AUTOMATED QUESTION-ANSWERING TECHNIQUES AND THE MEDICAL DOMAIN

Andrea Andrenucci
Department of Computer and System Sciences, Stockholm University/ Royal Institute of Technology,
Forum 100, SE-16440, Kista, Sweden
andrea@dsv.su.se

Keywords: Automated Question Answering, Natural Language Interfaces, Medical Applications.

Abstract: The question-answering (QA) paradigm, i.e. the process of retrieving precise answers to natural language (NL) questions, was introduced in late 1960-ies and early 1970-ies within the framework of Artificial Intelligence. The advent of WWW and the need to provide advanced, user-friendly search tools has extended the QA paradigm to a larger audience of people and a larger number of fields, including medicine. This paper reviews three research approaches utilized in automated QA in medical domains and discusses their application areas.

1. INTRODUCTION

The question-answering (QA) paradigm, i.e. the process of retrieving precise answers to natural language (NL) questions, was introduced in late 1960-ies and early 1970-ies within the framework of Artificial Intelligence. From the beginning it was mainly an academic research field and there were hardly any commercially applicable QA applications. The advent of WWW and the need to provide advanced, user-friendly search tools has extended the QA paradigm to a larger audience of people and a larger number of fields, including medicine, since medical content is one of the most retrieved types of information on the WWW.

This paper discusses which of three major QA approaches, i.e. deep Natural Language Processing (NLP), Information Retrieval (IR) enhanced by shallow NLP, and Template-based QA, better fit medical applications, eliciting their context of pertinence. To our knowledge, this is the first formal comparison of the three QA approaches that focuses on the medical domain.

The next three sections discuss the approaches and provide some examples of their application in the medical domain; section five and six pinpoints the application areas that fit each technique.

2. NATURAL LANGUAGE PROCESSING (NLP)

A common feature of deep NLP systems is that they convert text input into formal representation of meaning such as logic (first order predicate calculus), semantic networks, conceptual dependency diagrams, or frame-based representations (Jurafsky and Martin, 2000, p. 502). In other words deep NLP systems perform a semantic analysis of text in NL. Semantic analysis is the process of studying the meaning of a linguistic input and giving a formal representation of it.

Jurafsky and Martin (2000, p. 548) provide a possible approach for semantic analysis (see figure 1 on the next page): the user input is first passed through a syntactic parser, whose output, represented with a parse tree, is then processed by a semantic analyzer which delivers a meaning representation.

A medical QA system that implements this approach is the ExtrAns system (Rinaldi et al., 2004). The system derives logical representations of both user questions and the documents in the collection. The documents are analysed in an off-line stage and their semantic form is stored in a Database. In an on-line stage user questions are converted into their semantic representation, prior to being compared to the representations of the documents in the matching process. When a match

occurs, the sentences that originated the match are extracted as possible answers to the user question.

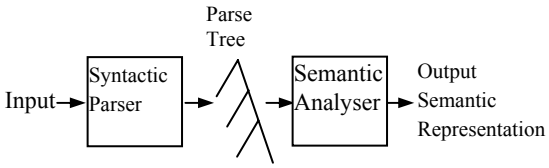

Figure 1: The steps in Semantic Analysis

Drawbacks of the deep NLP approach are its computational intensiveness and its high processing time (Andrenucci and Sneiders, 2005, Rinaldi et al., 2004) as well portability difficulties (Andrenucci and Sneiders, 2005, Hartrumpf 2006). Figure 2 (Androutsopoulos, Ritchie, and Thanisch, 1995) shows the possible architecture of a typical deep NLP QA system. Six components (linguistic front-end) change when the input language changes, and three components (domain-dependent knowledge) change when the knowledge domain changes. The domain dependent knowledge contains information specific for the domain of interest: a lexicon and a world model. The lexicon contains admissible vocabulary words from the knowledge domain. The world model describes the structure of the domain of interest, i.e. the hierarchy of classes of the domain objects, plus the properties and the constraints that characterize the relationship between them. The linguistic front-end parses and analyses the user input in NL.

