Multilingual Question/Answering System applied to trusted health information

Sarah CRUCHET[a], Arnaud GAUDINAT[a], Célia BOYER[a], Angelica MORANDI[b], Daniela MARINO[b]
[a] *Health On the Net Foundation, Geneva,* [b] *San Raffaele hospital,* Milan*o, Italia*

Abstract. *The Health On the Net Foundation is developing a multilingual question answering system adapted to trusted health information. The system is using supervised method to classify questions according to its medical type and its type of excepted answer. The French and English, the first languages added, have already been presented [6]. This paper enhances the tedious work of the addition of another language: the Italian. Medical questions have been collected on the Internet by the HSR and then classified in accordance to the categories found for French and English. Tests of supervised classification have been made to find the best classifiers for the recognition of Italian's question types. In this paper, the Italian results are compared to the French and English results. Currently, for a set of 100 Italian questions, 71 are well categorized according to the type of answer expected and 65 according the medical type in Italian. These results are similar to these obtained for the classification of French's questions but are lower than the English's ones. It confirms the linguistical differences between the Latin's languages and the English one. Latin languages seem to be more difficult to treat by an automatic tool.*

Keywords. Multilingual system, Supervised classification

Introduction

The medical domain is widely represented on the Web [1]. Health information comes from a wide range of sources – friends and family, general print media, specialist print media, television and radio, and the Internet. A North American survey, conducted in July 2008, found that 76% of all adults have online access and 81% of those online have looked for healthcare information at least once [2]. How can a common user judge the relevance of medical documents?

The Health On the Net Foundation (HON) is a leading organization in promoting and guiding the deployment of useful and reliable online medical and health information. For more than 12 years its code of conduct, the HONcode, is the oldest and the most used ethical code for health and medical information on the Internet [3, 4]. The HONcode targets: the general public, webmasters and medical professionals. It is composed of 8 principles of quality [5] which are respected by all certified websites. It is to answer this double problem that HON became involved in the development of a QA system specific to the medical domain which is a complex domain requiring the limitation of the framework of research. This applied research has been conducted within the European PIPS project [IST–2002-2.3.1.11]. The paper submitted at the MIE 2008 congress [6] exposed the framework of the QA and the results of classifiers for

English and French. Many articles about health QA are found on the Internet [7-11] but only one is available on the Internet since 2007: MedQA [12]. It targets biologists to facilitate their access to literature and performs definitional questions. It should be remembered that most QA use pattern matching to classify questions [13-14]. However HON chooses machine learning which is known to be capable of identifying data a human would not have though and whatever the language [15]. The QA developed by HON is intended for the citizen both patients or health professionals available in English, French and newly in Italian. In addition the user can choose the domain of research i.e. websites certified by HON or all the websites. By default the research is done in the database of certified websites as the quality of the available information is paramount, far more so than the quantity.

The following sections present the material and methods used to classify medical questions in Italian, the results obtained for the supervised method and the comparison with the French and English results.

1. Material and Methods

1.1. Corpus

We disposed of 101 questions in Italian. Questions have been collected in Frequently Asked Questions (FAQ) found in specialized health forum as well as in discussion on the Internet by HSR (San Raffaele hospital) respectively to our French and English model.

1.2. Human classification

The medical types and the types of excepted answers have already been defined with the classification of French and English questions [6]. Consequently the health professional in charge of classifying the set of Italian's questions has used the existing categories for both the medical type and the type of excepted answers.

Table 1. Repartition of questions for the medical type

	En	**Fr**	**It**
Causes	8	9	**13**
Diagnostic	6	2	**1**
Diet	18	14	**14**
Disease	25	12	**0**
Evolution	12	11	**6**
Physiology	15	11	**10**
Prevention	6	8	**4**
Routine	10	34	**22**
Symptoms	13	8	**3**
Treatment	**32**	**55**	**23**

Table 2. Repartition of questions for the type of expected answer

	En	Fr	It
Boolean	53	65	**26**
Causal	17	5	**3**
Definition	47	12	**5**
Factoid	6	13	**23**
Duration	4	3	**0**
List	7	15	**14**
Moment	4	6	**4**
Person	7	4	**1**
Place	9	4	**1**
Procedure	4	30	**5**
Quantity	15	7	**14**

