Worlds Apart – Ontological Knowledge in Question Answering for Patients

Lina Henriksen, Anders Johannsen, Bart Jongejan, Bente Maegaard, Jürgen Wedekind
University of Copenhagen, Center for Language Technology
Njalsgade 140, 2300 Copenhagen S, Denmark
{linah, ajohannsen, bartj, bmaegaard, jwedekind}@hum.ku.dk

Abstract

We present ESICT, a hybrid question-answering system building on formalized knowledge from a medical ontology (SNOMED CT) as well as text-to-text generation in the form of document summarization and question generation. The independent subsystems are queried in parallel and compete for delivering the best answer. The use of an ontology gives the patient access to information typically not found in other sources, but also exposes a gap between everyday language and the specialized terms and conceptualizations of health professionals. In this paper we describe the ESICT system and discuss for each strategy how well it deals with this gap.

Keywords: question answering, nlidb, eHealth

1. Introduction

There is an emerging focus on the active citizen taking responsibility for his own health and illness through actively seeking and using information. But two people suffering from the same disease and seeking the same information might not ask questions in the same way and might not find the same answers equally useful. Citizens have different backgrounds in terms of education, social circumstances, new diagnoses vs. long-term chronic disease histories, etc.—issues that put very heavy demands on eHealth systems.

In Denmark a number of eHealth information systems are available on the Internet such as *netdoktor.dk* and *sundhed.dk*. However, they are to a large extent based on Frequently Asked Questions, they contain text chunks in a flat structure, and their purpose is to provide the citizen with an overview of predefined topics. Question-answering (QA) technology is another approach to provide citizens with quick and easy access to health and disease related information.

Existing QA systems have different strengths and limitations (cf., e.g., Athenikos and Han, 2010, for a survey). Those based on text-to-text generation are limited by the relevance and accuracy of the underlying text collections. Among the strengths of these systems are that the underlying text collection can be very large and the retrieved answer's wording and style will typically reflect the question and therefore be intelligible to the user. QA systems based on terminologies, ontologies, and other kinds of formalized knowledge often command clear, unambiguous, and terminologically correct information that may, however, be very different from the words used and known by the user. Besides, these systems require at least a shallow understanding of the user's question in order to provide a meaningful answer; recognition of a few keywords is not sufficient. The question *Can diabetes result in blindness?* is different from *Will diabetes result in blindness?* and *Can blindness result in diabetes?* and they require different answers. Further, such systems provide facts and not linguistically well-formed answers, and often the user's question is the best choice as a starting point for the wording of an answer.

We present ESICT (Experience oriented Sharing of health knowledge via Information and Communication Technology) (Andersen et al., 2012), a hybrid QA system employing highly structured as well as less structured information and offering information about diabetes mellitus in Danish. The coverage of ESICT is based on a corpus of 321 diabetes questions collected from three different sources: (i) an online diabetes discussion forum; (ii) an outpatient clinic at a Danish hospital (Wizard of Oz sessions); and (iii) a workshop with health informatics students. These real-life questions collected from patients, relatives, and other citizens reflect many complexities such as modality as in *Can diabetes be hereditary?* ambiguity as in *Is diabetes curable?* and personal issues as in *Can I get blood clots from diabetes?* or *Can I eat chocolate?* Most questions are within the topics: molecular and biomedical facts, epidemiology, interventions (e.g., behavioral intervention in terms of diet, exercise and other life style issues), and diagnostics.

ESICT applies three different approaches to question processing and answer generation. One approach relies on SNOMED CT, a multilingual clinical healthcare terminology covering terms of anatomy, findings, procedures, etc. in sub-type (is-a) hierarchies supported by defining relationships.[1] SNOMED CT is considered the world's most comprehensive nomenclature of clinical medicine. Examples of computer applications using SNOMED CT include electronic patient journal systems, clinical decision support systems, laboratory reporting systems, and many more.[2]

Because SNOMED CT, with its hierarchical design and primarily definitional knowledge, only covers some types of user questions, we also investigated two alternative QA strategies that could backup or, in cases of failure, even replace SNOMED CT-based querying. These are text-to-text generation approaches that both draw on authoritative medical texts within the diabetes domain. One approach uses query-focused multi-document summarization whereas the other generates potential users' questions on the basis of the particular document collection. These approaches all work in parallel in a running prototype. The following sections include information about the status, challenges, and

[1] http://www.ihtsdo.org/snomed-ct

[2] http://www.ncbi.lm.nih.gov/pmc/articles/PMC3704061

perspectives of each approach.