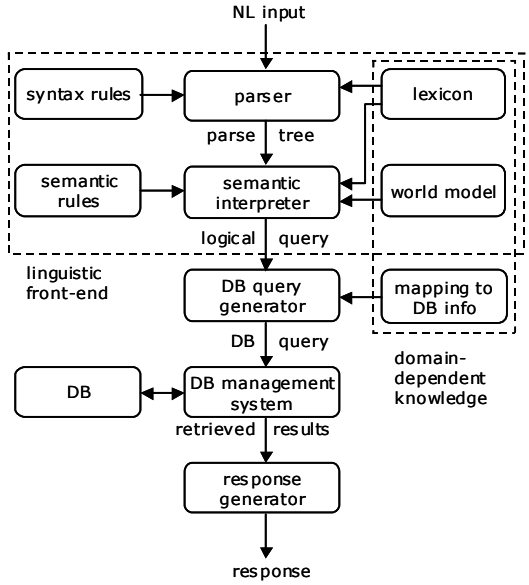

Figure 2 : Architecture of a typical deep NLP system, originally from Androutsopoulos et al., 1995

3. INFORMATION RETRIEVAL (IR) AND SHALLOW NLP

IR has evolved from document retrieval systems to passage retrieval systems, which focus on retrieving text passages rather than entire documents. Answers are extracted with the help of *shallow NLP*, which does not imply text understanding, i.e. semantic analysis of NL input. Instead it focuses on extracting text chunks, matching patterns or entities that contain the answer to user questions. For instance in a question like "Who discovered the polio vaccine?", the presence of the interrogative pronoun "who" implies the extraction of an entity of type "person name" associated with the keywords, "discovered", "polio", "vaccine".

This approach has been implemented in several biomedical systems. Rindflesch et al. (2000) utilized named entity recognition techniques to identify drugs and genes in biomedical documents, then the keywords which connect them (predicates). Craven and Kumlien (1999) utilized "bag of words" at the sentence level, to extract relations between proteins and drugs from the information stored in Medline articles (MEDLINE, 2006).

The IR approach is more domain-independent than traditional NLP, but requires the answer to be explicitly present in the text (Voorhees, 2001). Furthermore answers retrieved with IR techniques are less justified by the context, since they only focus on extracting text snippets containing words that are present in the user question (Laurent, Seguela and Negre, 2006).

This approach is typical for information extraction and is largely used in the Text REtrieval Conferences (Voorhees, 2001), which aim at comparing QA systems that retrieve mainly factoid questions. Several systems that implement the shallow NLP approach exploit data redundancy (Brill et al., 2001), i.e. a number of text passages that contain similar statements, in order to find reliable answers. For example Sekimizu, Park and Tsujii (1998) exploits domain specific verbs, which occur frequently in MEDLINE abstracts, in order to locate the biomedical terms that are respectively subject and object terms for the verbs, and thereafter classify their relations (e.g. Protein X *regulates* Protein Y). Similarly Spasic, Nenadic, and Ananiadou (2003) measure frequent co-occurrences of biomedical verbs with unclassified terms in order to extract domain specific terms.

A medical search system that implements both IR techniques and deeper NLP techniques is PERSIVAL (McKeown et al., 2001). The system supports user search and summarization of medical information with the help of representations of

medical texts and patient records. The system processes medical documents with part of speech tools and with a finite state grammar (that regulates syntactic constraints) in order to extract multi-word terms. This step is similar to the syntactic analysis provided with the help of syntax rules in fig. 2.

Also similarly to the deep NLP approach, this system utilizes a well defined world model (see section 2), provided by the UMLS medical knowledge base (McCray and Nelson, 1995), in order to define the semantic category and the level of specificity of the extracted terms. This is a kind of semantic analysis.

The user profiles and the medical documents are represented with vectors, which are typical IR representation models. The vectors enclose the semantic categories of the medical terms and their associated values. The representations are then compared calculating the cosine similarity of the vectors (Salton and Buckley, 1988), which is also a typical IR technique. Tests conducted with the system (Teufel et al., 2001) have shown that the semantic analysis enhances precision and recall of the system, compared to standard IR techniques.

