


Question Answering Systems: a partial answer

Luis Fernando Costa
Diana Santos


Linguateca
www.linguateca.pt



Outline

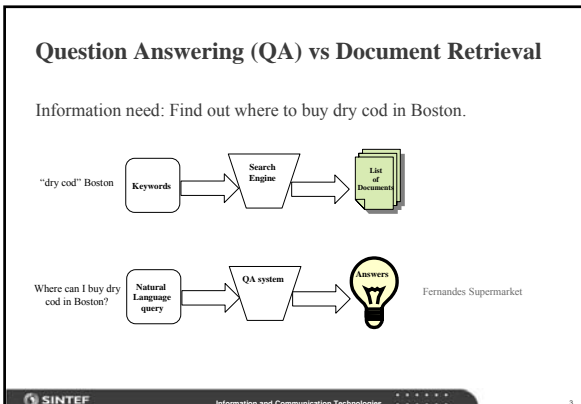
- What is question answering?
 - Question Answering (QA) vs Document Retrieval
 - History
 - Dimensions of QA
- The QA system Esfinge
- Evaluation of QA systems
 - CLEF from a participant point of view
 - CLEF from an organization point of view
- QoLA
- Other interesting systems/perspectives
- Evaluation of QA systems... much more

Linguateca




Question Answering (QA) vs Document Retrieval

Information need: Find out where to buy dry cod in Boston.




Keywords → Search Engine → List of Documents

Natural Language query → QA system → Answers
Fernandes Supermarket




What is a question answering system?

- (Maybury 2004) "Aims to provide natural language responses to natural language queries"
- (Maybury 2004) "Interactive human computer process that encompasses understanding a user information need, typically expressed in a natural language query; retrieving relevant documents, data or knowledge from selected sources; extracting, qualifying and prioritizing available answers from these sources; and presenting and explaining responses in an effective manner."
- (Jijkoun & Rijke 2005) : "One of several recent attempts to provide highly focused access to textual information."
- Wikipedia : "Type of information retrieval. Given a collection of documents (such as the World Wide Web or a local collection) the system should be able to retrieve answers to questions posed in natural language. QA is regarded as requiring more complex natural language processing (NLP) techniques than other types of information retrieval such as document retrieval, and it is sometimes regarded as the next step beyond search engines."
- (Luis 2007): System that gives a **precise** feedback to a **precisely** expressed information need.
- (Diana 2007): System that answers a natural language question, that is, a question in natural language, in a natural way for humans.




Integral parts in a QA context

- A user
- A collection
- In most laboratory experiments
 - a given collection
 - a "general" user
- For real applications, it is crucial
 - to have an adequate user model (and real users)
 - to have an appropriate/adequate collection/subject/topic of interest



History of QA

- Different origins
 - Old AI: Reasoning for answering complex questions (Bruce 72, Lehnert 78)
 - Natural language interfaces to databases: closed domain
 - More user-friendly information retrieval (IR): open domain
- Merging of IE (NLP) and IR in TREC-99 did it! (Voorhees 2001)
- The Web as an unlimited source of answers
 - Auto-FAQ, cyberspace leveraging (Whitehead 95)
 - FAQFinder (Burke et al. 97)
 - data mining (finding implicit knowledge in large data sets) -> text mining (finding implicit knowledge in large text collections)
- QA beyond "factoids": roadmap in 2000 (Burger et al. 2000)



Some sub-areas/dimensions in QA (Maybury 2004)

- Temporal QA
- Spatial QA
- Definitional QA
- Biographical QA
- Opinionoid QA
- Multimedia/multimodal QA
- Multilingual, or crosslingual QA
- Summarizing QA
- Interactive QA, etc

Temporal QA

- Questions with an additional temporal restriction, maybe using implicit publication date, requiring an underlying temporal reasoner
Who was the chancellor of Germany from 1974 to 1982?
Who played against Rosenborg in the 1994 UEFA Cup? (against a newspaper collection in 1994)

Spatial QA

- Questions involving spatial objects, attributes and relations, requiring an underlying spatial reasoner
What is the nearest pharmacy from Holbergs Plass?
What is the largest river in Portugal?
Who is the oldest president of a former Yugoslavia republic?

Definitional QA

- Creation of definitions or descriptions of objects or terms.
What is brunost ?
What is ESA?
Who is Jorge Amado?