In this paper we will also discuss the pros and cons of each approach with respect to coherence between question and answer in terms of style, word selection, accuracy, etc. Another aspect which will be discussed is an inherent difficulty of medical QA systems: Not only do laymen phrase questions in different ways, as mentioned in the introduction, but laymen and health professionals have very different conceptual models of the world (cf., e.g., Zhang, 2010). In this paper we will try to discuss this problem for each of the approaches and evaluate their capacity for bridging the gap between the layman's and the doctor's conceptual worlds.

The rest of the paper is organized as follows. Sections 2–4 describe the three strategies of the ESICT QA system: ontology-based QA, multi-document summarization, and question generation. Section 5 presents the evaluation of the ESICT system. The last section highlights our key observations.

2. Strategy A: Ontology-based QA

The main focus of this approach was to build a natural language interface to SNOMED CT. We transform natural language questions into queries on the SNOMED CT ontology and produce natural language answers based on the results of these queries. Both the natural language questions and their answers are systematically related to SNOMED CT interpretable expressions, in the following called SNOMED expressions. SNOMED expressions are composed of atomic relational statements (triplets of the form $\mathit{concept}_1 - \mathit{relation} - \mathit{concept}_2$) and the description-logical operators supported by SNOMED CT. (In the actual implementation SNOMED expressions, together with the dependency analyses of the questions, are compiled into ontological scripts enabling SNOMED CT to infer the requested information.)

2.1. Workflow

The user's question in natural language is mapped onto a SNOMED expression by rewriting the dependency analysis of the question. The dependency analysis is created through a chain of natural language processing tools.[3] Rewriting is accomplished by a set of transformation rules[4] and results in a semantic analysis of the question. For example, the question *Får man katarakt af diabetes?* (Do you get cataract from diabetes?) is analyzed and transformed into a SNOMED expression as illustrated in Figure 1. The figure shows the relation between the decomposed syntactic analysis and the corresponding SNOMED CT terms. The area delineated with a red line is mapped onto the SNOMED CT relation DUE TO, while the blue delineated areas are mapped onto SNOMED CT concepts. In many questions the relation is expressed by a transitive main verb and the concepts are expressed by the subject and the object. However, there are also many other questions, such as predicative constructions (e.g., *Is diabetes type 2 dangerous?*) or questions with ditransitive verbs (such as in Figure 1 with non-pronominal subjects) that require a more sophisticated analysis.

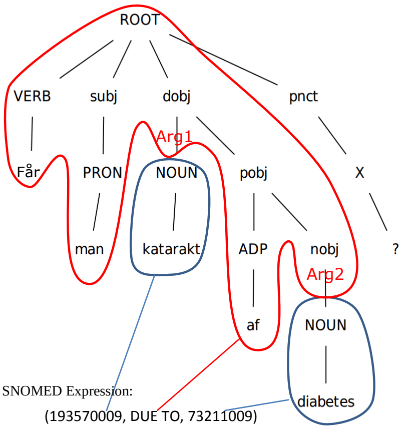


Figure 1: Dependency parse tree and semantic analysis of the question *Får man katarakt af diabetes?*

Each SNOMED CT relation accepts only arguments of certain types. In case of the DUE TO relation these types are Clinical Finding and Event for both the first and the second argument. These restrictions are context constraints on the application of the mapping rules and are used to disambiguate the semantic interpretation of the dependency structures.

The dependency parse tree is not only used as input for the semantic analysis but also as the raw material for the preparation of the answer. By replacing WH-phrases and pronouns, inserting function words, and moving constituents, the interrogative form of the user's utterance is transformed into a linguistically well-formed declarative answer.

2.2. Strengths and limitations

SNOMED CT has been developed using a variant of the relatively inexpressive description logic $\mathcal{EL}$. Thus it provides only very limited semantic expressivity and reasoning support. Modal, causal, and temporal reasoning, quantifier scoping, negations, comparatives, superlatives, etc. cannot be properly accounted for within an ontology like SNOMED CT. Therefore, for example, procedural how questions (e.g., *How do I prevent eye damage in diabetes?*), explanation questions (e.g., *Why does diabetes affect the liver?*), and advice-seeking questions (e.g., *How do I cope with diabetes?*) are generally out of scope of strategy A.