4. TEMPLATE-BASED APPROACH

Template-based QA extends the pattern matching approach and exploits a collection of manually created question templates, i.e. questions which have open concepts to be filled with data instances, mapped into the conceptual model of the knowledge domain. The templates generally cover the most frequently asked questions (FAQs) of the domain (Sneiders, 2002b), and can be either static, where each template is a question linked to a piece of static text, or dynamic and parameterized if they cover a structured database (Sneiders, 2002a). A question template is viewed as a predicate with variable and fixed parameters:

$$\exists\ data_1,\ldots data_n\text{:Q(fixed}_1,\ldots\text{fixed}_m,\text{variable}_1,\ldots \text{variable}_n)$$

During the process of matching a template to a user question, the fixed parameters ($fixed_1,\ldots fixed_m$) are bound to the user question. If there are database data instances ($data_1,\ldots data_n$) that fit the variable parameters ($variable_1, \ldots variable_n$) and make the statement Q true, then these data instances constitute the answer. This approach has been utilized on a medical portal aiming at providing cross language QA in matters of psychology and psychotherapy (http://www.web4health.info).

A similar approach, which focuses on classifying user questions with the help of pre-determined semantic patterns, is applied in a feasibility study for creating a QA prototype for the oral surgery domain (Jacquemart and Zweigenbaum, 2003). The patterns are created with triples that contain two concepts and their relation (Concept A – Relation – Concept B). The relation between the concepts is defined with the help of the UMLS Semantic Network (McCray and Nelson, 1995).

The Medline Button system (Cimino et al., 1992) tries to automate the question generation process creating semantic patterns of concepts that occur frequently in user questions. The system then instantiates the generic concepts in the templates with terms that are specific for the search context and user interests. For instance the template "Does <*procedure*> cause <*disease*>?" is instantiated to "Does *chest x-rays* causes *cancer*?" if the user is interested in those topics.

The PICO-format (Sackett et al., 2004), utilized in several medical QA systems (Niu et al., 2003, Demner-Fushman and Lin, 2005), consists of templates that classifies NL input with the help of a conceptual structure that represent the key elements of clinical questions: Problem (the primary problem of the patient), Intervention (medication or therapeutic procedure), Comparison (of the actual intervention to other possible interventions) and Outcome (the effect of the intervention).

A system that implements IR and templates-based techniques is the EPoCare QA system (Niu et al., 2003). Candidate answers are first retrieved with standard IR techniques and then classified with the PICO format, prior to being matched to PICO-formatted user questions. The system also tries to classify the relations between the PICO conceptual units, for instance individuating cause-effect relations between interventions and outcomes.

5. QA TECHNIQUES AND THE MEDICAL DOMAIN - DISCUSSION

As mentioned in section 3, the IR approach distinguish the expected answer type (e.g. person, place or time) with the help of the so-called "wh-words" in the user question (e.g. who, where, when). Niu et al. (2003) states that this classification is not appropriate for the medical domain for the following reasons:

1) Questions about patient care usually deal with diagnosis, treatments, prognosis and outcome of the treatments (Richardson et al., 1995). This require a methodology for identifying answer types that is

different from the traditional approach utilized for generic "factoid" QA systems.

2) Answers to "when"-questions in medical area are usually related to relative time (Q: "When should I eat my medicine?" A: "One hour before lunch") rather than absolute time/dates, which is typical for generic QA systems (Q: "When was America discovered?" A: "1492"), or address a clinical condition (Q: "When should I see a therapist?" A: "You should consider professional advice if your personal problems are affecting your quality of life and social functions at work or at home for more than a month"). This requires a deeper semantic interpretation of the user question.

3) Yes-no questions, i.e. questions that require yes or no answer (e.g. "Is cognitive behaviour a good therapy method for a person suffering from anxiety disorder?") are not considered by systems that focus on "wh"-questions.

Furthermore IR techniques extract answers containing words that are present in user questions, without considering contextual information in the text that could be relevant to provide and justify answers (Niu and Hirst, 2004, Laurent, Seguela and Negre, 2006). In medical applications correct answers may be missed or incorrect answers may be retrieved if contextual information is not understood, since the context may provide more evidence, clarify or even contradict the extracted snippets (Niu et al., 2003).

