Supervised approach to recognize question types in a QA system for Health

Célia Boyer, Sarah Cruchet and Arnaud Gaudinat
Health On the Net Foundation

MIE 2008
27 May 2008
What is a Question-Answering system?
- A system able to answer to a precise question
- A system able to retrieve answers rather than a list of documents for a natural language query

Why a Question-Answering?
- Alternative to:
  - Looking up answers in resulting documents proposed the search engine
  - Structured data
- Feasibility
  - Maturity of the approach (TREC results)
  - Data redundancy, abundance of structured data

Particularities
- Question-Answering applied to the health domain
- Trusted and controlled areas
- Search only in trustworthy resources
Part of PIPS (Personalized Information Platform for Life & Health services) – a 4 years IST Project

Goal: The overall objective is to develop and pilot a Health & Life Knowledge and Services Support Environment that shall enable:

- Support to HealthCare Professionals in the delivery of the spot/time just-in-time HC services that take into account individual personal preferences and rights.
- Support to Citizens in making informed decisions about therapy and nutrition, and assisting them in determining their health status in a timely manner.

HON Part: *Trust mechanism for e-Health Knowledge base* from *Trust and Security Management*

The QA system is one of HON developments
**Current Initiatives**

### TREC 2007

<table>
<thead>
<tr>
<th>Run Tag</th>
<th>Submitter</th>
<th>Accuracy</th>
<th>NIL Prec</th>
<th>NIL Recall</th>
</tr>
</thead>
<tbody>
<tr>
<td>LymbaPA07</td>
<td>Lymba Corporation</td>
<td>0.706</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>LCCFerret</td>
<td>Language Computer Corporation</td>
<td>0.494</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Isv2007c</td>
<td>Saarland University</td>
<td>0.289</td>
<td></td>
<td>0.000</td>
</tr>
<tr>
<td>UofL</td>
<td>University of Lethbridge</td>
<td>0.258</td>
<td>0.052</td>
<td>0.500</td>
</tr>
<tr>
<td>QASCU1</td>
<td>Concordia University</td>
<td>0.256</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>FDUQAT16A</td>
<td>Fudan University</td>
<td>0.236</td>
<td>0.053</td>
<td>0.312</td>
</tr>
<tr>
<td>pronto07run3</td>
<td>Universita di Roma “La Sapienza”</td>
<td>0.222</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>ILQUA1</td>
<td>State University of New York (SUNY) at Albany</td>
<td>0.222</td>
<td>0.000</td>
<td>0.000</td>
</tr>
<tr>
<td>Ephyra3</td>
<td>Carnegie Mellon University and Universitaet Karlsruhe</td>
<td>0.208</td>
<td>0.048</td>
<td>0.062</td>
</tr>
<tr>
<td>QUANTA</td>
<td>Tsinghua University (State Key Lab)</td>
<td>0.206</td>
<td>0.091</td>
<td>0.062</td>
</tr>
</tbody>
</table>

### QA on the Web
- **BrainBoost:** [http://www.brainboost.com/](http://www.brainboost.com/)

### QA in Health Domain (in scientific literature only)
- QA in biomedecine [Zwiegenbaum 2003]
- Beyond Information retrieval-Medical Question Answering [M. Lee & al 2006]

### Use of machine learning in the task of classification
- Classifying Medical Questions based on an Evidence Taxonomy [H. Yu & al 2005]
What is different?

Overview

- **Multilingual System**
  - Currently English, Italian and French

- **Choice of the field of research**
  - Trusted area
  - The whole web

- **Machine learning to better identify questions**

- **Decomposition in 3 steps**
  - Analysis and treatment of the query
  - Search of reliable documents and passages for the answer
  - Answer treatment
Architecture of the Trusted Health QA system

1. question analysis and treatment

Question Analyzer

Free text, question

Free text, question

Question Type
Medical Type
Answer Type

Question Model

2. documents and passages retriever

Query Generation

Query for retrieval

Contents List

Document Retrieval

Passage Retriever

Passage list

3. answer extractor

Answer Extraction

Answers
**Question Analyzer**

**Goal:** categorize the question

**Significance of the module**
- Upstream the system
- Other modules use its results to carry out their treatment

**How to categorize a question?**
- Medical type (10)
- Type of expected answer (11)

**Methods:**
- Non-supervised methods (clustering) to automatically identify classes (rapid miner, Yale)
- Manual classification by experts of health (corpus of 136 questions in English, 151 in French and 99 in Italian)
- Supervised methods to automatically classify a question
- Use of the UMLS semantic network
Question Analyzer

Example 1: “How to prevent lung cancer?”
- Type of expected answer: [procedure]
- Medical type: [prevention]
- UMLS sources to qualify the question
  [Therapeutic or Preventive Procedure] [prevents] [Disease or Syndrome]

Example 2: “What is diabetes?”
- Type of expected answer: [definition]
- Medical type: [disease]
- UMLS sources to qualify the question [Disease or Syndrome]
Corpus of questions