There are also many questions on topics not dealt with by the ontology. For instance, a diabetes patient might ask whether a particular food item is compatible with his illness and if not under what circumstances allowances may be made (e.g., *Can I eat pork rinds at the Christmas dinner?*). This is an out-of-topic question because diet recommendations are not included in SNOMED CT.

Our generation component is currently dependent on the input analysis. Thus we cannot generate answers that are structurally unrelated to the question (as, e.g., more elab-

[3] Consisting of a simple tokenizer, the OpenNLP tagger (http://opennlp.apache.org/), Bohnet's parser (http://code.google.com/p/mate-tools/), and CST's lemmatizer (Jongejan and Dalianis, 2009).

[4] Expressed as a script in a tree pattern matching language, Bracmat (https://github.com/BartJongejan/Bracmat).

orate answers to the question *Which treatments are available for diabetes?*). However, there are dependency-based stand-alone generators (e.g., Guo et al., 2011) that can, in combination with our mapping rules, be adapted to generate natural language expressions from SNOMED CT expressions without reference to the input analysis.

The presence of vagueness and ambiguity in ordinary language concepts presents an enormous challenge for disambiguating and interpreting layman's health-related questions in the SNOMED CT ontology. This is because there is usually a multi-to-one correspondence between natural language concepts and the medical concepts of SNOMED CT, and vice versa. This problem is exacerbated by SNOMED CT's relational sparseness (there is only a rather small number of relation types). Thus, there are usually numerous predicates that map to the same relation, and there are also many predicates that map onto different relations, depending on their arguments. This rather loose correspondence between SNOMED CT's conceptual model and ordinary language terms was not only a major bottleneck in processing questions, it presented also a particular challenge for generation where the mapping rules were reversed.

In our prototype system, approach A provides correct answers to 38% of the questions in the corpus. It performs best with simple *What is* and Yes/No questions that can be answered with a simple definition or *Yes* or *No* followed by the reordered input question, and it provided for this type of questions the most reliable answers.

3. Strategy B: Summarization

In our collection of questions on diabetes, we observed that patients often need to know something that cannot be answered by querying the SNOMED CT ontology, no matter how cleverly this is done. Multi-document summarization provides a robust, well-established way to address this problem (Demner-Fushman and Lin, 2006; Lee et al., 2006; Niu et al., 2006). It is a text-to-text generation method that answers questions by finding and manipulating text from documents in a reference corpus. The answer generation happens in a two-step process in which first the most relevant documents (with respect to the question) are found using information retrieval techniques. The top-ranking documents then become input to a summarization component responsible for compiling the final answer. The information retrieval component ranks documents based on the cosine similarity of bag-of-words vectors with tf-idf weighting. For the summarizer we use an implementation of the unsupervised, graph-based LexRank algorithm (Erkan and Radev, 2004) coupled with Maximum Marginal Relevancy (Carbonell and Goldstein, 1998) to ensure information diversity in the answer.

A summary of multiple documents is of course unlikely to be an exact answer to the patient's question. Indeed, for many specific questions, the answer is unlikely to be found at all in the reference collection. In these cases we consider the goal of this approach to be to deliver information pertinent to the question which hopefully allows the patient to infer the answer to his question.

The basic unit is a document in the information retrieval step and a paragraph of text in the subsequent summarization step. This choice implies that the answer cannot be shorter than a paragraph, making it considerably longer than answers obtained from SNOMED CT and question generation (Section 4), which are always a single sentence. Apart from this lower limit, the length of the answer is an adjustable parameter and in the prototype it has been set to a maximum of 100 words. The limit is optimized for questions that solicit advice or ask for explanations as human judges overwhelmingly prefer long answers for these types of questions (Kaisser et al., 2008).

Below we describe how we collected the reference corpus and our strategy for dealing with out-of-vocabulary words in the question.

3.1. Reference corpus

The reference corpus is compiled from various publicly available web sources. It collects the contents of 125 web pages, which have been manually curated and linked to individual questions in the question corpus.

For each document we automatically removed boiler-plate text (e.g., menus and copyright notices) and identified headlines and text content via heuristic rules,[5] resulting in a plain-text file. All downloaded pages were further segmented by headlines so that a section of the text below a headline (and before the next one) would be treated as a separate, more specific, document. Counted this way the total number of documents in the corpus is 556.