Deep NLP-based and Template based QA are the techniques that better fit QA in medical matters. Both approaches handle more advanced types of questions that implies understanding of their context, such as yes-no questions, and have shown better results when it comes to requests for "advice-giving" (e.g. "How to…" questions) since they perform a semantic interpretation of user input (Andrenucci and Sneiders, 2005, Laurent, Seguela and Negre, 2006). In the template based approach the interpretation is done manually, individuating for each single template the concepts that cover a part of the conceptual model of the knowledge domain. In the NLP approach the interpretation is done automatically by the system as questions are asked, mapping user questions and candidate answers into a formal semantic representation.

Unlike IR enhanced by shallow NLP, those techniques do not rely on data redundancy, which is more likely to be useful in large, open domains (Molla et al., 2003). Information in restricted domains, such as the medical one, is usually well structured and it is unlikely that answers to the same question are redundantly present in several places of the information source (Niu et al., 2003). Deep NLP and Template based QA are the techniques that are more often utilized to form interfaces to structured data (Andrenucci and Sneiders, 2005).

However there are some important differences that determine the context of application of the two afore-mentioned techniques. The NLP approach provides a natural flow in the user-computer dialogue that resembles human-to-human communication, thanks to the implementation of realistic discourse planning models; see for instance (Buchanan et al., 1995). NLP-based systems may also implement dialectical argumentation techniques in order to be more persuasive while giving advice in health matters. One example is the DAPHNE system (Cawsey, Grasso, and Jones, 1999), which provides advice for the promotion of healthy nutrition and implements a persuasive conversational model based on providing supports for its claims ("People who eat more fruit have less diseases") and anticipating possible counter arguments and exceptions ("Although you may not like all types of vegetables…").

So in dialoguing or counselling matters that have to resemble the patient-doctor communication, the NLP approach is preferable. The NLP-approach also delivers more reliable answers in comparison to the other approaches (Andrenucci and Sneiders, 2005, Molla et al., 2003, Teufel et al., 2001). For example in Power Answer (Moldovan et al, 2003), the best performing system for TREC 2004 and 2005, a logic proof based on abductive justifications is performed among the candidate answers prior to presenting the valid answers to the users, enhancing the quality and reliability of the results. Power Answer achieved an accuracy of 70% while other medical systems implementing approaches similar to the template based approach achieved 60% (Jacquemart and Zweigenbaum, 2003) of accuracy. Among IR systems Persival (McKeown et al., 2001) achieved precision results that varied between 65 and 89 %, but IR techniques were supported by syntactic and semantic analysis. Deep Semantic Analysis technique has also proved to improve the disambiguation of causal questions, boosting the precision results of the retrieved answers (Girju, 2003).

So in cases where the reliability of the answers is vital, systems enhanced by NLP approach are preferable; for instance medical systems that support practitioners in their decision making process and that provide evidence for the suggested answers (Lin and Demner-Fushman, 2005); the so-called evidence-based medicine (Sackett et al., 2000).

A major drawback of this approach is that development and maintenance of NLP systems are complex and require highly qualified personnel such as programmers, knowledge engineers and database administrators. For example when the NLP QA

needs to be adapted to multi-lingual environment, changes needs to be applied to the whole linguistic module, which includes the lexicon, the world model, the semantic interpreter and the syntactic parser (Androutsopoulos, Ritchie, and Thanisch, 1995). Another drawback is that this approach is computationally intensive and requires high processing time, which makes it difficult to adapt it to the Web (Rinaldi et al., 2004, Hartrumpf, 2006).

Template-based question answering is the most viable approach when it comes to medical information portals on the Web (Andrenucci and Sneiders, 2005). This is due to the following characteristics: 1) its suitability to support multi-lingual content, 2) the relatively easiness of maintenance, 3) its capacity to solve linguistic ambiguities such as word sense disambiguation without computationally expensive software, 4) and its capability to return answers in different formats.

The suitability to support content in several languages has a simple explanation: user questions are matched against question templates that match different interpretations of the same question and contain individual lexicons; this implies that it is only necessary to change individual templates to get a multi-lingual matching.

Template-based QA systems are also easier to manage since they do not require rare skills: the administrator must only have knowledge of the subject domain and possess basic linguistic skills (Sneiders, 2002a).

