Finding answers to questions, in text collections or web, in open domain or specialty domains

Brigitte Grau

Finding answers to questions, in text collections or Web, in open domain or specialty domains

Brigitte Grau
LIMSI-CNRS and ENSIIE, France

ABSTRACT

This chapter is dedicated to factual question answering, i.e. extracting precise and exact answers to question given in natural language from texts. A question in natural language gives more information than a bag of word query (i.e. a query made of a list of words), and provides clues for finding precise answers. We will first focus on the presentation of the underlying problems mainly due to the existence of linguistic variations between questions and their answerable pieces of texts for selecting relevant passages and extracting reliable answers. We will first present how to answer factual question in open domain. We will also present answering questions in specialty domain as it requires dealing with semi-structured knowledge and specialized terminologies, and can lead to different applications, as information management in corporations for example. Searching answers on the Web constitutes another application frame and introduces specificities linked to Web redundancy or collaborative usage. Besides, the Web is also multilingual, and a challenging problem consists in searching answers in target language documents other than the source language of the question. For all these topics, we present main approaches and the remaining problems.

INTRODUCTION

The large number of documents currently on the Web, but also in intranets, makes it necessary to provide users intelligent assistant tools to help them finding the specific information they are searching for. Relevant information at the right time is able to help solving a particular task. Thus, purpose is to be able to access the content of texts, and not only give access to documents. The document is the means to reach the knowledge it contains, not the goal of the research. Question-answering systems address this question and their purpose is to provide a user the information she is looking for instead of documents she will have to read to find the required answer.

This topic arose since the early work in Artificial Intelligence with systems dedicated for questioning knowledge base in natural language, as BASEBALL in 1963 (Green et al., 1986) LUNAR in 1973 and LADDER in 1977 (see (Barr et al., 1981) for a brief description of these systems). Afterward, Lehnert with her system QUALM (Lehnert, 1977) has posed the problem of the semantic modeling of questions in order to associate them different strategies to find answers. However, these works were based largely on manual modeling of knowledge and remained dedicated to limited domains. Thus, they have not led to realistic applications and the research for precise answers turn towards the development of database interrogation interfaces.

It is only recently that the problem has re-emerged at TREC, in 1999, with the first evaluation of question-answering systems in open domain dedicated to find answers to factual questions in texts. As in querying database, factual questions wait for short answers that give a precise information. Factual questions are those questions that ask for a short and concise answer about precise facts, as for example a person name as in "*What is the name of the managing director of Apricot Computer?*" or a date as in

"*When is Bastille Day?.*" However, this time, topics are not limited and knowledge is not structured previously, since these are the texts that are its repositories. Finding answers requires analyzing texts and this is made possible thanks to mature natural language processing tools. The wide availability of texts in numeric format has allowed to model and evaluate linguistic processes and led to the distribution of tools widely applicable, such as word syntactic category taggers (also called part-of-speech (POS) taggers) or robust syntactic parsers. Word syntactic category taggers is the process of identifying which word is used in a text, and which is its grammatical category, as noun, verb, adjective. Syntactic parsers realize grammatical analysis of sentences, highlighting the different phrases (noun phrases, verbal phrases, etc.) and their relations, as subject, direct object, etc. The dissemination of knowledge sources, such as lexicons, thesauri and ontology also enables the realization of advanced text processing.

Thus, the problem of finding answers to questions is now posed differently: it consists in extracting a piece of information from a text. The texts themselves are the sources of knowledge and can be structured and enriched by automatic processes. As first systems have found applications in natural language interface for querying databases by non expert users, QA systems are an answer each time there is a great amount of documents to interrogate for precise information needs, even in professional sectors: business analysis, technologic scouting, journalistic documentation, biography, etc.

Since their beginnings in TREC, question-answering systems have known a great interest from the community, either in Information Retrieval or in Natural Language Processing. Following TREC, the task was introduced in other conferences in IR evaluations: CLEF 2 in 2003, for European languages and multi-lingual approach, NTCIR 3 for Asian languages, in 2003 too. These researches have led to the realization of systems which differ from document retrieval systems.

Their first characteristic is the way to specify the information sought. When a user searches for specific information, the most explicit and easier way for her to give her request is to use her own language, without having to translate it in a query dedicated to a search engine. In fact, whatever the query language used, ranging from lists of words to more structured and constrained queries, all queries are intended to describe the type of document sought: documents that are similar to the query. In such queries, type of the expected information is not explicit, and it is not clear whether are searched all documents that refer to a subject or just a specific information or even a definition. QA systems start from a formulation in natural language and provide just the exact answer, and not documents, as a result.

This is the second characteristic of QA systems, and it is this that makes their specificity: they return as a result a set of answers, not a set of documents that the user has to read to find the information she looks for. When a classical search engine entails the need to read documents to assess their relevance, a QA system will prevent this work to the user. Thus, a QA system provides answers supported by excerpts of documents enabling the user to verify their validity. We will see that this notion of validity of an answer goes beyond the assessment of its relevance.

Depending on the application, search will be made in different resources. Technology scouting will lead to browse the Web to answer questions such as the list of companies whose turnover is down by June 2003, or companies that manufacture products X or Y. The search for technical information, such as "how to install a printer" or "what is the command to copy a file" should be made in manuals or on the Web, or will be addressed more specifically by research in FAQ. Knowing the winner of the Nobel Prize in 1965 is possible by consulting newspaper articles or the Web. With the semantic Web, it may also turn to interrogate factual or encyclopedic knowledge base, structured or semi-structured to obtain information. We will see that different media induce different retrieval processes.

In this chapter we will first present question-answering systems whose purpose is to extract answers from documents in a fixed collection, in response to open domain questions. They will be our reference systems. We will then present QA in specialty domains, focusing on the specific approaches they required and the need for using dedicated knowledge bases. We will see after how to search the Web and what particularities it induces by examining different points: i) the Web as a source of knowledge, with its characteristics in terms of size and kind of knowledge it holds; ii) the multilingual Web as the diversity of

languages makes it necessary to develop interlingual or crosslingual systems where the question is in a language and the answer is found in documents in another language; iii) finally, the collaborative Web, where the Web is the vector for providing collective answers to questions and entails new search processes to exploit these resources automatically.

QA IN OPEN DOMAIN

Question answering in open domain is the most studied domain, and has essentially focused on finding answers to factual questions. Such kinds of answer generally correspond to named entities. Named entities are multiword units that can be recognized in texts, according to surface criteria and gazetteers, and that refer to objects of the world as person, location etc. (Nadeau et al., 2007). However, answers can also be other kinds of entities, as in "*Which alphabet has only four letters, A, C, G and T?*" or in Why or How questions. Even if named-entity questions give a supplementary clue for finding the exact answer, all question types present same characteristics to account for. Thus, before describing question-answering systems, we will show the problems they have to address.

Relations between a question and an answer

Searching for specific answers in texts poses two major problems: finding the passage of text containing the answer and the extraction of the exact answer from this passage. Passages are the units preferred by users over documents for supporting an exact answer provided by a system. Thus, a relevant passage may be defined as a piece of text, usually one to three sentences, which contains the information given in the question and the expected answer. Very often, this information is not provided in the exact terms of the question, and there is a gap between the question wording and the text excerpt wording. So a passage will be considered as relevant if it paraphrases the question put into a declarative form and contains the answer. Often, relevant passages are not strict paraphrases of the question they answer: they may contain such a paraphrase plus other information, or they may entail the answer. Thus, our definition of paraphrase covers this larger phenomenon.

Depending on the question and texts phrasing, these paraphrases are more or less distant from the original question: either a passage contains exactly the terms and the syntactic form of the question, but it is pretty rare, or, and that is what question-answering systems have to face, there are linguistic variations in term of different wording of semantically equivalent contents.

At term level, variations involve use of:

- synonyms or other semantic relations between terms such as hypernyms or hyponyms[1] to designate entities;
- morphological variations, such as the transformation from verb to name as "*to meet and the meeting*" or vice versa;
- combinations of these variations that lead to deal with paraphrases of terms.

In the example Figure 1, matching question and passage requires tying "to take final decision" with "to have last words" and "authorize" with "permit", and we can see that a Who question does not always lead to search for a person name, but a person category.

[1] Hypernym is a term that refers to a more general concept, as fruit for apple, and hyponym is the opposite

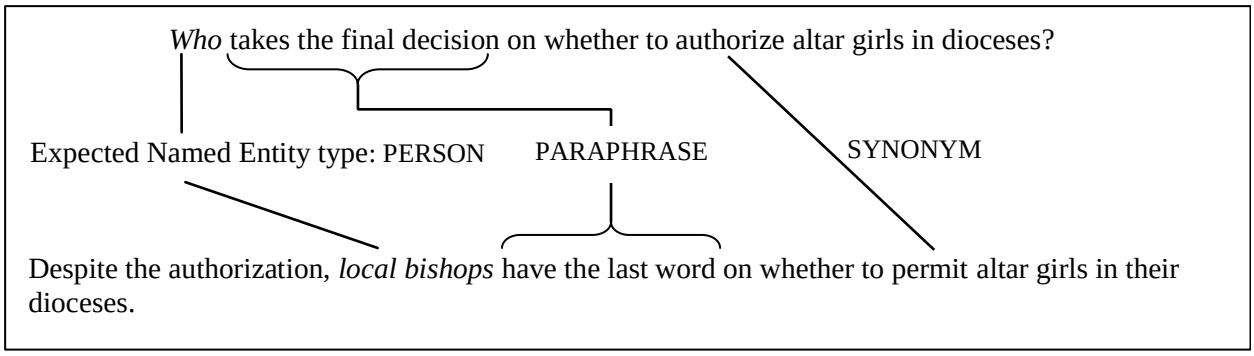

Figure 1: Lexical variations between question and passage

At sentence level, systems have to cope with anaphora and paraphrases, either paraphrases of subpart of question or of the whole question. Anaphora occurs when an entity of the previous discourse is referred by a personal pronoun or another name in a sentence, as Bill Clinton …, the president …, he ...