From a user's perspective, the reference corpus has a broader coverage of topics (e.g., diet) and provides richer information (e.g., how to check one's blood sugar) than SNOMED CT. Therefore, many questions that fall outside the scope of SNOMED CT have answers in the reference corpus. The reference corpus is well-suited for manner (*how*) and advice questions, because answers may convey useful information that cannot easily be formalized or is not universally true, e.g., practical advice and rules-of-thumb.

3.2. Vocabulary expansion

Although the reference corpus covers a broad range of topics, it does not have a specific answer for each and every question. For instance, many questions are on the topic of food and ask specifically about the feasibility of different dietary choices.

However, *some* of these specific questions have actually been asked before and answered before and so the corpus could easily have an answer for a related question, even though it lacks the answer for the question itself. In this case we say the question is *out of vocabulary* but not *out of topic*. We address the lack of recall due to out-of-vocabulary questions by performing vocabulary expansion at query time.

To motivate this, consider the two related questions in (1) and (2).

(1) Kan jeg spise leverpostej? (Can I eat liver paste?)

(2) Kan jeg spise rullepølse? (Can I eat "rolled meat" sausage?)

[5] https://pypi.python.org/pypi/jusText

Example (3) answers (1) and (2) equally well, but only the topic word of (1) *leverpostej* occurs in the text. Here, expanding the topic word of (2) *rullepølse* to *leverpostej* would allow us to consider the answer for that question as well.

(3) Spis mindre af fede kødprodukter som pølser, bacon, salami, leverpostej og frikadeller. (Eat less of fatty meat products like sausages, bacon, salami, liver paste and rissole).

We retrieve the word expansions by selecting the most similar words in a word embedding space (Mikolov et al., 2013), with a manually tuned threshold for similarity. The embeddings were learned on a corpus consisting of a general language corpus[6] and a specialized diabetes corpus. We obtained the specialized corpus by submitting all questions in our collection as queries to a search engine[7] and for each retrieve all documents in the top 50 matches. Using either of the corpora alone resulted in expansions of a much lower quality than when they were combined.

3.3. Strengths and limitations

The summarization approach as implemented here is fully unsupervised and requires no deep semantic processing, which is an advantage since only limited language resources exist for Danish and performance of state-of-the-art tools for, e.g., parsing and named entity recognition is well below that of comparable tools for English. The use of information retrieval techniques enables the summarization approach to recover relevant answer text from the reference corpus with high recall, and this is further improved by the use of a query expansion component, such as that described above. Unfortunately, the high recall comes at the expense of a lower precision, since the lack of semantic processing may result in, e.g., conflation of homographs, causing too much information to be retrieved. Also, for questions where brief and concise answers are appropriate, the summary will often be too verbose, because the answer length is set to a fixed value for all questions.

In patient question answering the overriding concern is patient safety: Providing an answer which is wrong is much worse than not providing an answer at all. From this perspective the trade-off between recall and precision made by summarization might seem unfortunate. There is, however, a limited degree of error detection built into a summary in that the summary usually provides enough context to allow the user to learn whether the text is in fact addressing the question. This is in contrast to many systems that use sophisticated semantic analysis but where errors in the analysis may go unnoticed by the user.

Further limitations of the summarization approach to QA arise from the need for an up-to-date corpus of verified documents and the fact that summarization, in contrast to knowledge-based approaches, cannot infer new answers on its own. It can, however, as we have seen, provide answers to related questions in case the correct answer cannot be found in the corpus.

In the prototype implementation, the summarization approach provided an answer to 248 out of 321 questions. Of these answers 30% were judged to be correct (covering 24% of the whole question corpus). Compared to the other other strategies, summarization performed best on complex questions involving multiple entitites and relations. However, due to the fixed answer size, the answers were generally too long, mixing correct answer text with peripheral or irrelevant content.

4. Strategy C: Question generation

Question generation (QG) (the task of automatically creating questions from various sources of inputs: texts, databases, etc.) was originally used for educational assessment and intelligent tutoring (cf., e.g., Heilman and Smith, 2010). More recently, it has also been exploited to improve closed-domain QA for languages with scarce resources. Since QG is a relatively new technique in QA, there exist at the moment only very few systems that take advantage of QG (e.g., Bernhard et al., 2012).