1.3. Supervised method

The set of questions defined previously have been divided into 2 parts: the training base (90% of the corpus) and the one for tests (10%). Cross evaluation have been realised. This was done to calculate a score of performance of the automatic tools and its relevance for the task in comparison with the results obtained for the French and the English classifiers [6]. As in our previous study, seven measures were used: macro (ma) and micro (mi) recall (R: Capacity of the system to report only relevant documents)/precision (P: Rate of relevant documents proposed by the system as compared to all the retrieved documents)/F1 and the error. It should be remembered that macro-measures compute values of precision/recall for each category and make a mean on these values; and that micro-measures gather data of each category in a same contingency table and compute values of precision/recall according to this table. The F1-measure is a function that is maximized when the precision and the recall are near. Finally, this set of measures has been calculated according to units of treatment (unigram, bigrams, trigrams and co-occurrences of words).

Table3. Summarization of the classifiers' results

	English Med Type (SVM W2*)	**French Med Type (SVM W2*)**	**Italian Med Type (SVM W2*)**	**English Rep Type (SVM W2*)**	**French Rep Type (NB W3*)**	**Italian Rep Type (SVM W2*)**
maR	0.478	0.475	0.632	0.805	0.774	0.763
maP	0.525	0.512	0.647	0.846	0.686	0.710
maF1	0.477	0.464	0.639	0.815	0.680	0.736
miR	0.694	0.313	0.437	0.779	0.714	0.539
miP	0.410	0.629	0.736	0.919	0.631	0.825
miF1	0.410	0.414	0.548	0.919	0.668	0.652
Erreur	0.078	0.086	0.080	0.027	0.066	0.058

*Wx = x grams of words, Med = Medical and Ans = Answer

2. Results and Discussion

Table 1. and **Table2.** give the repartition of questions by categories. We notice that, as for French and English, this repartition is not homogeneous although categories

were known in advance. It is due to the collection of data which is in accordance with the Internet users' requests. Indeed, the system aims to fit with the end users wishes. The well represented categories for the type of excepted answer in Italian is "Boolean" as for French and English. As a consequence the type of excepted answer which is the most frequent is "Boolean" for all the three languages. It points out that the "yes or no" is the most asked whatever the language used.

Table 3. gives the best results obtained with classifiers for both the medical type and the type of excepted answer. Results obtained for the Italian language are similar to those obtained for French. The system is better for the recognition of the type of excepted answers (71% of macro-precision) than the medical type (64.7% of macro-precision). This unbalance may come from the repartition of questions by categories. **Table 1.** shows that, in our set of questions, there are only 1 question of "Diagnostic", 3 of "Symptoms" and 0 of "Disease". As a consequence the system cannot classify a question in the category "Disease" because it is unknown for him. This disproportion is traced by the results of macro and micro measures. **Table 3.** provides us macro and micro measures of the Italian classifier. For the medical type, macro-recall, which gives an equal importance to all categories, is 63.2%. But micro-recall, which gives an equal importance to all questions (so gives more importance to well represented categories), is lower: 43.7%. To cope with this difficulty, the Unified Medical Language System sources are used. Actually it provides synonyms for all medical terms, which is very useful to surround the medical meaning of the question. In addition we should improve the training base with targeted questions which belong to the missing categories. These improvements could be quantified computing new scores of performance of the system.

These scores are closest to the French's ones (68.6% of macro-precision for the type of excepted answer and 51.2% for the medical type) but are less high than the English's one. It reinforces the idea that Latin's languages like Italian and French are complex (because they use accents for instance) and, as a consequence, more difficult to classify by an automatic tool. In addition, linguistically difficulties of languages like French and Italian are well known.