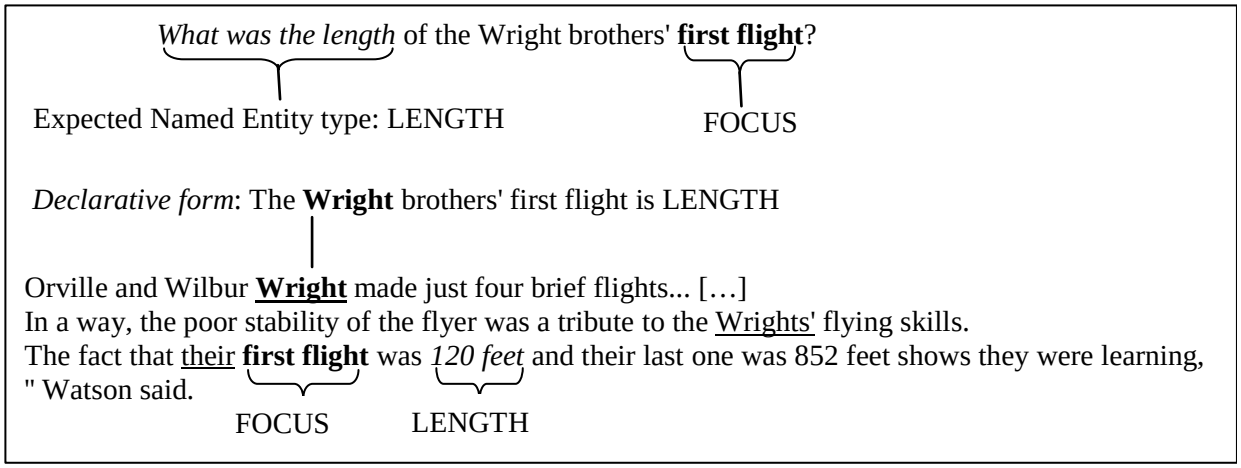

Figure 2: Syntactic phenomena between question and passage

In example Figure 2, the passage contains almost all the question words, but they are not in the same sentence: there exists an anaphoric chain that begins with ”Orville and Wilbur Wright”, then continues with “the Wrights” and ends at “their” in the sentence that contains the answer. Note that the brotherhood between *Orville* and *Wilbur* is not explicit, and should be verified in another document, or in an encyclopedia to be sure. In order to select the right answer, *120 feet*, and not *852 feet*, which are both lengths, some syntactic dependencies have to be checked: the subject of verb *be* is the focus *first flight*, in order to be a paraphrase of the information provided by the question.

Answering a question involves processes related to information retrieval (IR), information extraction (IE) and natural language processing (NLP) fields: NLP to analyze question, IR to search for documents or passages likely to contain the answer and IE and NLP to analyze them and extract the answer. It requires the implementation of various processes, modeling varying levels of understanding.

Components of a question answering system

Question-answering systems generally comprise three steps:

- Question analysis, that determines the characteristics of the answer;
- The selection and the analysis of passages, taking into account the elements identified in the analysis of questions;
- The extraction of the answer from the selected passages.

We will present the general principles implemented by the various systems for these three modules in open domain QA systems.

Question analysis

Analysis of questions makes explicit the information sought by the user as it can then be exploited by the following modules. An important feature deals with the expected type of the answer that systems are able to recognize in texts, outside the question context. These types are associated to classical named entities such as person and organization names, places, dates, quantities, etc. but also to types specific to the QA field, as the definition of different dates (birth, death), subcategories of organizations or persons as political parties, newspapers, universities, actors, politicians, etc. and new types that regularly arise in questions, as symbols of countries, titles of films or books. The number of types varies greatly from one system to another and can range from tens to hundreds.

Thus, Prager et al (2000) have identified 50 types of answers, the system Webclopedia (Hovy et al., 2001) 122, called qtargets, recognized by a set of rules or patterns, consisting of named entity types or semantic categories present into a knowledge base. A broad classification is further developed in (Harabagiu et al., 2000) based primarily on WordNet (Fellbaum, 1998).

Ittycheriah et al (Ittycheriah et al., 2001) make use of a statistical approach to type answers (31 types divided into 5 classes), but their performance remains limited (56% of items labeled, or 280 out of 500), and handwriting rule remains the best solution. Since, more corpus have been developed, and some work propose question type recognition by machine learning methods (for example (Day et al., 2007) for Chinese and English).

Another concern of question analysis is the representation of the information given in the question. Two main trends exist then. The first class of approaches produces a comprehensive analysis of sentence, both syntactic and semantic ((Moldovan et al., 2002), (Hartrumpf, 2005), (Bouma et al., 2006)) to find similar sentences in texts that are analyzed in the same way. The method coverage relies on the capabilities of syntactic parser to produce in depth analysis, and on the existence of semantic knowledge base for achieving semantic analysis.

The second kinds of approaches perform a surface analysis, and highlights certain features, like significant words, their POS[2] tags, the term pivot about which information is sought, called the focus in ((Ferret et al., 2002) , (Laurent et al., 2005), (Plamondon et al., 2003)), relations between the terms of the question. In some system, the term focus corresponds to the designation of the answer type (for example *president* in *Which president*), or the focus corresponds to one or several terms of the questions, ((Soubbotin et al., 2001), (Ittycheriah et al., 2001) (Hovy et al., 2001)). In particular, Hovy et al. identify the relevant question terms and expands them using WordNet, and Soubbotin and Soubbotin recognize primary words (the words which are indispensable for sentence comprehension).

Questions can also be categorized according to the type of information searched, such as a definition, a characteristic of the focus, a role in an event to determine how this information could be expressed and extracted ((Ferret et al., 2002), (Moldovan et al., 2002), (Grau et al., 2006)).

---

[2] POS: part-of-speech, the morphosyntactic category of a token

Table 1 and 2 exhibit some examples of question analysis.

Question	What is the chemical formula for sulphur dioxide?
Answer type	chemical formula
Focus	sulphur dioxide
Terms	chemical formula, sulphur dioxide, plus the single terms
Category	instance

Table 1: Example 1 of question analysis

The answer will be a kind of formula, associated to the focus. We will see in the answer extraction paragraph how these characteristics guide the extraction process.

Question	What female leader succeeded Ferdinand Marcos as president of the Philippines?
Expected named entity type	PERSON
Answer type	female leader
Focus	to succeed
Terms	female leader, Ferdinand Marcos, president of the Philippines, plus the single terms
Category	event + role subject

Table 2 Example 2 of question analysis

The answer will be a named entity that should correspond to the subject of the verb that designates an event.

Most QA systems develop shallow analysis that may involve use of syntactic parsers, and question analysis is usually performed by hand-made rules based on surface criteria (word order, type of words, standard expression, etc.). Some words, either nouns or verbs, play a triggering role to detect the expected named entity type and they are classified relative to this type. Thus, rules for determining this type of answer are based on the interrogative word, the class of the word it is linked to and the class of the main verb. The focus is often the subject of the main verb, except when the latter corresponds to the expected type, it is then the object. The category of the question can be determined by syntactic criteria on the form of the question.

Document and passage analysis

Most systems first retrieved documents with the help of a search engine, then extract relevant passages from them using a dedicated process. Queries are made of the significant question words, eventually expanded by synonyms. A first choice concerns the kind of search engine to rely on. Tellex et al. (2003) have conducted a series of experiments with a Boolean search engine on one hand (Lucene) and a vector model engine (Prise). A Boolean query is made of words related by AND, OR and NOT operators as in "president AND (USA OR American)" and relevant documents have to verify this query, e.g. contain the two words *president* and *USA* or the two words *president* and *American*. Nowadays Boolean search engines also provide an approximate verification when the query is not fully verified. A vector model search engine evaluates a similarity between a query made of a set of words and documents, represented by set of words also. They conclude that both engines produce similar results.
The methods applied for selecting passage from documents can vary widely from one system to another. Many systems develop a weighting scheme to select passages from the retrieved documents, whose size varies from one to three sentences. Prager et al. (2000), and Clarke et al. (2001) based their QA system on passage retrieval techniques, rather than on classical IR techniques and they directly select passages from

the whole corpus. However whatever the process is, the main criteria considered remain the same to score passages, only their combination differs. Thus, some systems annotate the whole collection in order to perform a fine grained collection indexing and search ((Laurent et al., 2005), (Rosset et al., 2005)), while others search for passages, then annotate and weight them. Weighting schemes are based on the following criteria ((Ferret et al., 2001), (Magnini et al., 2002), (Ittycheriah et al., 2001)):

- The number of significant words of the question, usually weighted either according to their degree of specificity in natural language or to their expected role in answer extraction (for example the focus);
- Variations of these words, in order to try to cover all formulations of the underlying concepts that can be found in answering passage;
- Expected Named Entities;
- The proximity of the question terms identified in the passage;
- Eventually syntactic relations between phrases.