Thanks to the usage of multiple lexicons, i.e. small individual lexicons attached to each template, linguistic ambiguities are solved at the micro-level rather than at the macro-level. Small lexicons identify mutually exchangeable words (synonyms and their grammatical forms) for every concept within the narrow context of a given template/document, rather than in the context of the whole knowledge domain (Sneiders, 2002a), which is typical for the deep NLP approach. This makes the individuation of word meanings in different contexts an easier and less error-prone process (Sneiders, 2002b, p. 262).

The template-based approach supports also the retrieval of answers in a variety of multimedia forms, such as spoken languages, audio-files and imagery (Andrenucci and Sneiders, 2005, McKeown et al. 2001).

The Template based approach has a high recall level, which fits users who are interested in retrieving complete sets of answers rather than few very precise answers.

A drawback of this approach is that manual creation of the templates is required. This is a tedious process, which poses great consistency demands among the persons who create the templates. Another drawback is that the template-based QA does not provide a natural flow in user/system dialogue or provides dialogues of poor quality. One of the first medical systems trying to use templates while dialoguing with users was Eliza (Weizenbaum, 1966), a conversational agent created to simulate the responses of a psychotherapist. The system did not contain any domain knowledge and the templates utilized regular expression in order to match user input and to create responses that exploited keywords from the input sentences. This resulted often in nonsense answers and nonsense dialogues (Copeland, 1993).

6. CONCLUSIONS

This paper has discussed three main techniques within QA and has pointed out the approaches that are more suitable for medical applications: the deep NLP approach and the template based approach.

The template based approach is the most viable commercially and fits Web-based medical applications that are aimed at retrieving multilingual content in different multimedia formats. Its high recall level makes it the technique that fits users who are more interested in retrieving complete sets of answers rather than few very precise answers.

The deep NLP approach provides a dialogue that better resembles the human-to-human conversation and also delivers more reliable answers. It fits areas where the precision of the retrieved information is crucial, e.g. in decision-support or evidence-based medicine.

IR enhanced by shallow NLP is more appropriate as a search tool for larger or open domains as the Web, since it exploits data redundancy. However it can mainly retrieve factual answers unlike the NLP and the template based approaches, which support more complex types of questions such as requests for “advice-giving”.

REFERENCES

Andrenucci, A. and Sneiders, E., 2005. Automated Question Answering: Review of the Main Approaches. In *ICITA’05*, IEEE press.

Androutsopoulos, I., Ritchie, G. and Thanisch P., 1995. Natural Language Interfaces to Databases: An Introduction. *Journal of Natural Language Engineering, vol. 1*. Cambridge University Press.

Brill, J., et al., 2001. Data-intensive question answering. In *TREC 2001,* NIST.

Buchanan, B.G., Moore, J.D., Forsythe, D., Carenini, G., Ohlsson, S. and Banks, G., 1995. An intelligent

interactive system for delivering individualized information to patients. In *Artificial Intelligence in Medicine, 7*.

Cawsey, A., Grasso, F. and Jones, R., 1999. A conversational model for health promotion on the WWW. In *AIMDM'99*, Springer-Verlag.

Cimino, J.J., Johnson, S., Aguirre A., Roderer N., and Clayton P., 1992. The Medline Button. In *16th Annual Symposium on Computer Applications in Medical Care*.

Copeland, J., 1993. *Artificial Intelligence: A philosophical Introduction.* Blackwell.

Craven, M., and Kumlien, J., 1999. Constructing biological knowledge bases by extracting information from text sources. In *ISMB'99, 8th International Conference on Intelligent Systems for Molecular Biology*.

Demner-Fushman, D. and Lin, J., 2005. Knowledge extraction for clinical Question Answering: preliminary results. In *AAAI'05 Workshop on Question Answering in Restricted Domains,* AAAI press.

Girju R. 2003. Automatic Detection of Causal Relations for Question Answering. In *ACL 2003 Workshop on Multilingual Summarization and Question Answering*. ACL press.

Hartrumpf, S., 2006. Adapting a Semantic Question Answering System to the Web. In *MLQA'06, EACL Workshop on Multilingual Question Answering,* ACL press.