The basic idea of the QG approach is to identify sentences of informative documents that can serve as answers to potential questions and transform these sentences into their interrogative forms. All question/answer pairs thus extracted are stored in a database. For question answering, a user's question is identified in the question-answer database, and the corresponding answer is returned.

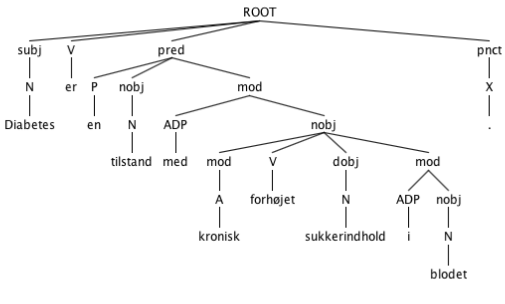
The minimum core resources that are required to produce an operable system are a collection of reliable (authoritative), informative documents, a grammar to parse the documents, and a set of transformation rules that generate questions from the syntactic parse of useful answer sentences. For our prototype we use documents from *Medicinhåndbogen* and *sundhed.dk*, a projective Danish dependency parser (Søgaard and Rishøj, 2010; McDonald and Pereira, 2006), and a set of manually created syntactic transformation rules. To execute the transformation rules, we use Tregex (a tree query language) and Tsurgeon (a tool for modifying trees) (Levy and Andrew, 2006). These tools match the left-hand sides of the transformation rules to syntactic analyses of the documents, apply the syntactic transformations to the document analyses, and output the resulting syntactic descriptions.

As a simple illustration, consider the Danish sentence (4).

(4) Diabetes er en tilstand med kronisk forhøjet sukkerindhold i blodet. (Diabetes is a condition with a chronically elevated level of sugar in the blood.)

and its dependency parse tree in (5).

(5)



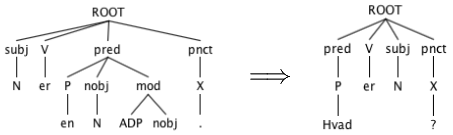
This sentence provides a useful answer to the question

[6] Korpus2000: http://ordnet.dk/korpusdk

[7] Microsoft Bing API

Hvad er diabetes? (What is diabetes?). The dependency tree for the question is then produced by running the Tregex and Tsurgeon scripts for the transformation rule in (6).

(6)



4.1. Refinements and optimizations

There are a number of refinements and optimizations to the basic setup that can improve the overall performance of the QG module. However, only some of them have been implemented (to some extent). This is because the required resources and tools were either not available for Danish or not adaptable within the project funding and timeframe.

Automatic recognition of potential question topic terms and question foci The term *question topic* is usually used to refer to the object or event someone intends to increase his/her knowledge about by using the question (like *diabetes*, *hypoglycemia*, etc.), while the *question focus* is the term/phrase indicating the semantic type of which the answer is an instance (e.g., *country* in *Which country is most densely populated?*). In order to identify potentially useful answer sentences in a collection of documents, it is therefore extremely beneficial to tag all topic terms and to annotate all named entities with their focus category. This can be accomplished with a named entity recognizer that supports the required classification (cf., e.g., Prager et al., 2006). However, since such an advanced named entity recognizer is not available for Danish, we relied for the prototype on an existing list of on-topic terms (from the Steno Diabetes Center) and a manually created list of focus terms/classes (like *disease*, *condition*, etc.).

Paraphrase generation In several QA systems paraphrasing has been used to better cope with the linguistic variability of questions and answers. Paraphrasing increases the likelihood of finding an answer if the user's question is meaning equivalent to a question in the database but varies from it in linguistic form. Because of time constraints, we focused first on simple paraphrases (taken from Microsoft Word's thesaurus) that could safely be produced by global substitution without much regard for the surrounding context. However, for the generation of more complicated context-sensitive and phrasal paraphrases, several promising approaches extract paraphrases automatically from the Web (cf., e.g., Duclaye et al., 2003) or exploit phrase tables from existing machine translation systems (cf., e.g., Kuhn et al., 2010).

Answer selection To further improve recall, a very rudimentary fuzzy matching procedure has been incorporated. This, however, leaves room for improvements. A lemmatizer, for example, can help in identifying the answer in cases where a user's question differs only morphologically from a question in the database, as in *Hvad er symptomer på diabetes?* and *Hvad er symptomerne på diabetes?* (What are (the) symptoms of diabetes?). There are, of course, many other algorithms that can be used to compute the best matching database question, ranging from relatively simple algorithms that compute the edit distance between two strings (e.g., Levenshtein distance computing algorithm) to more sophisticated syntactic similarity measures (cf., e.g., Croce et al., 2011).