All the systems annotate passages by named entity recognizers. In order to detect linguistic variations in passages, QALC ((Grau et al., 2006) (Chalendar de et al., 2002)) analyses them with Fastr (Jacquemin, 2001), a transformational shallow parser for the recognition of term occurrences and variants. Terms, which correspond to multiword units, as "president of the USA", are transformed into grammar rules and the single words building these terms are extracted and linked to their morphological and semantic families, in order to recognize for example "American president". This term recognition shows two advantages: i) documents that contain multiword units in place of single terms are often more relevant; ii) linguistic variations computed on multiword units are more reliable as these terms are less ambiguous than single words. If we consider the example given table 1, the WordNet synonyms of *formula* are: *expression, recipe, convention, normal, pattern, rule*. *Chemical* has only one synonym, *chemic*, and the meaning involved in this question "*chemical formula*" has no synonym found in the corpus by Fastr. By the way, all synonyms of single terms can be discarded as inappropriate. Disambiguation of words is a hard task, and QA systems rarely implement such a process. Thus, synonyms involved in multiword units will be in some manner disambiguated by each term of the unit, and will lead to less noise.

Studies about passage length (Gillard et al., 2005) recommend selecting passages of three to five sentences.

Recent works developed passage reranking techniques, and are mostly evaluated on collections of pairs (Question/Answering passages) and not fully integrated in QA systems. They are based on learning methods in order to take into account lexical and syntactic similarities between passages or questions/passages, or to classify passages (Moschitti et al., 2007) (see section about Collaborative QA).

Extraction of answers

The selection of passages is a first evaluation of the reliability of candidates by applying global criteria. For extracting the answer, more local criteria, related to its formulation, are necessary. The implementation of these criteria can be based on a parsing, syntactic or semantic, of the passage sentences. Within numerical approaches, the system relies on a measure of proximity of recognized terms with the candidate answer, selected according to its type. (Gillard et al., 2005) defines a standardized mean score of compacity of the realizations of the question words in the right and left neighborhoods of candidate answer. Other researchers have developed machine learning approach in place of a weighting scheme.

Systems that develop deep analysis of sentences rely on one sentence passages and have to define a distance between the syntactic representations of the question and each candidate sentence. Bouma et al. (2006) define similarity as the proportion of syntactic dependency relations of the question that match dependency relations of the candidate sentence. The answer is then extracted based on the knowledge of the type expected and additional criteria such as the frequency of the short answer. The determination of

the answer may also result from a logical proof of candidate answers ((Moldovan et al., 2003), (Hartrumpf et al., 2006)): sentences and questions are represented by logical formulas, and the proof relies on deduction rules that model world knowledge. However, such a prover must implement relaxing process when computing the proof to avoid silence.
The most common approach consists in applying extraction patterns to select the correct answer. In Soubbotin et al. (2001; 2002), these are regular expressions describing all types of expected answer. In (Ligozat et al., 2006a), patterns correspond to local syntax rules in the formalism of SCOL (Abney, 1996), written on POS tags (see Figure 3).

Level 1: Phrases
SP = "comma|parenthesis|dash";
NPFoc → DT ? RB ? (ADJ (CC ADJ) ?) ? (FC|FCS) RB ? ;
NPTG → DT ? RB ? (ADJ (CC ADJ) ?) ? (TG|TGS) RB ? ;
NPH → (DT? RB* ADJ* (NN|NNS)+ RB* ADJ* | DT ? RB* ADJ* (NP|NPS)+;

Level 2: Patterns
 # The answer is characterized by its type in an apposition phrase or by a modifier inside its phrase
RTsep → b= NPH SEP NPTG ;
RInTP → NPTG c=NPH;

 # Precision of the answer type
RDefTG → NPTG (IN NPFoc)? VB a= NPH;

 # The answer defines the focus (by using verb be or by an apposition)
RDefFoc → NPFoc VB a= NPH;
RAppFoc → NPFoc SEP a= NP SEP;

Legend:
DT: determinant, RB: adverb, ADJ: adjective, NP, NPS: proper noun(s), IN: preposition
FC, FCS: focus or focus variant, TG, TGS: answer type or a variant

Figure 3. Extraction patterns written in SCOL formalism, dedicated to instance or definition questions

These rules are articulated around the focus tagged FC or FCS or the expected type, tagged TG or TGS and associated to the category of the question. They are written by the definition of two levels: the first identifies different basic noun phrases in sentences NPH, the noun phrases that contain the focus (NPFoc) and the expected type (NPTG). The second level corresponds to the patterns themselves and is based on previously identified groups.
Labels are used to sort the patterns according to their reliability. Thus the answers recognized by pattern *a* are more reliable than those recognized by pattern *b*.
Returning to example Table 1, the following sentences can be retrieved:

S1: Sulfur dioxide (also sulphur dioxide) is the chemical compound with the formula SO2
S2: The structural formula of sulphur dioxide is SO2, and …
S3: The chemical formula for sulphur dioxide is SO2
S4: For example, sulfur dioxide (SO2) and nitric acid (HNO3) may ...

By applying pattern RDefTG on S2 or S3, the answer *SO2* is extracted, as VB stands for the verb *be*, while pattern RInTP allows to extract the answer in S1 and RAppFoc applies on S4.

After the extraction step, some systems apply a validation step, if the extraction approach itself does not entail this validation. Systems generally try to validate the answer by a confirmation coming from another source of knowledge.

Answer validation

A first approach consists to confirm the answer based on the size and redundancy of the Web. Magnini et al. (Magnini et al., 2002b; 2002c) have tested two approaches. The first is purely statistical and is based on the number of documents returned. The Web is queried by Altavista with a query made of keywords of the question and the answer to validate, linked by Boolean and proximity operators, AND, OR or NEAR. They do not search an exact match of the question in the documents found on the Web. The validity of an answer is calculated from the number of documents returned for three queries: one is made from the only question words, the other from the answer words and the third from the previous two. The second method tested is based on the content and relevance of answers relative to questions and is evaluated by a measure based on co-occurrence of words in the snippets returned by Google. These two methods are similar in term of gains and were incorporated into their system evaluated at TREC11 that tries to validate 40 answers per question (Magnini et al., 2002). The final weighting of answers is based on the coefficient of validity from the Web search and the reliability of the answer type. The best result is obtained with the second method and enabled them to find 38.4% of correct answers. This type of approach has been extended in (Awadallah et al., 2006), by adding more measures and applying it to the Arabic language. The test corpus consists of questions from TREC 11 and questions from the game "The Millionaire" that exists in English and Arabic. Results on the two languages are better with strategies based on co-occurrences in the extracts, although below the results of (Magnini et al., 2002c), and results on Arabic are low, probably due to two main factors according to authors: the greater ambiguity of the Arabic words and fewer documents found on the Web, for which the search engines have no linguistic approximation techniques. This fact shows that the applicability of methods often depends on the analyzed language and the resources available for it.
In QALC (Chalendar de et al., 2003), a similar search is performed on the collection and the Web, and only the query formulation changes (see section QA and the Web). Then, the results of the two systems are merged, to promote same answers found in the two sources of knowledge. This strategy allows QALC to validate 106 of the 165 correct answers to 500 questions from TREC 11.

Another form of validation consists in trying to validate missing information in external source of knowledge. Indeed, when an answer is extracted from a passage, its type is not always identified. This is the case of answers whose expected type is given in the question but do not fit exactly a general named entity type but a more specific one, as with the type *female leader* and named entity type PERSON. The verification of the answer can be driven by checking its type into a knowledge base (Bouma et al., 2006), or by exploiting external textual resources ((Grappy et al., 2010), (Schlobach et al., 2007)), as Wikipedia and the Web to compute different criteria giving some evidence about the validity of the answer type, and combine them by a machine learning approach. Such a case occurs in the following answering passage of example table 2:

> *In 1986,* ***President Ferdinand_E._Marcos*** *fled the* ***Philippines*** *after 20 years of rule in the wake of a tainted election;* ***{REP Corazon Aquino}*** *assumed the* ***presidency*** *<.>*

It has to be checked that Corazon Aquino is a female leader.

Verifying that an answer extracted from a passage answers a question may also be posed as a problem of "textual entailment" to find if a passage entails a hypothesis made of the question in a declarative form

plus the candidate answer. Evaluations RTE[3] (Recognising Textual Entailments) and AVE[4] (Answer Validation Exercise) at CLEF gave a frame to evaluate this kind of task. The RTE task consists in determining whether a passage implies a hypothesis while the AVE task whether a passage justifies the answer to a question. This last task can be resumed to the first question by considering the couple question plus answer as a hypothesis.
Systems rely mostly on machine learning approaches incorporating various criteria, most often of lexical order: terms of the hypothesis present in the passage, common named entities or similarity measures. To get a better fit when comparing terms, systems make use of external semantic knowledge such as WordNet (Fellbaum, 1998) or VerbOcean[5] (Chklovski et al., 2004). A criterion frequently used is the longest common substring between the question and the passage (Newman et al., 2005), that also may reflect linguistic variations ((Herrera et al., 2006), (Hickl et al., 2006), (Ligozat et al., 2007)). Such a criterion allows systems to take into account both syntactic and lexical similarities in a same measure, with common words and common syntactic roles, considering that if the hypothesis and the passage share an important subpart, there is a strong evidence that their topic is same. However, criteria based on syntactic dependencies can also be explicitly introduced as a criterion, and Moriceau et al. (Moriceau et al., 2008) compute the number of common syntactic relations. In order to develop a comparison of sentence structures, some systems developed syntactic graph matching ((Kouylekov et al., 2006), (Iftene et al., 2009)) or semantic graph matching (Wang et al., 2009b). As many occurring phenomena can be solved by different methods, Wang et al. (2008) combine all of them within a voting approach.
The last kinds of methods rely on logical proofs, which often lack of robustness since they depend on the completeness of the knowledge base. Thus, they are used in conjunction with the above methods ((Tatu et al., 2006), (Clark et al., 2009), (Bensley et al., 2008)). Best systems obtain an accuracy value around 70-75%.
Such paradigm supposes that the justification can be found in few consecutive sentences, and cannot allow studying justifying processes based on information found in different documents and resources.

