Developing Offline Strategies for Answering Medical Questions

Erik Tjong Kim Sang
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
erikt@science.uva.nl

Gosse Bouma
Computational Linguistics
Rijksuniversiteit Groningen
Groningen, The Netherlands
gosse@let.rug.nl

Maarten de Rijke
Informatics Institute
University of Amsterdam
Amsterdam, The Netherlands
mdr@science.uva.nl

Abstract

We describe ongoing developments on two offline strategies for automatically answering questions in the medical domain: one based on an analysis of the document structure, the other based on dependency parsing. We highlight differences with open domain question answering, and provide a preliminary evaluation of the current state of our strategies.

Introduction

IMIX (Interactive Multimodal Information eXtraction) is a national research framework in the Netherlands which is aimed at developing knowledge and technology for providing focused, multi-modal access to documents in the Dutch language (NWO 2004). The research framework combines work on question answering (QA), dialogue management, user interfaces, as well as multi-modal output; it attaches great importance to developing integrated demonstrators that showcase the project's achievements. The integrated demonstrators will be restricted to the medical domain, and, more specifically, to issues related to RSI (repetitive strain injury). A further restriction is that the envisaged demonstrator should work with documents in Dutch.

In this paper we focus on the question answering (QA) work being carried out within IMIX. The IMIX QA modules are to be deployed as part of a real-time interactive demonstrator, and to deliver the required responsiveness we have decided to implement so-called *offline* strategies for answering questions. While information retrieval techniques are very successful at providing near-instant access to vast amounts of data, the precision required of QA systems makes on-the-fly question answering from unstructured data sources too slow and impractical. Furthermore, for many popular types of (factoid) questions, the semantic information that (likely) answers these questions, occurs in very fixed patterns. For example, for questions like "Where are the islets of Langerhans located?" typical answer patterns are "special groups of cells in the pancreas" and "irregular microscopic structures, varying from a few to hundreds of cells, scattered throughout the PANCREAS and comprising its endocrine portion." Our strategy is to exploit such regularities for offline extraction of semantic data so as to make the data available for rapid and easy access. Finally, having access to such extracted data will allow us to more easily answer certain types of questions than we would be able to if we only did on-the-fly processing; examples include list questions such as "Name liver diseases."

Open domain QA for English has been extensively studied at TREC (Voorhees 2003), and even open domain QA for Dutch has received a fair amount of attention within the CLEF setting (Jijkoun, Mishne, & De Rijke 2004; Ahn *et al.* 2005). However, developing a restricted domain QA system requires more than the techniques used in general QA systems. For a start, question analysis needs to be done from scratch. Available general WordNets are of little usage in a restricted domain, and the available corpora are typically small and thus we cannot rely on redundancy-based search and extraction techniques but need very accurate ones. Furthermore, in our medical domain the required answers are usually larger than the named entities that systems return in response to TREC-style factoids; this obviously complicates the answer generation parts.

In our experience, different answering strategies tend to perform well on different types of questions (Jijkoun & De Rijke 2004). This is why we are developing two different offline strategies for answering Dutch medical questions. One of the aims of the paper is to describe these strategies, and another is to report on their relative performance. Specifically, after an outline of related work in section two, section three provides an analysis of the questions we expect in this domain and defines what the question types and the formats of the expected answers are. We then describe our two offline answering approaches: one based on layout cues, the other based on deep syntactic analysis. We provide an intrinsic evaluation of the underlying extraction approaches as well as an extrinsic evaluation of the usability of the extracted information in initial versions of our QA systems. In the final section we conclude.

Related Work

In recent years, work on medical QA has received a great deal of interest; Zweigenbaum (2003) provides an overview.

Offline methods have proven to be very effective for factoid QA. Different techniques have been used to do the actual extraction. The use of surface patterns to identify answers in corpus-based QA received lots of attention after a team taking part in one of the earlier QA Tracks at TREC showed that the approach was competitive at that stage (Soubbotin & Soubbotin 2002; Ravichandran & Hovy 2002). Different aspects of pattern-based methods have been investigated since. E.g., Ravichandran *et al.* (2003) collect surface patterns automatically in an unsupervised fashion using a collection of trivia question and answer pairs as seeds. These patterns are then used to generate and assess answer candidates for a statistical QA system. Fleischman *et al.* (2003) focus on the precision of the information extracted using simple part-of-speech patterns. Jijkoun *et al.* (2004) have shown for English that using syntactic patterns based on dependency relations can lead to significant improvements in recall over systems based on regular expression pattern matching; below, we use the same technique in our parsing-based extraction. The other extraction technique we use, and which we call layout-based extraction, exploits the regular layout of a fixed corpus for our extraction, and is sometimes called information extraction from semi-structured text (Soderland 1999).