Often there will be more than one answer to a question. Indeed, if the document collection is rich enough, there may easily be ten or twenty options for answering common queries such as *What is diabetes?* Even though we did not rank the answers, there are a number of algorithms for selecting the "best" answer, among them ranking algorithms that exploit syntactic and semantic features to account for both linguistic quality and information content at the same time (cf., e.g., Athenikos and Han, 2010).

Co-reference chain detection If sentences following an identified answer sentence further elaborate that sentence, these sentences are often linked through a co-reference chain. Consider, for instance, the following English translation of an extract from a document:

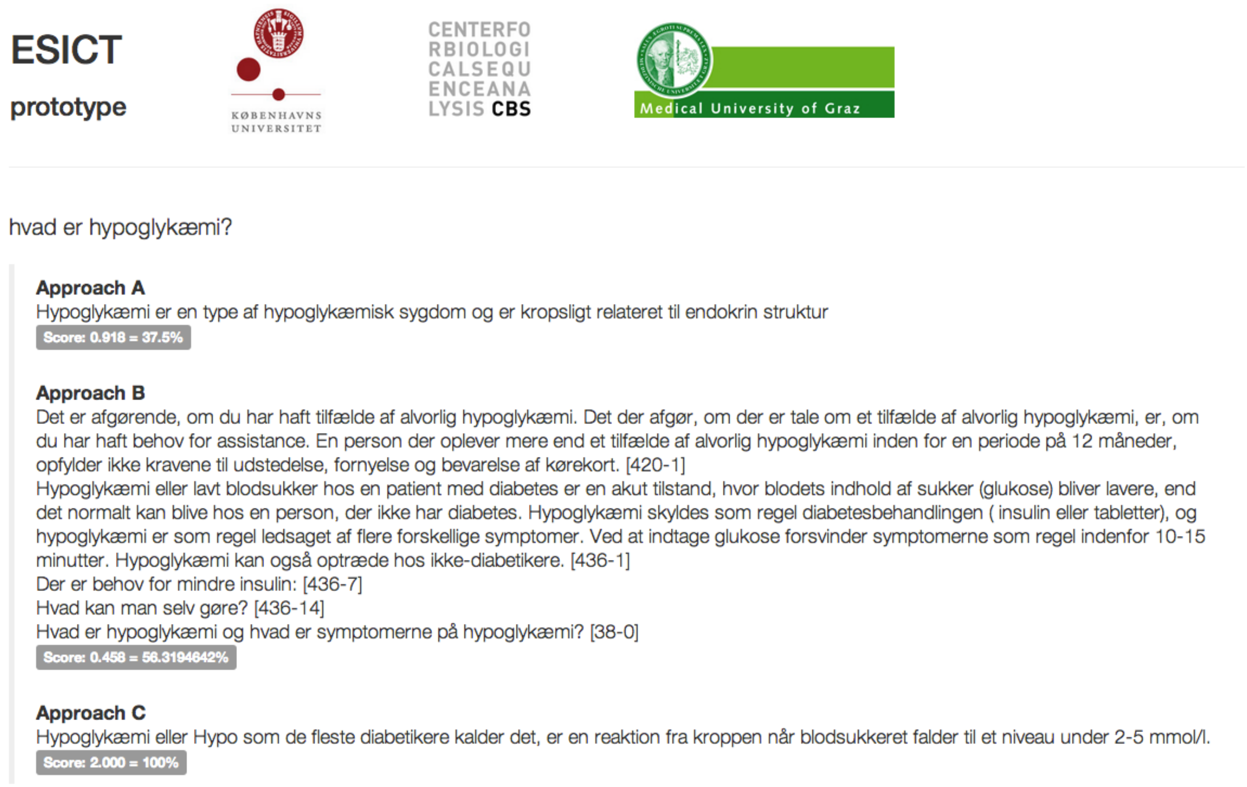
(7) In the pancreas *hormones*$_1$ are produced, which control, among others, the body's sugar balance. *These*$_1$ are glucagon and *insulin*$_2$. The *insulin*$_2$ enables sugar in the blood to enter more easily, among others, muscle and liver cells.