The ultimate verification of the validity of an answer is made by the user of the system. Given the answer and the justifying passage, she can usually judge the validity of the proposal. But she may have doubts about the confidence she can give to the materials from which is extracted the answer or to the behavior of the system. To this end, Inference Web (McGuinness et al., 2004a; 2004b) is a tool able to trace the reasoning process for finding an answer and to specify from which sources it is extracted. This tool requires that the reasoning can be modeled by documents PML (Proof Markup Language). By a less formal approach, Javelin (Calais et al., 2004) and REVISE (El Ayari et al., 2010) provide an environment for storing intermediate results and source documents in a relational database, to associate them XML elements and then view the processing steps and the results of modules via a Web browser.

Evaluation

While the problematic of question answering exists since the beginning of NLP, the introduction of a dedicated task in TREC in 1999, campaign organized by NIST, has renewed the topic, by focusing on open domain factual questions whose answer can be extracted from documents (Voorhees, 2001). The success known by this task and its growing complexity has shown the vitality of the researches. The synopsis of the evaluation proposes a set of questions the systems have to answer. Human judges evaluate system results, with several judgments for a same answer. A result consists of an answer along with a document that justifies the answer. Thus an answer with a right value, but that is not warranted in the proposed document will not be considered as correct.

---

[3] Recent RTE challenges held at TAC: http://www.nist.gov/tac/about/index.html
[4] http://nlp.uned.es/clef-qa/ave/
[5] http://demo.patrickpantel.com/demos/verbocean/

In the first campaign, TREC8, the organizers selected two hundred questions among a set proposed by the twenty-six participants. Systems had to return five ordered excerpts of 250 characters as answers, extracted from a corpus of 1.9 gigabytes, or 528,000 documents. The documents came from American newspapers, the Los Angeles Times, the Financial Times, FBIS and the Federal Register. At this first attempt, around 50% of answers were found by the best systems. This first campaign makes in evidence the need to use NLP approaches and semantic knowledge. The TREC9 campaign, the following year, proposed two subtasks to the participants, around 25, one still focused on the extraction of passages, the other requiring short answers (50 characters). A set of 700 questions, including 200 rewritings, were built from logs provided by search engines and selected by the organizer according to their scope (general enough) and leading to evaluable answers. The size of the collection has nearly doubled since it contained 980 000 texts of 3 gigabytes.
The TREC10 campaign in 2001 complicated the task since only short answers were allowed, and some questions had no answers in the documents. Many questions focused on definitions that have caused some problems in their assessment. Indeed, answers could be quite disparate, ranging from the proposition of a generic concept to a part of the definition. For example, the question, "What is an atom?" or "Who is Colin Powell?" were answered by very different levels of granularity and different answer completeness. This is why such questions were deleted from TREC11 when they were not precise enough. The number of participants has stabilized around 35.
The difficulty of TREC11 focused on two points: i) to give only the exact answer and not a short passage, ii) to give only one answer per question and iii) to classify the answer according to a degree of confidence. The collection has been replaced by the corpus AQUAINT. Most systems have searched answers on the Web. Although some works developed a fine grained analysis of sentences, the broadest topping approach relies on criteria to approximate such an analysis. Apart from the LCC system (Moldovan et al., 2002) that gets more than four hundred correct answers of five hundred, the other systems, which certainly differ in their modules but all try to marry surface NLP processes, use of semantic knowledge and techniques of information retrieval, got results that could still be significantly improved.
In parallel to the main task, a track addressing answering questions by multiple answers (list questions) existed since 2000. An attempt to held chained questions was abandoned; the aim was to move towards an evaluation of successive couples of questions and answers related to each other as a simulation of dialogue. Best systems obtained an accuracy value around 70-80%.
From TREC12 (2003) to last campaigns in 2007, questions of definition were reintroduced, assuming a same context defined *a priori* for several questions. After this time, the QA track held in the Text Analysis Conference (TAC) with opinion questions in 2008, and then closed.
In Europe, the campaign CLEF created in 2003 a multilingual question answering track, whose evaluation was conducted similarly to TREC. The difference comes from questions and documents in two different languages, in addition to monolingual QA tracks. The NTCIR evaluation followed analogous specifications, but for Asian languages.

QA IN SPECIALTY DOMAIN

QA dedicated to ontologies on the Web

An evolution of the Web is the vision that it would allow to store and access structured semantic knowledge represented by ontologies. This view brings out new forms of interrogation and search and some research in QA explore this field. As we already said, solutions based on a logical representation of questions and documents for answering open domain factual questions have been proposed and evaluated in (Moldovan et al., 2002; 2003), who developed extendedWordNet for representing inference rules, and in (Hartrumpf, 2006), based on its ontology MultiNet interfaced with the German language through

HaGenLex to build and match semantic graphs. Zajac (2001) and Lopez Garcia et al. (Lopez et al., 2006; 2005) explored the formalization of the process of finding answers in a formal ontology, but it is the work of (Atzeni et al., 2004) and (Calais et al., 2006) which defines the problem for the Web and explore the querying of several ontologies. The first work takes place in the multilingual project MOSES which aims at querying a federation of university websites, each in a different language. To this end, it proposes to merge the ontologies in order to relate concepts described in two different languages, based on the structure of the ontologies as well as translation of labels associated to concepts. The second work envisions the problem differently and search answers in several ontology and merge them in order to increase the completeness of the system proposition.
In a more particular context, some initiatives propose to interrogate semi-structured database (often RDF[6] triplets) build from manual or semi-manual entries[7] [8].
However there is few works in this domain, according to its restrictive application field and more systems were developed to interrogate specialty or restrictive domains.

QA on restrictive domains

It is interesting to see that QA on restrictive domains (RDQA) regained an interest in the research community, and is back since the early years of AI and first QA systems, presented in the introduction section. Some examples of such domains are services of telecommunication corporation (Doan-Nguyen et al., 2006), Biomedical domain ((Sang et al., 2005), (Rinaldi et al., 2004), (Demner-Fuschman et al., 2007)), practical domains as weather information (Chung et al., 2004) or geographical domain (Ferrés et al., 2006). Information is obtained from documents, or semi-structured knowledge or databases. In this latter case, databases are built from documents, generally with offline processes.
Some particularities distinguish this field from open domain QA ((Doan-Nguyen et al., 2006), (Minock et al., 2005)):

- Restricted domain collection, and thus scarcity of answers;
- Domain specific terminology;
- Complex questions and different types of answers than factual ones.

As the domain knowledge is better delimited, and thus can be modeled formally into a conceptual representation that supports inferences, deep analysis of text can be applied to transform documents and questions in such a representation and questions are mapped over the knowledge base ((Sang et al., 2005), (Rinaldi et al., 2004), (Frank et al., 2007)). Besides classical problem related to natural language processing, the main problem concerns the recognition of terms of the domain, as they play a pivotal role. Without having terminological resources or ontologies having a broad covering of the domain, systems performances remain low, as shown when applying a non dedicated QA system.
While open domain QA system are designed to answer factual questions, questions in restricted domain lead to different type of answers, and are often formulated in a more complex manner. Thus, developing a RDQA system requires new classifications of questions (see (YU et al., 2005) for the medical domain) and another level of granularity for answers: precise answer but also passages ((Sang et al., 2005), (Doan-Nguyen et al., 2006)).
Thus, even if their general architecture remains the same in each domain, many processes have to be redesigned or adapted to develop a RDQA system ((Minock et al., 2005), (Jacquemart et al., 2003)). However these systems provide new tools in information processing and management in corporations for example. The reader can report to the overview of Molla and Vicedo (2007) for more references.

---

[6] RDF: The Resource Description Framework (RDF) is a family of World Wide Web Consortium (W3C) specifications originally designed as a metadata data model (from Wikipedia).
[7] http://www.wolframalpha.com/, April 2011
[8] http://www.trueknowledge.com/, April 2011

QA AND THE WEB

Searching the Web or searching a much smaller collection introduces differences in query formulation and QA systems make the hypothesis that there must be at least a document on the Web, and even more documents, which provides the answer in a form closed to the question wordings. The redundancy of the Web also gives some clues to select the correct answer without having to implement complicated extraction techniques.

Searching the Web for QA

One of the first systems that allowed questions in natural language on the Web was AskJeeves. This system was looking for answers as documents in its database. One of the first QA systems designed to query the web using an existing search engine is MULDER (Kwok et al., 2001). With the creation of the QA track at TREC, many systems have early used the Web as a resource for answer validation. These systems performed a rewriting of questions to bring the relevant documents in the top, or to extract only the excerpts provided by the engine (i.e. the snippets). All of them exploit the idea that the huge size and the high redundancy of information allows to find relevant documents even with a very specific query.

Queries for getting precise answers

MULDER and QALC (Berthelin et al., 2003) generate queries as specific as possible, while keeping the capability of relaxing constraints. A rewriting of questions aims at building queries close to the wording of the answers, where the verbal phrase is converted in a declarative form, clues dedicated to introduce the answer are added, and groups of words are kept. This rephrasing is realized by hand made rules and their conception relies on the same principles that guide the conception of extraction pattern. An evaluation of the contribution of these reformulations for MULDER can be found in (Kwok et al., 2001). The two systems make use of Google. Hermjacob et al. (2002) generate paraphrases of the question using syntactic and semantic rules. These paraphrases are used to build Boolean queries (three paraphrases per question on average) to search the Web. Brill et al. (2001) implement simpler question reformulations by keeping the question words in their original order and moving the verb in all possible positions. They demand the search engine to make comparisons at the string level.