3. Conclusion and perspectives

The Health On the Net Foundation chooses a supervised approach to realise a QA system applied to health whereas pattern matching are mostly used. To obtain a multilingual system, one set of questions by language has been used. The evaluation of the "QuestionAnalyser" module for Italian confirms the feasibility of the supervised method for our multilingual QA system. The SVM classifiers are impressive for the task of classification according to the type of excepted answer with only 5.8% of error. Adding a language to the QA is tedious work because of the collection of questions and then, their manual classification. But it is necessary to target the answer. The specificity of our system is the quality of answers. They only come from the 6'500 certified websites of the Health On the Net foundation. For the moment there are more than 660 Italian websites in the HON database of which 550 are certified. It shows our wish to provide trustworthy and reliable answers to the users. A prototype of the multilingual QA system in English, French and Italian has to be evaluated in term of the retrieved answer before providing it to the users of the HON web site.

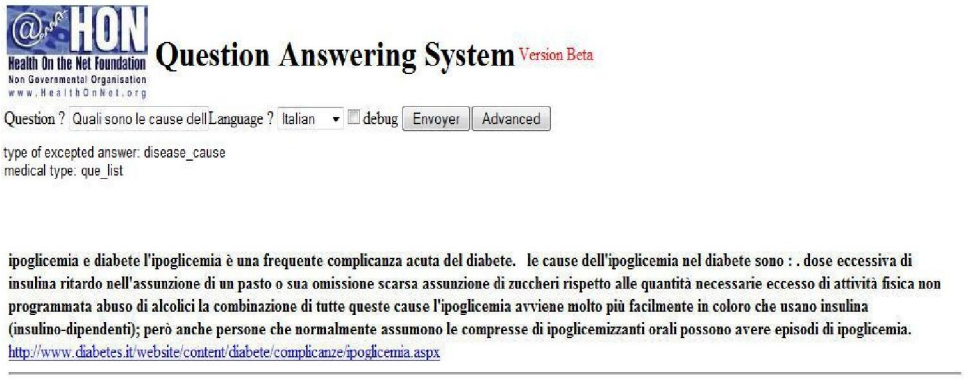

Figure 1. Example of an Italian question: Quali sono le cause dell'ipoglicemia?

Thanks to the SER and the EU who have supported PIPS and more particularly, thanks to the collaborators of the HSR.

References

[1] Hege K ANDRESSEN & AL, European citizens' use of E-health services: A study of seven countries, BMC Public Health 2007.

[2] http://www.harrisinteractive.com/news/newsletters/healthnews/HI_HealthCareNews2008Vol8_Iss8.pdf

[3] M. SELBY & AL, Health On the Net Foundation Code of Conduct for Medical and Health Websites. MedNet 96 - European Congress on the Internet in Medicine, Brighton, U.K., Oct. 14 to 17, 1996.

[4] C. BOYER & AL, Health On the Net foundation: assessing the quality of health web pages all over the world, MedInfo, 2007.

[5] http://www.hon.ch/HONcode/Conduct.html

[6] S.CRUCHET & AL, Supervised approach to recognize question type in a QA system for Health, MIE, 2008.

[7] P. ZWEIGENBAUM, Question Answering in Biomedicine, in DE RIJKE M., WEBBER B., Eds., ACL, p. 1–4, 2003.

[8] M. LEE & AL, Beyong information retrieval-Medical Question Answering, AMIA, 2006.

[9] D. DEMMER-FUSHMAN & AL, Answering Clinical Questions with Knowledge-Based and Statistical Techniques, Computational Linguistics, 2007.

[10] H. YU & AL, Classifying Medical Questions based on an Evidence Taxonomy, AAAI, 2005.

[11] W. WEIMING & AL, Automatic Clinical Question Answering based on UMLS Relations, Proceedings of the 3rd International Conference on Semantics knowledge and Grial, 2007.

[12] Y. HONG & AL, Development, implementation, and a cognitive evaluation of a definitional question answering system for physicians, Journal of Biomedical Informatics, 2007.

[13] P. JACQUEMART & AL, Towards a Medical Question-Answering System: a Feasibility Study, Stud Health Technol Inform. 2003.

[14] E. ALFONSECA & AL, A prototype Question Answering system using syntactic and semantic information for answer retrieval, TREC, 2002.

[15] R. MAY & AL, Building a Question Classifier for a TREC-Style Question Answering System, The Stanford Natural Language Processing Group, Final Projects 2004.