Jacquemart, P. and Zweigenbaum, P., 2003. Towards a medical Question Answering system: a feasibility study. *Studies in Health Technology and Informatics*, IOS Press.

Jurafsky, D., and Martin, J.H., 2000. *Speech and language processing*, Prentice Hall, NJ, USA.

Laurent, D., Seguela P., and Negre, S., 2006. QA better than IR?. In *MLQA'06, EACL Workshop on Multilingual Question Answering*, ACL press.

Lin, J. and Demner-Fushman, D., 2005. "Bag of Words" is not enough for Strength of Evidence Classification. In *AMIA 2005, the Annual Symposium of the American Medical Informatics Association*.

MEDLINE database, 2006. National Library of Medicine. http://www.ncbi.nlm.nih.gov/PuBMed

McCray, A.T. and Nelson, S. J., 1995. The semantics of the UMLS knowledge sources. *Methods Inf Med 34, 1-2*.

McKeown, K., et al., 2001. PERSIVAL, a System for Personalized Search and Summarization over Multimedia Healthcare Information. In *JCDL 2001,* ACM press.

Moldovan, D., et al., 2003. LCC Tools for Question Answering. In *TREC 2003,* NIST.

Molla, D., et al., 2003. NLP for Answer Extraction in Technical Domains. In *EACL 2003,* Morgan Kaufmann.

Niu, Y., and Hirst, G., 2004. Analysis of Semantic Classes in Medical Text for Question Answering. In *ACL Workshop on Question Answering in Restricted Domains*. ACL press.

Niu, Y., Hirst, G., McArthur, G. and Rodriguez-Gianolli, P., 2003. Answering clinical questions by identifying roles in medical texts. In *ACL Workshop on Natural Language Processing in Biomedicine, 41st annual meeting of the Association for Computational Linguistics,* ACL press.

Richardson, W. S., Wilson M., Nishikawa J. and Hayward R., 1995. The well-built clinical question: a key to evidence-based decisions. In *American College of Physicians Journal Club 123, 3*.

Rinaldi, F., Dowdall, J., Schneider, G., and Persidis, A., 2004. Answering Questions in the Genomics Domain. In *ACL 2004 Workshop on Question Answering in Restricted Domains*. ACL press.

Rindflesch, T.C., Tanabe, L. Weinstein, J. N. and Hunter, L., 2000. Edgar: Extraction of drugs, genes and relations from the biomedical literature. In *Pacific Symposium on Biocomputing*.

Sackett, D., Straus, S., Richardson, S., Rosenberg, W., and Haynes, R., 2000. *Evidence-based medicine: how to practice and teach EBM*. Churcill Livingstone.

Salton G., and Buckley, C., 1988. Term weighting approaches in automatic text retrieval. *Information Processing and Management, 25, 5.*

Sekimizu, T., Park, H. and Tsujii, J., 1998. Identifying the interaction between genes and gene products based on frequently seen verbs in Medline abstracts. *Genome Informatics*, Universal Academy Press.

Sneiders, E., 2002a. Automated Question Answering Using Question Templates that Cover the Conceptual Model of the Database. In *NLDB'2002,* Springer-Verlag.

Sneiders, E., 2002b. *Automated Question Answering: Template-Based Approach*. PhD thesis, Royal Institute of Technology, Sweden.

Spasic, I., Nenadic, G., and Ananiadou, S., 2003. Using domain-specific verbs for term classification. In *ACL 2003 Workshop on Natural Language Processing in Biomedicine,* ACL press.

Teufel, S., Hatzivassiloglou, V., McKeown, K., Dunn, K., Jordan, D., Sigelman, S., and Kushniruk, A., 2001. Personalised medical article selection using patient record information. In *AMIA 2001, the Annual Symposium of the American Medical Informatics Association*.

Toulmin, S., 1958. *The uses of argument*. Cambridge University Press.

Voorhees, E., 2001. The TREC Question Answering Track. In *Natural Language Engineering 7*, Cambridge University Press.

Weizenbaum, J., 1966. Eliza – A computer program for the study of natural language communication between man and machine. *Comm. of the ACM 9*, ACM press.