The last approach presented here concerns reformulation learning, related to each of the search engines used to make the query, Google and AltaVista (Agistein et al., 2001). Lexical clues that might introduce answers according to their expected types are learned. The training corpus is coming from the Web and was created from FAQ (Frequently Asked Question). The authors have restricted their work to definition questions, i.e. "Who is" and "What is" questions, and questions "how" and "where", covering a specialty domain, computer science.

Use of redundancy

Extraction of exact answers can rely on the concept of redundancy and avoid implementing strategies based on the development of extraction patterns. Clarke et al. (2001a) select answers according to their redundancy and study more particularly this factor, focusing on answers corresponding to names of person, by evaluating the impact of the number and size of the selected passages on the results.

To quantify their strategy, Dumais et al. (Dumais et al., 2002) have applied an adaptation of the approach they have developed for the Web on the TREC collection. Their system exploits primarily redundancy of the Web. This enables to rewrite questions simply (see (Brill et al., 2001) in previous paragraph) and implement an extraction technique as simple, since it retains the string the most frequent in the snippets returned by the search engine, after applying some filters based on the types of questions as an alternative to tagging named entities: existence of uppercase, numbers, for example. The system, applied to the Web with TREC9 questions, finds 61% of answers within the first 5 ranks. Applied to the collection

AQUAINT, the system, after some modifications in the extraction of short passages, is 24% right, against 53% for the previous system applied under the same conditions. A similar technique was used on Portuguese (Costa, 2006), whose pages indexed by Google are estimated at 60.5 million. The system performs a simple rewriting of questions, eliminates noise caused by some sites that have been manually selected, select the first 100 snippets returned by Google and extract the answer based on a technique of frequent n-grams. The system found only 30% of correct answers to questions from CLEF 2004 and 2005, but these issues are dedicated to a collection dating from 1994 and maybe the answers are not all on the Web.

Another approach is to use the Web to assist a search in a reference corpus against finding answers exclusively on the Web. Clarke et al. (2001b) select 40 passages among the top 200 documents returned by two Web search engines and 20 passages in their reference corpus, in which the answer is extracted, provided it belongs to the reference corpus. The Web is used here to increase the redundancy factor of candidates. This approach has improved the results of their system from 25 to 30%.

The Web can also be seen as a repository of knowledge from which information can be extracted that will populate knowledge bases. Thus, Fleischman et al. (2003) have built a large corpus of 15GB, consisting of newspaper articles and documents from the Web to extract concept-instance relationships, in order to answer questions like "Who is the Mayor of Boston "and" Who is Jennifer Capriati. ". 2,000,000 of such relationships were obtained after filtering by a classifier to eliminate noise caused by patrons of extractions. Questions were collected on the site www.askjeeves.com, available in 2003, and the evaluation was calculated on 100 questions. The base can improve the performance of QA system of 36% answers.

Crosslingual QA

The richness of the Web is also its multilinguism, and an important challenge concerns the ability to ask a question in its own language and receive an answer extracted from texts written in any language. When looking for a fact relating to a particular event in a country, it is more likely to find the answer in texts written in the language of that country. That is the purpose of crosslingual systems. They should then not only solve the problem of searching for answers in a language different from the question but also, for completeness, consider their translation. This last point is less problematic than machine translation in all its generality, see (Bos et al., 2006) for example for translating answers. Currently, most QA systems that implement crosslingual solutions leave the answer in the target language. These systems, evaluated at CLEF, provide answers in English from different question source languages, French, Italian, etc. or inversely.

Translation of questions

Some systems make use of machine translation to translate the questions and apply a monolingual system thereafter ((PER04), (Jijkoun et al., 2004a), (Neumann et al., 2004; 2005), (Ahn et al., 2004)). (Perret, 2004) and (Jijkoun et al., 2004a) have also applied their monolingual system to the same set of questions to compare results. The first, in its English-Dutch version obtained a decrease of 10.5% on its results: 91 (45.5%) to 70 (35%) correct answers, and the results of the second, in its English-French version, saw its percentage of correct answers decrease of 13.5: 49 (24.5%) to 22 (11%) answers. BiQue (Neumann et al., 2004; 2005) made use of several tools for translating German into English to get a good coverage. Alignment of translated questions provides the translation of the source words that are put together in a "bag-of-words" representation used for expanding the query. This set is completed with synonyms, after disambiguation. The disambiguation module uses EuroWordNet to find correspondences between words in the two languages (English and German) and, for each ambiguous word, it looks at which of its meaning are expressed both in the source question (in German) and its translations (in English). Their system has achieved 25.5% correct answers at CLEF 2005, and from English into German, 23%. Bouma

et al. (2006b) complete the question translation by automatic translations of named entities and bi-word expressions found in Wikipedia as these types of terms are poorly processed by translators that do not contain them in their dictionaries. Their system, which in the monolingual Dutch task obtained 31% of correct answers, gets 20% in bilingual, English-Dutch.
The two major problems in using machine translations for the questions lie in the bad resolution of ambiguity of the question word and in syntactically incorrect translations. If a word relevant to the search of the answer is badly translated, this error cannot usually be compensated by other words of the question, because questions are often quite short, and the mistranslation of a word changes its meaning.

Translation of the question terms

Another solution consist in analyzing the question in the source language, extract all the useful features, i.e. the type of the expected answer, the words and phrases (nominal, verbal and prepositional), focus and question category. This information remains the same regardless of the language, so only words have to be translated. This brings out the only problem of managing multiple meanings of words. This solution has been chosen by many systems. Tanev et al. (2004), considering that the results of machine translation, especially for questions, were not quite encouraging, managed to translate the keywords of the question: after a step of removing irrelevant words, keywords are translated. To eliminate the noise inherent to such a process, they only retain the combinations of translations the most plausible, i.e. those that appear most frequently in two reference corpus (AQUAINT and TIPSTER). This type of approach was already used in (Grefenstette, 1999) in the context of machine translation for validating translations of noun phrases on the Web. They get a score of 45 (22.5%) correct responses in bilingual cons 56 (28%) in monolingual, so with a loss of 6% of correct answers only.
A combination of translators and the validation of translated multi-terms of the question in a corpus can be found in (Sutcliffe et al., 2006) and (Ligozat et al., 2006b). In Ligozat et al., instead of relying only on co-occurrences, English translations of biterms (terms made of two words) and their possible variations are sought, using Fastr (Jacquemin, 2001)(see results in table 3). For example, from the 777 bi-terms extracted from the questions of CLEF 2005, 39.5% are found in a subset of the collection, 54% only as variants of the given form. This means that the translation of the biterm is not found as such, and therefore does probably not fit with a correct translation, but allow to finding the correct expression. The only bi-terms found alike are often proper names, usually names of people. Each bi-term found in corpus entails to validate translations of its single words, thus relative to the context of the questions. The system found 25% of responses in 2006 in the French to English track.

Total number of bi-terms formed from the questions	777	
Number of bi-terms found	307	39.5%
Number of bi-terms found only in their original form	52	17%
Number of bi-terms found only by semantic derivation	150	54%

Table 3. Validation of bi-terms translations by Fastr

Synapse (Laurent et al., 2005b; 2006) also translates words and idioms, and their system found 44% of French answers from questions in English, while their monolingual French system is 69% correct. In the same paradigm, the crosslingual system English-Spanish BRILIW (Ferrandez et al., 2009) uses EuroWordNet and Wikipedia for translating common words and named entities, and obtains better results than by translating the questions.

Translation of documents

The latter technique explored by (Bowden et al., 2006) is quite rare as they translate the documents into the language of the question, in this case French documents translated into English. The answer is then

extracted and "re-translated" by aligning documents. The system is 40% correct in the source language, translation introduces a loss of 50% of responses.

The choice of an appropriate method relies on available resources for translation from a language to another: machine translator, bilingual dictionary or none of them, using then aligned or parallel corpora to find translations. As machine translation does not provide tools able to produce always well written texts, an experiment (Lopez et al., 2006) was made with users who have to search manually for exact answers in a monolingual frame with documents written the source language (Spanish), and in a crosslingual frame with documents translated automatically from the target language (English to Spanish), using a same QA tool. The authors found that the performance of users for searching answers in the monolingual experiment was only 11.4% better, but they performed the task 40% faster on average. Given these facts for human beings, QA systems have a difficult task to achieve consisting in attaining performances of monolingual systems, even if some burden would be admitted by users. However, a more realistic task that would consists in searching different documents in different languages and merging results has not be proposed in QA, while this kind of task was proposed in information retrieval evaluation conferences.

Collaborative QA

User-generated contents become more and more popular on the Web since the last decade, and community-driven question-answering portals gain a large audience. Recent works are dedicated to provide tools for retrieving existing answers to users' questions, as Yahoo answers. In this specific context, focus is turn towards detecting questions similarities between user's question and existing ones, or similarities between the user's question and existing answers.
Similarity is often posed as a paraphrase problem detection ((Agistein et al., 2008), (Wang et al., 2009), (Cui et al., 2005)), on the intrinsic content of the compared extracts, based on lexical and syntactic information. However, links between contents and rating of them can also be considered for selecting better answers (Agistein et al., 2008), provided that content found on such sites are less trustful, and systems have to consider this kind of problems.
Dealing with paraphrases become more crucial in this context given that users’ formulations of a same need vary and that questions and answers do not possess same properties as in open domain answer extraction: answers are given in response to a question, and not extracted from texts to answer questions, thus their distances in term of vocabulary and forms of sentences are greater.