Question Analysis

The IMIX demonstrator will cater for *non-expert users* with information needs concerning repetitive strain injury (RSI), i.e., the kind of users that would probably consult medical encyclopedias. We want to avoid diagnosis questions. To better understand the requirements and the kind of answers that should be returned in this setting, we have manually analyzed a corpus of 435 questions on RSI which were collected by participants of the IMIX project (mostly computer scientists without a specific medical background). In our analysis we have examined two aspects of each question. The first was the expected answer format and the second was the question type; both are discussed below.

Answer Formats

For the expected answer formats we have made a distinction between four basic answer formats: named entity (NE), list, paragraph, and yes/no. Here are examples to questions requiring different types of answers (translated to English for this paper):

- *What does RSI mean?*
  (NE) repetitive strain injury
- *What causes RSI?*
  (list) RSI is caused by repetitive tasks, awkward postures, forceful movements and insufficient resting times.
- *What is a common misconception about RSI?*
  (paragraph) Many people believe that RSI is caused by working with a computer. Actually, there are many other activities that may cause RSI. RSI-related complaints have been frequently reported in professions like cleaners, welders, musicians and butchers.
- *Is RSI a fatal disease?*
  (yes/no) No.

We do not expect that users of a QA system will be satisfied with a basic yes/no answer. We aim at providing additional information to yes/no questions and therefore we will expect a paragraph answer for such questions.

Frequency	Relation
4303	treatment
3847	symptom
3548	definition
3195	cause
1137	diagnosis
993	occurs in
743	property
574	synonym
479	prevention
401	kind of
321	side effect
201	transferred by
116	similar to

Table 1: The 13 relation types in the annotated RSI corpus with their frequencies. The five most frequent types (*treatment*, *symptom*, *definition*, *cause* and *diagnosis*) have been used for question classification as well as the type *prevention* which occurred frequently in the questions.

Frequency	Q. Type	Frequency	A. Format
89	cause	157	list
73	prevention	148	yes/no
69	treatment	94	NE
68	symptom	36	paragraph
55	definition		
19	diagnosis		
62	other		

Table 2: The seven question types and the four answer formats of 435 RSI questions. The question types are more or less uniformly distributed with the exception of *diagnosis*, a question type which we do not deal with. Among the answer types, list and yes/no are more frequent than NE, which is most common in open domain QA.

Question Types

In order to find out what the useful question types are in this domain, we have examined an annotated corpus of Dutch RSI-related text. The corpus contains 32,248 sentences (about 500,000 tokens) and has been annotated at Tilburg University with concept and relation tags (Van den Bosch 2004). Frequent concept types in the corpus are *body part*, *disease* and *treatment*. The relation classes are the most interesting group for the question analysis. An overview of these classes can be found in Table 1.

We found 13 relation types in the corpus. The most frequent ones are *treatment*, *symptom*, *definition*, *cause* and *diagnosis*. These five types have been used for the question classification. Two other question types were added: *prevention* and *other*. The *prevention* type occurred frequently among the questions and the *other* category was created in order to be able to classify questions like *Is there a Dutch* *organization for RSI patients?*.

Classification of Questions and Answers

The 435 RSI questions have been manually classified with respect to question type and expected answer format; see Table 2. The question types are more or less uniformly distributed with the exception of the less frequent diagnosis type.[1] In the answer formats, list and yes/no are more frequent than the most common answer type in open domain QA: NE.

The lists of question types and answer formats presented in Table 2 indicate how the question answering task for this restricted domain can be performed. First, there is a limited number of question types (all except *other*) which makes up the majority of the questions (86%). We will focus our information extraction work on these six types. Second, unlike in many open domain question collections, only a minority of the questions (23%) requires a named entity as an answer. The most frequently requested answer format is the list format, and the combination of yes/no and paragraph answers also exceeds the number of required NE answers. For this particular domain, a successful information extraction approach needs to be capable of extracting answers of all four formats.

We did not find interesting relations between the question types and the expected answer formats. For three question types, the list format was the major required category (*diagnosis*, *prevention* and *treatment*). For three others, the yes/no format was most frequent (*cause*, *symptom* and *other*). Only the *definition* type had the NE answer format as the most frequently required category.

