, Washington DC (2006)
3. Yu, H., Kaufman, D.: A cognitive evaluation of four online search engines for answering definitional questions posed by physicians. In: Pacific Symposium on Biocomputing, vol. 12, pp. 328–339 (2007)
4. Zweigenbaum, P.: Question answering in biomedicine. In: Rijke, Webber (eds.) Proceedings Workshop on Natural Language Processing for Question Answering, EACL 2003, pp. 1–4. ACL, Budapest (2005)
5. Costa L.F., Santos, D.: Question Answering Systems: a partial answer (SINTEF, Oslo) (2007)
6. Blair-Goldensohn, S., McKeown, K., Schlaikjer, A.H.: Answering Definitional Questions: A Hybrid Approach. New Directions in Question Answering 4, 47–58 (2004)
7. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Language resources used in Multi-lingual Question Answering Systems. Online Information Review 35(4) (forthcoming, 2011)
8. Cui, H., Kan, M.Y., Chua, T.S., Xiao, J.: A Comparative Study on Sentence Retrieval for Definitional Question Answering. In: SIGIR Workshop on Information retrieval for Question Answering (IR4QA), Sheffield (2004)
9. Tsur, O.: Definitional Question-Answering Using Trainable Text Classifiers. PhD Thesis. University of Amsterdam (2003)
10. Olvera-Lobo, M.D., Gutiérrez-Artacho, J.: Question-Answering Systems as Efficient Sources of Terminological Information: Evaluation. Health Information and Library Journal 27(4), 268–274 (2010)
11. Diekema, A. R.: Translation Events in Cross-Language Information Retrieval: Lexical ambiguity, lexical holes, vocabulary mismatch, and correct translations. PhD Thesis. University of Syracuse (2003)
12. Cruchet, S., Gaudinat, A., Rindflesch, T., Boyer, C.: What about trust in the Question Answering world? In: AMIA 2009 Annual Symposium, San Francisco (2009)
13. Blair, D.C.: Searching biases in large interactive document retrieval systems. Journal of the American Society for Information Science 31(4), 271–277 (1980)

14. Peters, C.: What Happened in CLEF 2009: Introduction to the Working Notes. In: Peters, C., Di Nunzio, G.M., Kurimo, M., Mostefa, D., Penas, A., Roda, G. (eds.) CLEF 2009. LNCS, vol. 6241, pp. 1–12. Springer, Heidelberg (2010), http://www.clefcampaign.org/2009/working_notes/CLEF2009-intro.pdf
15. Raved, D.R., Qi, H., Wu, H. Fan, W.: Evaluating Web-based Question Answering Systems. Technical Report, University of Michigan (2001)
16. Salton, G., Mc Gill, M.J.: Introduction to Modern Information Retrieval. Mc Graw-Hill, New York (1983)