Web documents present specific structures with informative content, as tables, lists, frames, etc. QA systems that want to exploit them have to recognize such structures and to develop specific analysis (Lerman et al., 2004). That is why some projects have emerged to extract information from structured Web documents. Lixto (Baumgartner et al., 2005) allows to writing wrappers dedicated to sites in order to extract information from the tree representation of the pages, wrappers that can adapt to changing sites. More specifically to answer questions, Katz et al. (Katz et al., 2003), (Lin et al., 2003) have developed a hybrid approach to find answers, based either on simple extraction techniques exploiting the redundancy of the Web, or on the interrogation of certain sites, then functioning as dedicated knowledge bases. Answers to questions about the characteristics of countries, elements of biography of famous people, film is sought on some sites listed and for which wrappers have been developed., Their system answers 30 questions about 42 dedicated to be solved by this technique, from the questions of TREC 2002, and 153 from 458 by a conventional extraction technique, so 16% of answers are found by the use of some sites.

CONCLUSION AND PERSPECTIVES

The evolution of processes used in the question answering field is significant of the evolution in the field of NLP. From first systems, operating on high-level conceptual representations to infer information from their knowledge base within a narrow scope, now it came to systems that can answer questions concerning any field. Differences rely on the types of questions addressed and the sources of knowledge. QA systems answer factual questions when the answer exists in a text, even in an altered form in relation to the question and terms used. Thus, current approaches are working on unstructured knowledge bases, i.e. text collection or Web, and all processes are dedicated to structure this knowledge, primarily through use of NLP. Questions and especially candidate passages are analyzed to identify named entities, noun and verbal phrases, syntactic or semantic relations. Systems often apply surface analysis allowing the identification or the approximation of such information, and approaches as closed to those used in information extraction, where the use of extraction patterns, more or less fixed, was widely chosen, leaving aside the generic analysis of sentences and texts.
However, Moldovan et al. (2002; 2003) showed that it was also possible to apply fine grained analysis, while preserving good coverage. Thus, using a version of extended WordNet, a robust parsing and a relaxed logic prover, their system is over 80% of answers. Systems that use approximate methods to treat the same phenomena are closer to 70% of answers. The great lesson we can draw is that implementing elaborated linguistic processes in order to answer factual questions is possible, even in open domain, without damaging the overall performance. When such skills are missing, another way would be to implement different strategies and to apply them dynamically according to their performance. Systems that seek answers in different sources of knowledge (structured databases, the Web, documents from the collection of reference) show important gains. Others have tried using different strategies, by using two systems with different approaches or with different sources of knowledge (Chalendar de et al., 2003) or by combining in depth analysis and surface numeric processes (Jijkoun et al., 2004b) and get better results than a system or a strategy operating alone. The effectiveness of the techniques described is less tied to the type of question than the difficulty of solving the question, which depends on the sources of knowledge, the number of responses that are present and their formulation.
Monolingual QA evaluations show that the rate of correct answers for a given language is strongly related to the existence of resources, and solutions to overcome this limitation could be found in their acquisition from texts. All these approaches have found their utilization in the IBM system, WATSON[9], dedicated to participate to the American game Jeopardy, which consists in finding an answer from information related to it. WATSON is the integration of multiple search strategies and resources in a parallel environment.

A question-answering system does not represent a self content system and should depend on the application in which it operates. According to the application frame, it involves definition of user type, her degree of knowledge and expertise, the usage context of the system (why do we ask questions, what is the level of answer expected), and what are the searched knowledge bases.

Protocols of answers were little studied until now. It is often assumed there is only one correct answer to a question, or that different answers are complementary or equivalent, and thus are all correct. However, the question of the relevance of the information returned often arises. In TREC, the problem was partially solved by considering an answer in relation with a supporting document that should help to assess the veracity of the answer according to the context it provides. Thus, answers that are different depending on the period covered by the documents will be all accurate. This means that an answer cannot be considered correct by itself, without the passage that justifies it. Thus, to the question "Who married Tom Cruise?", Nicole Kidman is the correct answer in the collection AQUAINT. This raises the problem of presentation of the answer inside its justifying context and a correct answer would be "in 2001, Nicole Kidman". The problem of management of claims should be resolved as well, producing responses indicating "according

[9] http://www-03.ibm.com/innovation/us/watson/ April 2011

to X, the answer is Y, as it is not always possible to choose among several answers. This problem will be even more crucial if searched sources do not have same reliability and this leads to the differentiation of the search for answers on the Web and in a reference collection certified as to its content.

The search for specific information on the Web has its own peculiarities and the problem is a little different than looking for specific information in a collection of smaller size. Characteristics that induce strategies dedicated to the Web are a) its size: the information sought is likely to exist in a form similar to that of the question b) its redundancy: the correct answer is probably the proposition which is found in several documents and c) its multilinguism: the answer can be found in a language different from the language question, and it would encourage under-represented languages, since the information provided in these languages would be sources of answers, regardless of the interrogation language.
Question-answering systems have thus adapted their strategies to the particularities of Web search, whether to formulate queries, which can often be very specific, or to extract answers: intensive use of redundancy to select the more probable answer. Note that QA systems usually use Google to search for documents, and very often only select the returned snippets. Indeed, criteria for selecting passages in systems are based on common words between question and passage as well as their proximity and Google implements these criteria in its selection process of documents. With such a light approach, systems can claim to answer 60% of the questions. It remains that the 40% unanswered questions require more elaborate treatments, to deal with linguistic variation, ambiguity of natural language and finding rare information.
Systems of question answering on the Web have to offer crosslingual search. Current solutions show performances less than the monolingual frame, as are added translation ambiguities and problems of lexicon coverage, particularly as regards proper names and acronyms. Currently, only the translation problems have been studied in QA systems. Other interesting possibilities would be to use multilingual sites, or to guide the research on the language depending on each question: a question about a particular culture is more likely to be found in its language.
Finally, it is important to overcome factual questions and interrogation process limited to a single exchange, to study other types of questions and especially to integrate the notion of context in the process. First, the application context: why does one perform a search, for what purpose? What level of knowledge of the user? Considering these aspects will lead to produce different answers to the same questions. Then, if the questioning process is iterative, this should lead to be able to treat more questions and give more accurate and complete answers. Users do not always perceive the implicit that exists in their own request. QA systems would become closer to dialogue systems, at least for managing the interaction ((Rosset et al., 2005), (Quarteroni et al., 2009)). This raises the problem of the evaluation methodology of such contextual system for evaluating systems under the same conditions.

REFERENCES

Abney S. (1996). Partial Parsing via Finite-State Cascades, *J. of Natural Language Engineering, 2*(4): 337-344.

Agichtein E., Lawrence S., Gravano L. (2001). Learning Search Engine Specific Query Transformations for Question Answering, *proceedings of WWW10*.

Agichtein, E. and Castillo, C. and Donato, D. and Gionis, A. and Mishne, G. (2008). Finding high-quality content in social media, *Proceedings of the international conference on Web search and web data mining*.

Ahn K., Alex B., Bos J., Dalmas T., Leidner J.L. et Smillie M.B. (2004). Cross-lingual Question Answering with QED, *Working Notes of CLEF Cross-Language Evaluation Forum*, Bath UK, pp. 335-342.

Atzeni, P., Basili, R., Haltrup Hansen, D., Missier, P., Paggio, P., Pazienza, M. T., Zanzotto, F. M. (2004). Ontology-Based Question Answering in a Federation of University Sites: The MOSES Case Study. *NLDB 2004*, pp. 413-420.

Awadallah R., Rauber A. (2006). Web-based Multiple Choice Question Answering for English and Arabic Questions, *European Conference on Information Retrieval (ECIR 06)*, pp. 515-518.

Barr A., Feigenbaum E. A., (Eds) (1981). *The Handbook of Artificial Intelligence*, vol1, William Kaufmann, Inc, pp. 281-316.

Baumgartner R., Frölich O., Gottlob G., Harz P., Herzog M., Lehmann P. (2005). Web Data Extraction for Business Intelligence: the Lixto Approach, *BTW 05*.

Bensley J., Hickl A. (2008). Application of LCC's GROUNDHOG System for RTE-4, *TAC 2008 Proceedings*.

Berthelin J.B., de Chalendar G., Ferret, O., Grau, B., ElKateb F., Hurault-Plantet, M., Illouz G., Monceaux L., G., Robba I., Vilnat A. (2003). « Trouver des réponses sur le Web et dans une collection fermée », workshop Recherche d'Information: un nouveau passage à l'échelle, *INFORSID*.

Bos, J., Nissim, M. (2006). Cross-Lingual Question Answering by Answer Translation, *Workshop CLEF, ECDL conference*.

Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedemann, J. (2006). Linguistic Knowledge and Question, *n° spécial de TAL, Répondre à des questions*, dir. B. Grau et B. Magnini, Volume 46, Numéro 3.

Bouma, G., Fahmi, I., Mur, J., van Noord, G., van der Plas, L., Tiedemann, J. (2006b). The University of Groningen at QA@CLEF 2006: Using Syntactic Knowledge for QA, *workshop CLEF, ECDL conference*.

Bowden, M., Olteanu, M., Suriyentrakorn, P., Clark, J., Moldovan, D. (2006). LCC's PowerAnswer at QA@CLEF 2006, *Workshop CLEF, ECDL conference*.

Brill E., Lin J., Banko M., Dumais S., Ng A., (2001). "Data-Intensive Question Answering", *Proceedings of TREC10*, Gaithersburg, MD.