Biographical QA

- The answer should provide the most significant characteristics and events in the life span of a person, group or organization, possibly as a sort of coherent summary or even biography.

Opinionoid QA

- Detection of opinions (of individuals, groups or institutions).
What is the position of the political parties in Norway regarding Norway's role in NATO?
Is FrP in favour of homosexual adoption?

Multilingual, or crosslingual QA

- Answering questions using multilingual sources.
Name titles written by Jules Verne.
In which country abroad was the Pope longer in 2003?
Where did Jens Stoltenberg graduate?

Multimedia/multimodal QA

- Processing queries about and extracting answers from sources that may be expressed in a range of media (pictures, interviews, movies, maps, etc.)

Who is Bjørn Skjellaug? (in a picture of SINTEF)
Who is this guy?

Quem é Iqbal Masih?
o rapazinho da foto: R, W, U



Find me pictures where Bush is near Clinton

Esfinge Question Answering system

- Development started in the end of 2003.
- Starting point:
 - Apply the approach described in [Brill 2003] to Portuguese.

Brill, Eric. "Processing Natural Language without Natural Language Processing", in A. Gelbukh (ed.), *Computational Linguistics and Intelligent Text Processing: 4th International Conference, CICLing 2003 Mexico City, Mexico, February 16-22, 2003. Proceedings*, LNCS Volume 2588, Springer-Verlag Heidelberg, 2003, pp. 360-9.

- Using the information redundancy in the Web and less linguistic resources which take a long time to create and are expensive.

Motivation for the development of Esfinge

- The Web is ever growing and Portuguese is one of the most used languages there.
- There are few people working in Portuguese QA, and none of these people applying the information redundancy in the Web to QA.

Esfinge aims to answer

- Questions with a short answer (or which can be answered with several short snippets) (1 to 5 tokens):
 - *Quantos estados tem a Alemanha?* (How many states belong to Germany?)
 - *Quem é Stephen Hawking?*
- Cannot answer questions which require a long answer:
 - *Como preparar lutefisk?*

Main modules in Esfinge

- Question analysis and reformulation
- Retrieval of relevant passages
- Answer extraction and ranking using a named entity recognition system and n-gram harvesting.
- Answer filtering

Question analysis and reformulation

- Questions can give an hint about the answer type.

Quem era o pai de Isabel II? => (person)
Quando viveu Franz Liszt? => (period of time),
Onde é a sede da Interpol? => (place)

- Search patterns derived from the question.

"a sede da Interpol é", score=10
a sede da Interpol é, score = 1

Retrieval of Relevant Passages

- Search patterns are then used to query search engines (Google and Yahoo APIs) and/or document collections.
- The first 50 document snippets returned by each search engine are retrieved (excluding joke sites, blogs, etc identified by their address)
- Sets of 3 sentences are retrieved when a document collection is searched.

Answer extraction and ranking using a named entity (NE) recognition system

Quem é o presidente da França? => Human

"No dia 14 de setembro de 2004 o presidente da França, Jacques Chirac visitou a Finlândia"

No dia <TEMPO TIPO="DATA">14 de setembro de 2004</TEMPO> o <SERES TIPO="CARGO">presidente da França</SERES> , <SER TIPO="HUM">Jacques Chirac</SER> visitou a <LOC TIPO="PAIS">Finlândia</LOC>

NE score = \sum (NE frequency * Passage score * NE length)

Answer extraction and ranking using n-gram harvesting

"No dia 14 de setembro de 2004 o presidente da França, Jacques Chirac visitou a Finlândia. Teve uma breve reunião com o presidente finlandês."

N-gram	Frequency
o presidente	2
o	2
presidente	2
No dia 14	1
No dia	1
dia 14	1

$$\text{N-gram score} = \sum (\text{N-gram frequency} * \text{Passage score} * \text{N-gram length})$$

Answer filtering

- Filter candidate answers:
 - Contained in the question:
 - Ex: *Quem era o pai de Isabel II?* (Isabel II is not a good answer)
 - Based on the part of speech of the words in the answer (answers which first and final token are not adjectives, common nouns, numbers or proper nouns are filtered).
 - Ex: o presidente , o , presidente (OK) , No dia 14, No dia, dia 14 (OK)

Answer presentation (1)

Resposta(s) do Esfinge

Tue Nov 21 09:46:34 CET 2006

Pergunta: *onde é a esfinge?*

Hitos

1768 - 19092003. TVZ Esfinge é o nome de Helsinki, RAJA, 22 mai (AFP) - O TVZ Esfinge comparece a nível mundial de futebol...