Layout-Based Information Extraction

The corpus that we use for information extraction in this section consists of a Dutch medical encyclopedia: the Medische Winkler Prins (about 0.3 million words). The layout of this corpus is very regular and this enables us to apply text-layout-based information extraction techniques to the corpus. For encyclopedic text, the most straight-forward application is extraction of definition information. A single rule can be used for this:

> An encyclopedia keyword is defined by the first sentence of its description.

With this rule we extracted 2870 definitions from the text. We have estimated the quality of the definitions by manually checking a random sample of 100 extracted definitions, all of which were found to be correct. The definition extraction rule works very well in combination with this corpus. However, this does not mean that it will perform equally well with all corpora: a test with the Dutch Wikipedia encyclopedia, which is not as rigidly structured as the Medische Winkler Prins, resulted in an estimated precision of 72%.

Many of the larger descriptions in the encyclopedia are divided into sections with appropriate headings. The section headings often contain words which are identical or related to the words used for the question classification described in the previous section, for example: the three most frequent headings are *Treatment*, *Symptoms* and *Cause*. We used this fact for defining a second structure-based information extraction rule:

> The relation between a encyclopedia keyword and a question type is defined by the first paragraph of the section with an appropriate heading in the description of the keyword.

We defined a total of 19 heading words related to the five remaining question types: *cause* (7 words), *symptom* (5), *prevention* (3), *treatment* (3), and *diagnosis* (1). With these words, 1848 patterns have been extracted from the corpus for the five remaining question types: *treatment* (768), *symptom* (659), *cause* (333), *prevention* (56) and *diagnosis* (32). We randomly selected 100 of the patterns and a manual evaluation of these showed that 99% were correct.

Information Extraction from Parse Trees

In parallel to the layout-based extraction strategy we are also developing an offline extraction method that uses syntactic patterns based on dependency relations. So far, we have applied the latter method to a 2.1 million word corpus of Dutch medical texts (1.6 million words from structured and edited reference material, including the previously mentioned Medische Winkler Prins, and 0.5 million words from web-sites related to RSI). To this end, the full corpus was parsed using Alpino, a robust wide coverage, dependency parser for Dutch (Bouma, Van Noord, & Malouf 2001; Malouf & Van Noord 2004). The resulting dependency parse trees were stored as XML. We defined patterns for three of the seven question types listed in the second section: *definition* (What is a CVA?), *symptom* (What are the symptoms of breast-cancer?), and *cause* (How does one develop RSI?).

As definitions, we extracted ⟨**Concept**,*Definition*⟩ tuples from sentences headed by the auxiliary verb *be*, where **Concept** was the subject of the sentence, and *Definition* a predicative complement. Sentences where either the concept or definition contained the words *cause* or *consequence*, or which contained modifiers at the sentence level, were excluded. Concepts were not restricted to names, but could be any nominal phrase. Examples of sentences matching the extraction pattern are '**cystic fibrosis** is *the most common fatal genetic disease*' and '**Amoebiasis** is *a type of gastroenteritis (gastro) caused by a tiny parasite*. We extracted 7,600 definition tuples. To extract ⟨**Effect**,*Symptom*⟩ tuples, we defined syntactic patterns for sentences containing specific phrases (*a sign of, an indication of, is recognizable, points to, manifests itself in, etc.*). Again, we stored both full phrases and the head nouns. 630 tuples were extracted. To extract ⟨**Effect**,*Cause*⟩ tuples, we used the patterns roughly equivalent to *causes, cause of, result of, arises, leads to*. We extracted 6,600 ⟨**Effect**,*Cause*⟩ tuples.

We randomly selected 150 extracted tuples (50 of each type), and evaluated them for correctness. Overall precision was 57%. The precision of the extracted definition tu-

[1] The fact that there are so few diagnosis questions is not a coincidence; the questions were generated by IMIX project members; IMIX does not deal with diagnosis questions, as stated previously.

ples was only 18%, whereas the precision of the cause and symptom tuples was 76% and 78%, respectively. A common source of errors is the fact that the extracted information is context-dependent (i.e., *these reactions cause chronic heart problems*). In addition, many of the extracted definition tuples either are of a more specific type (i.e., *A common treatment of psoriasis is light therapy* describes a treatment) or are simply not definitions (*Waiting lists are an important problem in health care*).