Calais Pedro V., Ko, J., Nyberg E., Mitamura, T., (2004). An Information Repository Model for advanced Question Answering Systems, *LREC*.

Calais Pedro V., Nyberg E., Carbonell, J., (2006). Federated Ontology Search, *SIIK 2006*.

de Chalendar G., Dalmas T., Elkateb-Gara F., Ferret O., Grau B., Hurault-Plantet M., Illouz G., Monceaux L., Robba I., Vilnat A. (2002). The Question Answering System QALC at LIMSI, Experiments in Using Web and WordNet, *Trec 11*, Notebook page 457-467

de Chalendar G., Ferret, O., Grau, B., ElKateb F., Hurault-Plantet, M., Monceaux L., G., Robba I., Vilnat A. (2003). « Confronter des sources de connaissances différentes pour obtenir une réponse plus fiable », *TALN*, Nancy.

Chklovski Timothy and Pantel Patrick (2004). VerbOcean: Mining the Web for Fine-Grained Semantic Verb Relations. In *Proceedings of Conference on Empirical Methods in Natural Language Processing (EMNLP-04)*. Barcelona, Spain.

Chung, H. and Song, Y.I. and Han, K.S. and Yoon, D.S. and Lee, J.Y. and Rim, H.C. and Kim, S.H. (2004). A practical QA system in restricted domains, Workshop on Question Answering in Restricted Domains. *42nd Annual Meeting of the Association for Computational Linguistics (ACL-2004)*.

Clarke C.L.A., Cormack G.V., Lynam T.R. (2001a). Exploiting Redundancy in Question Answering, *SIGIR'01*.

Clarke C.L.A., Cormack G.V., Lynam T.R., Li C.M., McLearn G.L. (2001). "Web Reinforced Question Answering (MultiText Experiments for TREC 2001)", *Proceedings of TREC10*, Gaithersburg, MD.

Clark P., Harrison P. (2009). An Inference-Based Approach to Recognizing Entailment. *In Proceedings of 2009 Text Analysis Conference (TAC'09)*, Gaithsburg, Maryland.

Costa L. F., Esfinge (2006). A Question Answering System in the Web using the Web, *Proceedings of 11th EACL*.

Cui H., Sun R., Li K., Kan M.-Y., Chua T.-S. (2005). Question answering passage retrieval using dependency relations. *In SIGIR '05: Proceedings of the 28th annual international ACM SIGIR conference on Research and development in information retrieval*, pages 400–407.

Day, M.Y. and Ong, C.S. and Hsu, W.L. (2007). Question classification in English-Chinese cross-language question answering: an integrated genetic algorithm and machine learning approach, *IEEE International Conference on Information Reuse and Integration, IRI 2007*, p. 203—208.

Demner-Fushman Dina and Lin Jimmy (2007). Answering clinical questions with knowledge-based and statistical techniques. *Computational Linguistics*, 33(1):63–103.

Doan-Nguyen, H. and Kosseim, L. (2006). Using Terminology and a Concept Hierarchy for Restricted-Domain Question-Answering, *Research on Computing Science, Special issue on Advances in Natural Language Processing*, vol. 18.

Dumais, S., Banko, M., Brill, E., Lin, J., Ng A. (2002). Web Question Answering : Is More Always Better ?, *Proceedings of SIGIR'02*.

El Ayari S., Grau B., Ligozat A.-L. (2010). Fine-grained linguistic evaluation of question answering systems, *LREC Conference*.

Fellbaum C. (1998). WordNet: *An Electronic Lexical Database*, MIT Press, Cambridge, MA.

Ferrandez, S. and Toral, A. and Ferrandez, O. and Ferrandez, A. and Munoz, R. (2009). Exploiting Wikipedia and EuroWordNet to solve Cross-Lingual Question Answering, *Information Sciences*, vol. 179, n°20, p. 3473—3488.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz, G., Jacquemin, C. (2001). "Document selection refinement based on linguistic features for QALC, a question answering system", *Proceedings of Recent Advances in Natural language Processing (RANLP)*, Tsigov Chark, Bulgaria.

Ferret, O., Grau, B., Hurault-Plantet, M., Illouz G., C. Jacquemin, Monceaux L., G., Robba I., Vilnat A. (2002). "How NLP Can Improve Question Answering", *journal Knowledge Organization, Vol. 29*, N°3-4, pages 135-155.

Ferrés, D. and Rodriguez, H. (2006). Experiments adapting an open-domain question answering system to the geographical domain using scope-based resources, *Proceedings of the Workshop on Multilingual Question Answering (ACL)*.

Fleischman M., Echihabi A., and Hovy E.H. (2003). Offline Strategies for Online Question Answering: Answering Questions Before They Are Asked. *Proceedings of the ACL Conference*. Sapporo, Japan.

Frank, A. and Krieger, H.U. and Xu, F. and Uszkoreit, H. and Crysmann, B. and Jörg, B. and Schäfer, U. (2007). Question answering from structured knowledge sources, *Journal of Applied Logic, vol. 5:1*.

Gillard L., Sitbon L., Bellot P., El-Bèze M. (2005). Dernières évolutions de SQuaLIA, le système de Questions/Réponses du LIA, *n° spécial de la revue TAL, Répondre à des questions*, dir. B. Grau et B. Magnini, Volume 46, Numéro 3.

Grau B., Ferret O., Hurault-Plantet M., Monceaux L., Robba I., Vilnat A., Jacquemin C. (2006). Coping with Alternate Formulations of Questions and Answers, in *Advances in Open-Domain Question-Answering*, Strzalkowski & Harabagiu (eds), Series: Text, Speech and Language Technology, Vol. 32, Springer.

Grappy A., Grau B. (2010). Answer type validation in question answering systems, *9th RIAO Conference (RIAO 2010)*.

Green B., Wolf A., Chomsky C., Laughery K. (1986). "BASEBALL: An Automatic Question Answerer", in *Readings in Natural Language Processing*, eds by Barbara J. Grosz, Karen Spark Jones, Bonnie L. Webber, Morgan Kaufmann Publishers, Inc, pp. 545-550.

Grefenstette, G. (1999). The world wide web as a resource for example-based machine translation tasks. In *ASLIB Conference on Translating and the Computer*, volume 21, London, UK.

Harabagiu, S., Pasca, M., Maiorano, J. (2000). "Experiments with Open-Domain Textual Question Answering". *Proceedings of Coling'2000*, Saarbrucken, Germany.

Hartrumpf, S. (2005). University of Hagen at QA@CLEF 2005: Extending knowledge and deepening linguistic processing for question answering. In *Results of the CLEF 2005 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2005 Workshop*.

Hartrumpf, S. and Leveling J. (2006). University of Hagen at QA@CLEF 2006: Interpretation and normalization of temporal expressions. *In Results of the CLEF 2006 Cross-Language System Evaluation Campaign, Working Notes for the CLEF 2006 Workshop* (edited by Nardi, Alessandro; Carol Peters; and Jos Luis Vicedo). Alicante, Spain.

Herrera, J., Rodrigo, A., Penas, A., Verdejo, F. (2006). F. UNED submission to AVE 2006, *Working Notes for the CLEF 2006 Workshop (AVE)*.

Hermjakob U., Echihabi A., Marcu D. (2002). "Natural Language Based Reformulation Resource and Web Exploitation for Question Answering", *proceedings of TREC11*, Gaithersburg, MD.

Hickl, A., Williams, J., Bensley, J., Kirk Roberts, Y. S. & Rink., B. (2006). B. Question Answering with LCC's Chaucer at TREC 2006, *Proceedings of The Fifteenth Text Retrieval Conference (TREC 2006)*.

Hovy, E. , Hermjacob, U., Lin C-Y., Ravichandran, D. (2001). "Towards Semantics-Based Answer Pinpointing", *DARPA Human Technology Conference (HLT)*, San Diego.

Iftene, A., Moruz, A.M. (2009). UAIC Participation at RTE5. *In Text Analysis Conference (TAC 2009) Workshop - RTE-5 Track. National Institute of Standards and Technology (NIST)*.

Ittycheriah, A., Franz, M. & Roukos, S. (2001). "IBM's Statistical Question Answering System – TREC-10". *Proceedings of the Text retrieval conference, TREC 10*, Gaithersburg, MD. NIST Eds..

Jacquemin, C. (2001). *Spotting and Discovering Terms through NLP*. Cambridge, MA: MIT Press.

Jacquemart P., Zweigenbaum P. (2003). Towards a medical question-answering system: a feasibility study. In Robert Baud, Marius Fieschi, Pierre Le Beux, and Patrick Ruch, editors, *Proceedings Medical Informatics Europe*, volume 95 of *Studies in Health Technology and Informatics*, pages 463-468,

Jijkoun V., Mishne G., de Rijke M., Schlobach S., Ahn D. et Muller K. (2004a). The University of Amsterdam at QA@CLEF2004, *Working Notes of CLEF Cross-Language Evaluation Forum*, Bath UK, pp. 321-325.

Jijkoun, V. and De Rijke, M. (2004b). Answer selection in a multi-stream open domain question answering system, *Advances in Information Retrieval, Lecture Notes in Computer Science, Springer*, Volume 2997/2004, 99-111.

Katz B., Lin J., Loreto D., Hildebrandt W., Bilotti M., Felshin S., Fernandes A., Marton G., Mora F. (2003). Integrating Web-based and Corpus-based Techniques for Question Answering, *Proceedings of the Twelfth Text Retrieval Conference (TREC 2003).*

Kouylekov, M., Negri, M., Magnini, B., Coppola, B. (2006). Towards Entailment-based Question Answering: ITC-irst at CLEF 2006, *Working Notes for the CLEF 2007 Workshop (AVE).*

Kwok C. C. T., Etzioni O. (2001). Weld D. S., Scaling Question Answering to the Web, *proceedings of WWW10*.