Oficina de Arte - Alameda da 1ª rua - Odeon e Regatas - Esfinge é o pai do esportista de origem catalã, e Esfinge é a capital...

Uma coisa: cada vez que visito, entre outras atrações, incluído, também - Helsinki 19092003. TVZ Esfinge é o pai de 19092003 por causa do...

Mão

Liga dos Campeões TVZ Esfinge é a abreviação de Mão na "bola" O TVZ Esfinge é o nome abreviado de AC Mão, de Rua Certa, na Liga dos Campeões...

se refere-se uma série de resultados estatísticos, a que é associado de mãos como Lempit, TVZ Esfinge, Real Madrid, Brasa, Cuba, Mão e Láb...

Esfinge - Helsinki, 19092003. Odeon da 1ª de Janeiro de 2005. Frontal de 1 a TVZ Esfinge, Odeon da 1ª de Janeiro de 2005. Frontal de 1 a Mão...

Respostas: comentários e respostas

Answer presentation (2)

Resposta(s) do Esfinge

Tue Nov 21 09:23:39 CET 2006

Pergunta: *Quem viveu o Mundo?*

News

O livro de José Carlos de Almeida é fundamentalmente sobre uma Bíblia Sagrada. O livro é que, para quem o conhece, trata-se de uma obra que contém...

News

Escrito a história de José Carlos de Almeida é fundamentalmente sobre uma Bíblia Sagrada. O livro é que, para quem o conhece, trata-se de uma obra que contém...

Por Eduardo Lima SAC FAZ O Odeon - O jornal dos Odeões publicou, como a tradição sempre, o mundo como uma mensagem sobre o mundo 16 que não é real...

O jornal Notícias revelou que o Vaticano tem duas estruturas e é isso O mundo como uma mensagem de Deus XVI que não é real - Ele como responsável e...

Respostas: comentários e respostas

Esfinge Results

- Participation at QA@CLEF 2004 e 2005 e 2006

2004 (30/199 right answers, 67/199 in the current version)

2005 (48/200 right answers, 74/200 in the current version)

2006 (50/200 right answers)

- Web interface at <http://www.linguateca.pt/Esfinge/>
- Source code available at <http://linguateca.di.uminho.pt/cgi-bin/viewcv.s.cgi/Esfinge/>
- Available as web service at <http://acdc.linguateca.pt/Esfinge/EsfingeWS.wsdl>

QA@CLEF

- Given questions in a source language, and a target language collection, the goal is to find exact answers in the target language, with a justification snippet, as well as document references that support the answers.
- 200 questions / 1 week to return the results
- Question & Answers example:

Question: Where was Naguib Mahfouz stabbed?
Answer 1: Cairo Document: PUBLICO-19941015-175
Answer 2: pescoco (neck) Document: PUBLICO-19941015-172

Format of the results

Question Number	Run identifier	Confidence Score	Doc	Answer	Text snippet 1	Text snippet...
-----------------	----------------	------------------	-----	--------	----------------	-----------------

- 0139 esfg063ppt 0 PUBLICO-19941015-175 Cairo A prová-lo está o drama do cabo Waxman , o soldado israelita raptado pelos fundamentalistas do Hamas , que anunciavam matá-lo ontem e adiaram o ultimato por 24 horas . A ameaça dos extremistas à paz acabou por ter ontem um outro sinal , por coincidência relacionado com os Nobel . O escritor egípcio Naguib Mahfouz , Nobel da Literatura , foi apunhalado e ferido no Cairo .
- 0139 esfg068ppt 0 PUBLICO-19941015-172 **apunhalado no pescoço** Escritor egípcio apunhalado O ESCRITOR egípcio Naguib Mahfouz , único árabe a ganhar o prémio Nobel da Literatura , em 1988 , foi ontem apunhalado no pescoço quando saía de um automóvel do diário «Al-Ahram » numa rua do bairro residencial de Agouza , na parte ocidental do Cairo , para entrar em casa , junto ao Nilo . Mahfouz , de 83 anos , autor do livro «A Busca » , foi abordado por um homem que saiu de outra viatura e que , de repente , lhe desferiu duas facadas no pescoço , pelo que logo o levaram para o vizinho hospital da polícia . Aí foi operado de emergência e ficou na unidade de cuidados intensivos , se bem que tenha sido considerado livre de perigo .