Here, QG produces from the first sentence the question *What is produced in the pancreas?* Co-reference chain detection can then be used to identify the whole extract as a more detailed answer than the one provided by only the first sentence. Unfortunately, there are no practically applicable co-reference resolution systems specifically designed for Danish; there are a few language independent systems (cf., e.g., Zhekova and Kübler, 2010; Zhekova et al., 2012) that have been demonstrated to work effectively for several different languages. However, because of the scope of the project, it was not possible to adapt the language dependent modules to the given language data.

Ordering for multiple sentence answers If there are many options for answering common queries such as *What is diabetes?* often several of them can be combined to a more comprehensive answer. Consider, for example, the translations of two Danish sentences that generate *What is diabetes?* (in Danish).

(8) Diabetes is a condition with chronically high level of sugar in the blood. Diabetes is a serious disease because it causes a significant risk for developing complications of the eyes, kidneys, nerves, and blood vessels.

In this case a combination is desirable. However, nothing prevents two sentences mapping to the same question from being uninformative variants of each other. To ensure that new information is actually added, techniques from automatic summarization can be used. To maximize the information content of combined answers, we can, for instance, select the answer that, for a fixed size, has the largest possible number of different concepts (e.g., bigrams) (Gillick and Favre, 2009).


ESICT
prototype
KØBENHAVNS UNIVERSITET
CENTERFORBIOLOGICALSEQUENCEANALYSIS CBS
Medical University of Graz

hvad er hypoglykæmi?

Approach A
Hypoglykæmi er en type af hypoglykæmisk sygdom og er kropsligt relateret til endokrin struktur
Score: 0.918 = 37.5%

Approach B
Det er afgørende, om du har haft tilfælde af alvorlig hypoglykæmi. Det der afgør, om der er tale om et tilfælde af alvorlig hypoglykæmi, er, om du har haft behov for assistance. En person der oplever mere end et tilfælde af alvorlig hypoglykæmi inden for en periode på 12 måneder, opfylder ikke kravene til udstedelse, fornyelse og bevarelse af kørekort. [420-1]
Hypoglykæmi eller lavt blodsukker hos en patient med diabetes er en akut tilstand, hvor blodets indhold af sukker (glukose) bliver lavere, end det normalt kan blive hos en person, der ikke har diabetes. Hypoglykæmi skyldes som regel diabetesbehandlingen (insulin eller tabletter), og hypoglykæmi er som regel ledsaget af flere forskellige symptomer. Ved at indtage glukose forsvinder symptomerne som regel indenfor 10-15 minutter. Hypoglykæmi kan også optræde hos ikke-diabetikere. [436-1]
Der er behov for mindre insulin: [436-7]
Hvad kan man selv gøre? [436-14]
Hvad er hypoglykæmi og hvad er symptomerne på hypoglykæmi? [38-0]
Score: 0.458 = 56.3194642%

Approach C
Hypoglykæmi eller Hypo som de fleste diabetikere kalder det, er en reaktion fra kroppen når blodsukkeret falder til et niveau under 2-5 mmol/l.
Score: 2.000 = 100%

Figure 2: Screenshot of the system's output for the question *Hvad er hypoglykæmi?* (What is hypoglycemia?).

4.2. Strengths and limitations

Even without most of the refinements mentioned above the QG approach is able to produce a variety of more complex interrogative questions, including causative questions, that are assumed to be inherently more difficult than other interrogative questions, like, for example, factoid and definitional questions. English translations of a few typical examples are given in (9a–f).

(9) a. *What are the symptoms of diabetes?* The classic symptoms of untreated diabetes are weight loss, increased urination (polyuria), and increased sense of thirst (polydipsia) and hunger (polyphagia).

b. *Is diabetes contagious?* Diabetes is not contagious.

c. *Can diabetes be cured?* Diabetes cannot be cured, but there are now drugs that can increase insulin production and increase the cells' sensitivity to insulin.

d. *What is the diabetes treatment aiming at?* The diabetes treatment aims at bringing the level of blood glucose as near as possible to normal to eliminate the symptoms of high blood glucose and to prevent the development of complications.

e. *What causes diabetes?* Diabetes is caused by defective insulin secretion, reduced insulin action or a combination of these factors.

f. *When is a person diagnosed with diabetes?* A person is diagnosed with diabetes when the glucose levels are not normally controlled and the concentration is too high.