Evaluation

We have tested early versions of two question answering systems based on the strategies outlined above by supplying them with 50 randomly selected questions from our set of 435 RSI questions. The first system, Felix, uses tables built from Medische Winkler Prins encyclopedia with layout-based information extraction, with a backoff to paragraph retrieval based on the nouns in the question. The second system, Joost, employs the information tables extracted from parse trees. If the question type does not match with a table, or if no answer is found in the tables, a generic QA system (i.e., using Information Retrieval for finding relevant paragraphs and NLP for identifying and ranking potential answers) is used as fall-back system. Here is a translated sample of three questions and the top ranking answers produced by Felix and Joost:

Q48. Provide me with the causes of RSI.

Felix: The most important causes of RSI are high work pressure, stress and a long-lasting repetition of similar movements. In The Netherlands the work pressure increased with 1.5% in the eighties and the nineties, often caused by working time reduction policies which left less time for doing the same job.

Joost: Seven tests for detecting the cause of RSI in computer-related jobs.

Q27. Which disorders are counted as RSI?

Felix: A complex disorder like RSI fits the following complex description (click here for a clear explanation of this complex definition):

Joost: Specific RSI includes measurable disorders like tendenitis, epicondylitis, thoracic outlet syndrome (TOS), carpal tunnel syndrome, rotator cuff syndrome, tension neck syndrome, and De Quervain's disease.

Q20. Which phases can be identified in the development of RSI complaints?

Felix: (no answer)

Joost: In her opinion the often used three-phase segmentation (I, II and III) has not been clearly defined and there is no proven relation between the different phases and the expected RSI complaints.

The question-answer pairs produced by Felix and Joost were evaluated manually by two judges according to a three-point score: correct, wrong and incomplete. We only considered the highest ranked answer of each system. The results of this evaluation can be found in Table 3. Clearly, the two question answering systems do not perform well yet. Neither of them manages to find a reasonable answer for more than 15% of the questions. Felix finds more correct answers (11%) than Joost (4%) but the sum of the correct and incomplete answers of the two systems is in the same range (14% and 12%). Joost provides answers for all 50 questions but 88% of the answers are wrong. Felix missed 8 questions and thus a smaller fraction of its answers was wrong (70%). The two judges that performed the evaluation had an unweighted kappa agreement score of 0.6±0.1, which is reasonable.

	Felix	Joost
correct	0.11	0.04
incomplete	0.03	0.08
wrong	0.70	0.88
missed	0.16	0.00

Table 3: Performance of two question answering systems, Felix (semi-structured IE) and Joost (parsing-based IE), on 50 randomly selected RSI questions measured in fractions of correct/incomplete/wrong/missed answers based on a manual evaluation of two judges (kappa 0.6±0.1). Only the highest ranked answers were considered. point).

Error Analysis

There are three reasons for the non-optimal performance. First, the questions include many difficult ones which require careful question analysis, something which both systems currently lack. The question classification part of Felix only identifies the correct question type and main keyword for 25 of the 50 test questions. The type of the questions is determined by comparing them with a list of about 30 predefined phrases. Joost uses a question analysis module developed for factoid QA, and adapted minimally for the present task. It only classifies one question as a definition question, and none as cause or symptom questions, even though 13 questions in total belong to these three types.

The second problem for the two systems is the lack of coverage of the extracted tables. The question analysis part of Felix correctly identifies seven of the test questions as requiring an answer related to the treatment of RSI. However its tables do not contain such an answer. In fact the information tables of the system only generated answers to 6 of the 50 questions, a recall of 12% (the other answers were proposed by the fall-back strategy). The problem lies in the size of the system's training corpus: it is too small. For example, it does not contain any information on the treatment of RSI.

The third difficulty for the two systems lies with the inappropriateness of the employed fall-back strategy for those cases in which the tables fail to provide answers. Open domain QA can achieve impressive results by using techniques which basically spot constituents of the right type (i.e., a name of a person, organization, or location, a date, an amount phrase, etc.) within the context of keywords given in the question. For most of the questions in the medical domain, such an approach is not applicable. This means that identification of answers must rely solely on lexical and syntactic similarity between question and answer. In such a situation, ontological knowledge can be very useful. However, suitable Dutch resources for the medical domain are lacking.

The problem of limited available material will remain a challenge in this restricted non-English domain, especially for the techniques which rely on semi-structured text. More success can be expected of the parse-tree-based information extraction techniques which do not require formatted input and thus can be applied to text extracted from web sites. However, parsing techniques are typically applied to sentences and as we showed in the question analysis section, medical questions often require answers which are larger than a single sentence. We expect that a successful system in this domain will need to combine parsing techniques with topic detection methods for paragraphs, similar to the tiling work of Hearst (1997).