Tipos de perguntas

from Paulo Rocha's presentation in 2006

- I keep six honest serving men
They taught me all I knew
Their names are What and Where and When
And How and Why and Who*
Rudyard Kipling
- Tipos de perguntas
 - FACTOID (±150; PT: 143)
 - Quem escreveu Miguel Strogoff?
 - DEFINITION (±40; PT: 47)
 - Quem foi Júlio Verne?
 - LIST (±10; PT:9)
 - Diga livros de Júlio Verne.

Question types (Factoid, Definition)

- LOCATION (F,D)
 - Onde é que se afundou o petroleiro «Prestige»?
 - O que é Naurn?
- PERSON (F,D)
 - Quem era o pai de Isabel II?
 - Quem é Hugo Chavez?
- TIME (F)
 - Quando é que a Itália se tornou uma república?
- ORGANIZATION (F,D)
 - Que agência americana foi fundada em 1958?
 - O que é a OMS?
- MEASURE (F)
 - Quanto pesa Ronaldo?
- OTHER (F,D)
 - Em que embateu o Titanic?
 - O que é o samovar?

adapted from Paulo Rocha's presentation in 2006



Temporally restricted questions (I)

adapted from Paulo Rocha's presentation in 2006

- Cerca de 40 perguntas (PT: 23)
- EVENT
 - Quem era presidente do Brasil durante a Segunda Guerra Mundial?
- DATE
 - Quem era presidente do Brasil em 1944?
- PERIOD
 - Quem foi presidente do Brasil de 1951 a 1954?
- Antes, durante, após

NIL questions

adapted from Paulo Rocha's presentation in 2006

- No answer at all (false pressupositions)
 - Quem é o rei da República Checa?
- No answer in the collection
 - Quem é o sucessor de João Paulo II?
 - Quem é o presidente da Câmara de Plovdiv?

Definition questions

adapted from Paulo Rocha's presentation in 2006

- O que é "Margarita e o Mestre"?
- O que é o comunismo gulash?
- O que é Nauru?
- O que é o cirílico?
- O que é Roque Santeiro?
- O que é o efeito de estufa?
- O que é a Mona Lisa?
- O que é o fauvismo?
- O que é a cachupa?
- O que é o xelim?
- O que é a Marselhesa?
- O que é um mujique?
- O que é kabuki?
- O que é um ornitorrinco?
- O que é um sátrapa?

Linguatca organizing QA@CLEF

- Part of the organizing committee since 2004
- Full texts (1994, 1995) from the Portuguese newspaper *Público* and from the Brazilian newspaper *Folha de S. Paulo* copyright cleared and packed into the CHAVE collection, publically available (37 groups have ordered it).
- Created questions and answers in Portuguese.
 - 100 original, 600 translated from the other organizers
 - with a list of documents supporting the answers
 - and their translation into English
- Evaluated the systems 3(2), 5(3), 11(5) providing answers in Portuguese
- Suggested the QoLA pilot for 2007

QoLA (Collaborative QA)

- Pilot: a different setup that should not replace the main track, but whose (positive or negative) results influence future main tracks
- Goals:
 - Investigate strategies for the **combination** of different answers to a particular question (taking into account support justifications, confidence scores, resources used, accuracy values, etc.)
 - Investigate the use of **multiple collections** (and multiple languages) to answer questions
 - Foster **collaboration** between different approaches and different groups

QoLA participants

- **Providers** (without whom QoLA will not exist): participants who agree in deploying/converting their QA systems as a Web service according to the QoLA guidelines and make it available to QoLA.
- **Consumers**, who will only be concerned with strategies to employ -- to the best of their abilities -- all (or some) of the QA services available.