- Manual research on the Internet (FAQ, health forums…)
- 4 sets of approximately 100 questions
  - One set by language
  - One set for each type of categorisation (medical type and type of expected answer)

Gold standard

- Classification of reference made by experts
  - For learning
  - For evaluation
Non-supervised approach
- Elimination of human expert bias
- Automatic classification with no information about their categories
- Algorithms of clusterisation
  - Untreated form
  - Bigrams or cooccurrences

Supervised approach
- Questions gathered according to expert classification
- Cross validation
- Five algorithms tested
  - SVM, Naive Bayes, Knn, ROCCHIO and Decision Tree

* The two approaches are complementary
UMLS (Unified Medical Language System)

- Developed by the NLM (National Library of Medicine)
- To understand the significance of biomedical and health concepts
- To improve the medical knowledge of the question
  - Use of the Semantic Network

Example: “How to prevent lung cancer”

Semantic relation “prevents”
Semantic type “Neoplastic Process”

Mapping between Semantic types/relations of the UMLS network and medical types

Medical type of the question = prevention
Examples of question classification

- How long can you live with cirrhosis if not treated?
  - Medical type = Routine
  - Type of excepted answer = Duration

- What organ produces insulin?
  - Medical type = Physiology
  - Type of excepted answer = Factoid

- Can we prevent diabetes?
  - Medical type = Prevention
  - Type of excepted answer = Boolean
Question Analyzer

Medical type (10):
- Disease
- Symptoms
- Causes
- Prevention
- Diagnostic
- Evolution
- Physiology
- Diet
- Treatment
- Routine

Type of expected answer (11):
- Boolean (yes/no)
- Causal
- Definition
- Factoid
- Duration
- List
- Temporal
- Person
- Place
- Quantity
- Procedure

27.5% of the English corpus and 39.6% of the French corpus
Results Obtained (3/3)

Question Analyzer

Results comparison...

Better results for the type of expected answer
- How is a metabolic syndrome treated?
- How many bones are in a human body?
- What is rubella?

Better results in English

UMLS enhance results for the medical type
- In English: + 10% of precision
- In French: + 11% of precision

<table>
<thead>
<tr>
<th></th>
<th>maR</th>
<th>maP</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ans type (SVM, W2)</td>
<td>0.81</td>
<td>0.85</td>
<td>0.027</td>
</tr>
<tr>
<td>Med type (SVM, W2)</td>
<td>0.48</td>
<td>0.52</td>
<td>0.078</td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th></th>
<th>maR</th>
<th>maP</th>
<th>Error</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ans type (NB, W3)</td>
<td>0.77</td>
<td>0.67</td>
<td>0.066</td>
</tr>
<tr>
<td>Med type (SVM, W2)</td>
<td>0.48</td>
<td>0.51</td>
<td>0.086</td>
</tr>
</tbody>
</table>

French results

maR=macro Recall, maP=macro Precision; WN=N-grams of words; NB=NaiveBayes, SVM=Support Vector Machine; Ans=Answer, Med=Medical
Performance for the task of categorisation

- Non supervised classification did not emphasize new classes of questions compared to expert

- Supervised method is effective
  - Categorisation according to the type of expected answer better than the medical type one

How to improve the actual system?
- Adding new sample of questions for under represented categories
- Add “ad hoc” rules for French and some for English
Question Answering System

Answers

Type of excepted answer: que_proc
Medical type: disease_prevention

Answer 1:
sunday october 30, 2005 when people think of lung cancer prevention, they automatically think that it has to do with smoking cessation. while this is very true, it is more to lung cancer prevention than quitting smoking. remember, lung cancer affects non-smokers, too! check out ways to prevent lung cancer in "top 5 ways to prevent lung cancer".

Answer 2:
on medicinenet. browse centers allergies alzheimer’s arthritis asthma blood pressure cancer cholesterol chronic pain cold & flu depression diabetes dictionary digestion eyesight health & living healthy kids hearing & ear heart hiv/aids infectious disease lung cond.

Answer 3:
does tuberculosis increase the chance for developing lung cancer.
Question Answering System

Answers

Type of excepted answer: que_fac
Medical type: disease_routine

Answer 1:
perché è importante inserire l'attività motoria nelle abitudini quotidiane: così è più facile che il nuovo stile di vita sia mantenuto nel tempo.

Answer 2:
fino al mercato per comprare verdura fresca: ci guadagna anche il portafoglio. per ulteriori informazioni: * eurobarometer 183.

Answer 3:
in salute, dunque, per stuzzicare la volontà e l'impegno di queste persone che vengono invitata a riprendere e continuare i momenti di attività fisica anche al di fu del progetto ospedaliero. abbiamo già avuto dei riscontri positivi, ? racconta il dott.
Perspective of the QA

- For the answer extraction
  - N-grams on medical terms for the answer extraction
  - Use of the UMLS semantic network to enhance the extraction of answers

- Ajax interface
  - Proposition of pre-requests where answers are known

- Clustering of proposed answers

- Cross languages QA
Thanks to PIPS

THANK YOU FOR YOUR ATTENTION

Any questions