While examining the output of the two systems, we observed that they provide correct answers for different questions. In fact, for each question for which one system provided a correct or an incomplete answer according to a judge, the judge always rejected the answer of the other system; the three example questions and answers (Q20, Q27, Q48) illustrate this. This fact suggests an interesting method for improving the performance of the systems: by combining the answers. If we had an oracle which could make a perfect decision about which of a pair of answers is the best, a combined system would answer 15% of the questions correctly and provide a partial answer for an additional 11%.

Concluding remarks

We have compared two methods for extracting information to be used in an offline approach for answering questions a medical domain: layout-based and parsing-based. When evaluated in isolation both techniques obtain good precision scores. However, when we evaluated the extracted tables in a question answering environment, we found the overall performances of the systems was lower than could be expected from the table precision scores. The main problem was the lack of coverage of the extracted tables.

Offline information extraction is essential for achieving our goal of a real-time question answering system. The layout-based extraction approach has a limited application because it requires semi-structured data. The parsing approach does not share this constraint and therefore it is more promising. However, we have observed that the two techniques are complementary, and on its own the sentence-oriented nature of our parsing-based approach is insufficient in a domain in which the required answers are often larger than sentences. In our future work we will therefore combine these techniques with automatic paragraph classification techniques which may use but are not dependent on structural information like headings.

References

Ahn, D.; Jijkoun, V.; Müller, K.; De Rijke, M.; Schlobach, S.; and Mishne, G. 2005. Making stone soup: Evaluating a recall-oriented multi-stream question answering stream for Dutch. In *Multilingual Information Access for Text, Speech and Images: Results of the Fifth CLEF Evaluation Campaign*. Springer.

Bouma, G.; Van Noord, G.; and Malouf, R. 2001. Alpino: Wide-coverage computational analysis of Dutch. In *Computational Linguistics in The Netherlands 2000*. Amsterdam: Rodopi.

Fleischman, M.; Hovy, E.; and Echihabi, A. 2003. Offline strategies for online question answering: Answering questions before they are asked. In *Proc. 41st Annual Meeting of the Association for Computational Linguistics*, 1–7.

Hearst, M. 1997. TextTiling: Segmenting text into multi-paragraph subtopic passages. *Computational Linguistics* 23(1):33–64.

Jijkoun, V., and De Rijke, M. 2004. Answer selection in a multi-stream open domain question answering system. In *Proceedings 26th European Conference on Information Retrieval (ECIR'04)*, volume 2997 of *LNCS*, 99–111. Springer.

Jijkoun, V.; Mishne, G.; and De Rijke, M. 2004. How frogs built the Berlin Wall. In *Comparative Evaluation of Multilingual Information Access Systems, CLEF 2003*, volume 3237 of *Lecture Notes in Computer Science*, 523–534. Springer.

Jijkoun, V.; Mur, J.; and De Rijke, M. 2004. Information extraction for question answering: Improving recall through syntactic patterns. In *Coling 2004*, 1284–1290.

Malouf, R., and Van Noord, G. 2004. Wide coverage parsing with stochastic attribute value grammars. In *IJCNLP-04 Workshop Beyond Shallow Analyses - Formalisms and statistical modeling for deep analyses*.

NWO. 2004. *Interactieve Multimodale Informatie eXtractie (IMIX)*. URL: `http://www.nwo.nl/nwohome.nsf/pages/NWOP_5ZLCE8`.

Ravichandran, D., and Hovy, E. 2002. Learning surface text patterns for a question answering system. In *Proceedings of the 40th Annual Meeting of the ACL*.

Ravichandran, D.; Ittycheriah, A.; and Roukos, S. 2003. Automatic derivation of surface text patterns for a maximum entropy based question answering system. In *Proceedings of the HLT-NAACL Conference*.

Soderland, S. 1999. Learning information extraction rules for semi-structured and free text. *Machine Learning* 34(1):1–44.

Soubbotin, M., and Soubbotin, S. 2002. Use of patterns for detection of likely answer strings: A systematic approach. In *Proceedings of the TREC-2002 Conference*.

Van den Bosch, A. 2004. *Annotated Dutch RSI corpus for IMIX*. Tilburg University.

Voorhees, E. M. 2003. Overview of the trec 2003 question answering track. In *The Twelfth Text Retrieval Conference (TREC 2003)*. Gaithersburg, Maryland.

Zweigenbaum, P. 2003. Question answering in biomedicine. In *EACL 2003 Workshop on Natural Language Processing for Question Answering*. URL: `http://www.science.uva.nl/~mdr/NLP4QA/`.