QA web service definition

- Using WSDL (Web service definition language)
- Main operations:
 - GetSupportedLanguagePairs
 - GetSupportedEncodings
 - GetQuotaInfo
 - GetResourcesUsed
 - GetAccuracy
 - GetAnswer

QoLA plan/schedule

- Create at least one MT provider
- Create fake QA systems that simply provide last year's answers, for training purposes
- Provide a robust architecture for running QoLA runs *until March*
- Evaluate a lot more runs
- Semi-automatically create more questions
- Semi-automatically evaluate a lot of answers *until June*
- Plus the usual CLEF stuff (improving Esfinge and coming up with QA pairs)

More details about QoLA: Task

- Answer a set of 200 questions in a particular language
- Use a strategy/program for invoking and deciding among different answers
- Do not call systems by name (the strategy has to be fully automatic)
- Simple MT services can be invoked to test crosslinguality
- Organization is supposed to create several baselines, both as an illustration and for evaluation purposes
 - random invocation, random choice, concatenation, majority voting
- We expect participants to automatically create questions from answers, yes/no questions, and slightly related questions as well

Evaluation contests for QA systems

- Main evaluation fora (IR)
 - TREC (mainly for English) (started 1992, **QA since 1999**)
 - CLEF (European languages and multilinguality) (from CLIR track of TREC in 1998, started 1999, doing **QA since 2003**) (Linguatca since 2004)
 - NTCIR (Asian languages and multilinguality) (started 1999, **QA since 2002**)
- Best accuracy at TREC >70%, at CLEF 2006 (69% in monolingual QA and 49% in bilingual QA)
- AQUAINT program (ARDA Question Answering for INTelligence) 2002-2006 (2 phases): dialogue QA, multimedia and crossmedia QA
- Interactive QA
 - since TREC-6 (documents answering a question), TREC-9 (any N X's; comparison; largest; first or last); CLEF iQA

Evaluation contests, contd.

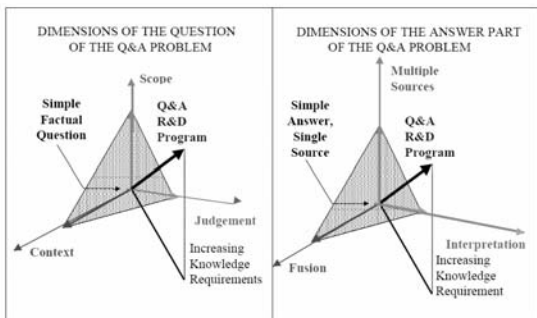
- TREC 2001: in addition to factoids, complex questions, list questions and context questions
- TREC 2003: list and definition questions
- TREC 2004: series around a person, organization, thing, plus "other" question
- TREC 2005: plus events; document ranking (for pooling); and asking for evidence for a particular relationship
- TREC 2006: =TREC 2005 - document ranking; questions in the present tense = now, not the document date

Evaluation contests, contd.

- NTCIR-3 (2002): 3 different tasks for QAC-1 for Japanese news texts
 - Task 1: System extracts five answers from the documents in some order. 100 questions. System is required to return support information for each answer of the questions. We assume the support information as a paragraph, 100 letter passage or document which includes the answer.
 - Task 2: System extracts only one answer from the documents. 100 questions. Support information is required.
 - Task 3: Evaluation of a series of questions. The related questions are given for the 30 of questions of Task 2.
- NCTIR5 CLQA 2005: cross-lingual question answering
 - questions only about named entities
 - CH-EN, JP-EN, EN-JP, EN-CH, CH-CH

A roadmap for QA (Burger et al. 2000)

- A broad spectrum of **questioners**
 - casual questioner
 - template questioner
 - cub reporter
 - professional information analyst
- **questions**: from about simple facts to complex, using judgement terms, knowledge of user context, broad scope
- **answers**: from simple answers in a single document to multiple sources, fusion, conflict resolution, multiple alternatives, adding interpretation, drawing conclusions
- **usefulness**: timeliness, accuracy, usability, completeness, relevance



(from Burger et al, 2000:4)

Use of structured knowledge sources

- Example: Omnibase system used in START QA system (<http://start.csail.mit.edu/>)

A Clockwork Orange (1971)



Name a film by Stanley Kubrick.

Which film of Stanley Kubrick has Malcolm McDowell acted in?

Who is the director of the film "A Clockwork Orange"?