Laurent, D., Nègre, S., Séguéla, P. (2005). QRISTAL, le QR à l'épreuve du public, *n° spécial de la revue TAL, Répondre à des questions*, dir. B. Grau et B. Magnini, Volume 46, Numéro 3.

Laurent, D., Séguéla, P., Nègre, S. (2005b). Cross lingual question answering using QRISTAL for CLEF 2005. *In Working Notes, CLEF Cross-Language Evaluation Forum*, Vienna, Austria.

Laurent, D., Séguéla, P., Nègre, S. (2006). Cross Lingual Question Answering using QRISTAL for CLEF 2006, *Workhop CLEF, ECDL conference*.

Lehnert W. (1977). "Human and computational question answering", *Cognitive Science*, vol. 1, p. 47-63.

Lerman, K., Getoor, L., Minton, S., and Knoblock, C.A. (2004). Using the structure of web sites for automatic segmentation of tables. *In Proceedings of ACM SIG on Management of Data (SIGMOD-2004).*

Ligozat A.-L., Grau B., Robba I., Vilnat A. (V). L'extraction des réponses dans un système de question-réponse, *TALN*, Leuven.

Ligozat A.-L., Grau B., Robba I., Vilnat A. (2006b). Evaluation and Improvement of Cross-Lingual Question Answering Strategies, *Workshop MLQA, EACL*

Ligozat A.-L., Grau B., Vilnat A., Robba I., Grappy A. (2007). Towards an automatic validation of answers in Question Answering, *19th IEEE International Conference on Tools with Artificial Intelligence - Vol.2 (ICTAI 2007)* pp. 444-447.

Lin J., Katz B. (2003). Question Answering from the Web Using Knowledge Annotation and Knowledge Mining Techniques, *CIKM'03*.

Lopez, V., Pasin, M., Motta, E., (2005). AquaLog: An Ontology-portable Question Answering System for the Semantic Web. *In the proceedings of ESWC 2005 (European Semantic Web Conference).*

Lopez, V., Motta, E., Uren, V., (2006). AquaLog: An ontology-driven Question Answering System to interface the Semantic Web, *Proceedings of the Human Language Technology Conference of the NAACL*, Companion Volume: Demonstrations, pp. 269-272.

Magnini B., Negri M., Prevete R., Tanev H., (2002). "Mining Knowledge from Repeated Co-occurrences: DIOGENE at TREC 2002", Proceedings of the Text retrieval conference, *TREC11*, Gaithersburg, MD. NIST Eds.

Magnini B., Negri M., Prevete R., Tanev H., (2002b). "Is It the Right Answer? Exploiting Web Redundancy for Answer Validation", in *Proceedings of the ACL*.

Magnini B., Negri M., Prevete R., Tanev H., (2002c). Comparing Statistical and Content-Based Techniques for Answer Validation on the Web, *Proceedings du VIII Convegno AI*IA*.

McGuinness, D. L., Pinheiro Da Silva, P., (2004a). Trusting Answers on the Web, in *New Directions in Question Answering*, Mark T. Maybury ed., Chap. 21, AAAI/MIT Press.

McGuinness, D. L., Pinheiro Da Silva, P., (2004b). Explaining Answers from the Semantic Web: the Inference Web Approach, *Journal of Web semantics, vol1*, n°4, pp. 397-413

Minock, M. (2005). Where are the 'killer applications' of restricted domain question answering?, in *Proceedings of the IJCAI Workshop on Knowledge Reasoning in Question Answering*, page 4, Edinburgh, Scotland.

Moldovan, D., Harabagiu S., Girju R., Morrarescu P., Lacatusu F., Novishi A., Badulescu A., Bolohan O., (2002). "LCC Tools for Question Answering", *Proceedings of the Text retrieval conference, TREC11*, Gaithersburg, MD. NIST Eds.

Moldovan D., Clark C., , Harabagiu S., Maiorano S., (2003). "COGEX: A Logic Prover for Question Answering", *proceedings of HLT-NAACL*, Edmonton, pp. 87-93.

Mollà, D. and Vicedo, J.L., (2007). Question answering in restricted domains: An overview, *Computational Linguistics*, vol 33:1.

Moriceau V., Tannier X., Grappy A., Grau B., (2008). Justification of answers by verification of dependency relations – The French AVE task, Working Notes of CLEF Workshop, ECDL conference.

Moschitti, A. and Quarteroni, S. and Basili, R. and Manandhar, S., (2007). Exploiting syntactic and shallow semantic kernels for question answer classification, *ACL*.

Nadeau D., Sekine S., (2007). A survey of named entity recognition and classification, *Journal of Linguisticae Investigationes, 30:1,*.

Neumann G. et Sacaleanu , B., (2004). Experiments on Robust NL Question Interpretation and Multi-layered Document Annotation for a Cross-Language Question / Answering System, *Working Notes of CLEF Cross-Language Evaluation Forum*, Bath UK, pp.311-320.

Neumann G. et Sacaleanu B., (2005). DFKI's LT-lab at the CLEF 2005 Multiple Language Question Answering Track, *Working Notes of CLEF Cross-Language Evaluation Forum*.

Newman, E., Stokes, N., Dunnion, J. , Carthy, J., (2005). UCD IIRG Approach to the Textual Entailment Challenge, *Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment*, 53-56

Prager J., Brown E., Radev D. R., Czuba K., (2000). "One Search Engine or two for Question-Answering", *proceedings of TREC9*, Gaithersburg, MD, p 235-240.

Perret L., (V). Question Answering System for the French Language, *Working Notes of CLEF Cross-Language Evaluation Forum*, Bath UK, pp. 295-305.

Plamondon, L. and Lapalme, G. and Kosseim, L., (2003). The quantum question answering system at TREC 11, *proceedings of TREC11*.

Quarteroni, S. and Manandhar, S., (2009). Designing an interactive open-domain question answering system, *Language Engineering Journal*, vol. 15, n°1, p. 73—95.

Rinaldi, F. and Dowdall, J. and Schneider, G. and Persidis, A., (2004). Answering questions in the genomics domain, *ACL 2004 Workshop on Question Answering in restricted domains*.

Rosset S., Galibert O., Illouz G., Max A., (2005). Interaction et recherche d'information : le projet RITEL, *n° spécial de la revue TAL, Répondre à des questions*, dir. B. Grau et B. Magnini, Volume 46, Numéro 3.

Sang, E.T.K. and Bouma, G. and de Rijke, M., (2005). Developing offline strategies for answering medical questions, *Proceedings of the AAAI-05 Workshop on Question Answering in Restricted Domains*.

Schlobach S., Ahn D., de Rijke M., Jijkoun V., (2007). Data-driven type checking in open domain question answering, *J. Applied Logic, volume 5 :1*.

Soubbotin, M. M., Soubbotin, S. M., (2001). "Patterns of Potential Answer Expressions as Clues to the Right Answers". *Proceedings of the Text retrieval conference, TREC 10*, Gaithersburg, MD. NIST Eds.

Soubbotin, M. M., Soubbotin, S. M., (2002). "Use of patterns for Detection of Likely Answer Strings: a Systematic Approach", *Proceedings of the Text retrieval conference, TREC 11*, Gaithersburg, MD. NIST Eds.

Sutcliffe, R. F. E., White, K., Slattery, D., Gabbay, I., Mulcahy, M., (2006). Cross-Language French-English Question Answering using the DLT System at CLEF 2006, *Workshop CLEF06*.

Tanev H., Negri M., Magnini B. et Kouylekov M., (2004). The DIOGENE Question Answering System at CLEF-2004, *Working Notes of CLEF Cross-Language Evaluation Forum*, Bath UK, pp.325-333.

Tatu, M., Iles, B., Slavick, J., Novischi, A., Moldovan, D., (2006). COGEX at the Second Recognizing Textual Entailment Challenge, Proceedings of the PASCAL Challenges Workshop on Recognising Textual Entailment.

Tellex, S., Katz, B., Lin, J., Fernandes, A., Marton, G., (2003). Quantitative evaluation of passage retrieval algorithms for question answering, *Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval*, Toronto, Canada, p. 41-47.

Voorhees E. M., (2001). "The TREC question answering track", *Journal of Natural Language Engineering, vol 7:4*.

Wang R., Neumann G., (2008). An Accuracy-Oriented Divide-and-Conquer Strategy for Recognizing Textual Entailment, *TAC 2008 proceedings*.

Wang, K. and Ming, Z. and Chua, T.S., (2009). A syntactic tree matching approach to finding similar questions in community-based QA services, *Proceedings of the 32nd international ACM SIGIR conference on Research and development in information retrieval*.

Wang R., Zhang Y., Neumann G. (2009b). A Joint Syntactic-Semantic Representation for Recognizing Textual Relatedness. *In Text Analysis Conference TAC 2009 WORKSHOP Notebook Papers and Results*, Pages 1-7, National Institute of Standards and Technology (NIST), Gaithersburg, Maryland, USA.

Yu, H. and Sable, C. and Zhu, H.R., (2005). Classifying medical questions based on an evidence taxonomy, *Workshop on Question Answering in Restricted Domains. 20th National Conference on Artificial Intelligence (AAAI-05).*

Zajac, R., (2001). Towards Ontological Question Answering, *Workshop on Open Domain Question Answering, ACL*.