From which year is the

Lexical ontology use (Harabagiu et al 2000)

- *What do penguins eat?*
- from processing previous sets of answers that are classified in a answer type hierarchy, the most widely noun concept associated with the verbs {eating, feeding} is *food*
- use all sorts of concepts under *food* in WordNet to semantically restrict the answer
- also, use keyword alternations to add terms to queries
 - morphological: invented, invent, inventor
 - lexical: killer, assassin
 - semantic: like to, prefer

Web: almost every name is ambiguous or vague

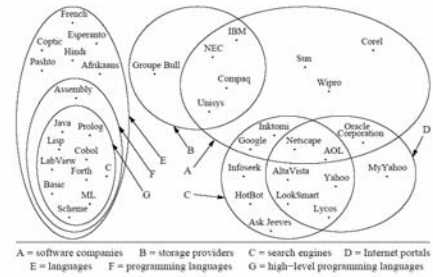


Figure 3: Instance set overlap indicates related categories

Pasca 2004

Metonymy coercion (Harabagiu 2006)

What kind of assistance has North Korea received from USSR for its missile program?

- **named entity metonymy** names of a certain class are used to reference an entity of another name class –country name (place), met/lit=met, government (organization)
- **predicate metonymy**: logical (*John enjoys the book* -> *John enjoys reading the book*) and troponymic metonymy (*receive assistance* -> *modernize*; *stride* -> *walk with long steps*): add a new predicate vs insert a new predicate

Reformulate a question into a possible answer

- Transform the question into an answer, using machine learning over 30,000 question-answering pairs (Agichtein et al. 2004)
- Evaluation
 - 290,000 real user questions *where, what, how and who* from 2.5 mill query logs
 - task: assessors (two panels) should assign good, bad or ignore to 424 questions (113 were evaluated) to the 50 first documents (randomized) with and without reformulation by Google, Altavista and AskJeeves
 - metrics: precision, helpfulness (the % of questions where a given system provides the best performance of all systems tested) and MRR

Find the best answer to a question

- Berger et al. (2000)'s problem: Given a set of 4,700 QA pairs, what is the best answer to a question
- randomly selected 10% of the documents for test
- evaluate: how close a correct answer is to the top of the returned list
- measures: median rank and MRR

What kind of answer to *why* questions?

- Verberne et al (2006) manually classified 122 questions according to whether they ask for cause, motivation, circumstance or generic purpose and developed a machine learning classifier to find that
- asked 10 subjects to read texts from Reuters and the Guardian and create questions and provide answers as well as answer the questions posed by another participant: 395 questions, 769 answers
- use the RST-annotated treebank to elicit 327 QA pairs
- strategy to answer *why P* questions: look for P in the text (83%) as the nucleus (62%) of one of 12 RST relations, and select the satellite as the answer, comparing to people's answers (59%)
- baseline (*because, since* function words or the sentence after): 24%

Find a QA pair for a question

- FAQFinder (Burke et al. 1997) uses a set of QA pairs from FAQ's
- Evaluation (rationale: if a user's question is answered, it is irrelevant that there was also another QA pair that answered it)
 - **Rejection**, the % of questions that a system correctly reports as being unanswered in the file.
 - **Recall**, the % of questions for which FAQ finder returns a correct answer when one exists
- If the answer is not in the collection, send it to an expert board...

QA reuse (Maybury 2006)

- Reuse of questions (question-oriented reuse)
 - same (reformulated),
 - subquestion,
 - embedded,
 - fact-contained, clarified (in follow-up)
- Reuse of answers (answer-oriented reuse)
 - inferring from multiple answers
 - learning the limits of an answers
- Reuse of QA pairs
 - topic characterization
 - answer revision

Other QA systems / approaches

- On the shallow/deep divide
 - Use of structured knowledge sources (ontologies)
 - Use of redundancy and high-precision patterns for bootstrapping
- On the NLP effort to understand the question
 - Use of deep parsing (Bouma) or rhetorical parsing (Verbene)
 - Use of statistical methods, usually machine learning techniques (Brill)
 - Use summarization techniques (Nunes), NER techniques, MT techniques
- Where to start from
 - Use of previous repositories of Q and A (de Rijke)
 - cached questions and their reformulation
 - Use of "intelligent" passage retrieval
 - Use of heavily annotated collections

Other QA systems / approaches (2)

- Use of people
 - social tagging
 - query expansion based on logs
- Use of reasoning
 - derive answers from several pieces of information
 - discard answers or find conflicts
- Answer justification
- Auto-evaluation/improvement (learning)
 - provide a confidence measure based on many factors
 - interact/reformulate the question
 - automatically create its own rules based on user feedback
 - user profiling