However, the scope of QG is limited to encyclopedic one sentence questions. In some cases it is possible to produce multi-clausal questions such as *What happens if the blood sugar is too low?* but these typically make up only a small portion of the generated questions.

We created altogether 45 transformation rules (the number could have been slightly reduced by fully exploiting the notational devices that Tregex and Tsurgeon provide). These generated from a small document collection (altogether 750 sentences) a total of 148 questions. Surprisingly, only 16% of them were contained in the question corpus (7.5% of the entire corpus), although all generated questions are perfectly reasonable to ask. Thus for real-life health-related questions posed by lay people (Kilicoglu et al., 2013), QG seems to be more like a back-up strategy when other strategies, for whatever reason, fail to provide meaningful answers to Wikipedia-related questions.

5. Evaluation

The three approaches are embedded in a prototype system which runs them in parallel and provides results for all of them. Figure 2 shows a screenshot of the system's output for the question *Hvad er hypoglykæmi?* (What is hypoglycemia?). The output shows (i) the suggested answer from each approach, (ii) the confidence scores, and (iii) the comparable scores in percentages. In this example approach C provides the answer that is given to the user (because it received the highest score). The English translations are as follows (less relevant information provided by approach B is displayed in gray):

(A) Hypoglycemia is a type of hypoglycemic disease and is physically related to the endocrine structure,

(B) It is essential whether you have had severe hypoglycemia. Hypoglycemia is assessed as severe hypoglycemia if you have needed assistance. A person who experiences more than one case of severe hypoglycemia within a 12-month period, does not meet the requirements for the issuance, renewal, and maintenance of a driver's license. [420-1] Hypoglycemia or low blood sugar in a diabetes patient is an acute condition where the level of blood sugar (glucose) is lower than normally observed in people without diabetes. Hypoglycemia is usually caused by diabetes treatment (insulin or tablets), and hypoglycemia is usually accompanied by a variety of symptoms. By consuming glucose, symptoms usually disappear within 10–15 minutes. Hypoglycemia can also occur in non-diabetics. [436-1] You need less insulin: [436-7] What can you do yourself? [436-14] What is hypoglycemia and what are the symptoms of hypoglycemia? [38-0],

(C) Hypoglycemia, or hypo as most diabetics call it, is a physical reaction when the blood sugar drops to a level below 2–5 mmol/l.

The approaches were evaluated on our corpus of 321 questions. 306 questions got an answer from at least one approach, and at least one of the approaches answered correctly in 45% of these cases (43% of the total corpus). This quality is too low for a running system. However, as Danish is a less-resourced language with much fewer medical knowledge resources than English, the prototype is a step toward the implementation of a medical QA system for the Danish citizens.

6. Consequences and observations

The goal of our strategy A (Section 2) was to explore to what degree it is possible to build a natural language interface to the SNOMED CT ontology. This required a mapping between layman's everyday language and the terminology of professionals. However, this mapping is not first and foremost a technical challenge, but rather a conceptual challenge, because professionals and patients understand and conceptualize medicine in fundamentally different ways.[8] Because of the divergent conceptualizations of the world and SNOMED CT's relational sparseness, it has been an extremely difficult task to disambiguate and interpret laymen's questions in the SNOMED CT ontology. Moreover, a considerable amount of user questions could not be processed since they simply could not be adequately interpreted in the lightweight description logic SNOMED CT is based on.

The problem of differences in terminology is less severe in the text-to-text generation strategies B (summarization) and C (question generation), for two reasons. First, answers are located by matching words in the question with words in the documents, ensuring a common vocabulary. Second, the documents in our collection are written with a layman audience in mind, addressing the concerns of patients and not professionals. But this issue does also affect these systems and may become more pressing when we move beyond small curated document collections and use text written for a broader range of audiences, including health professionals. However, even based on our rather small document collections, these approaches revealed serious limitations: summarization suffers from precision issues and question generation from recall issues.

Moreover, by considering the questions of our corpus of real-life layman's questions that our strategies were not able to adequately deal with, it became quite obvious that their complexity is way beyond currently known QA technology.

7. Acknowledgments

The ESICT project, running from 2010 – 2013, was funded by the Danish Strategic Research Council. We would like to thank our project partners: Ulrich Andersen, Anna Braasch, Søren Brunak, Anders Jensen, Lars Kayser, Ole Norgaard, and Stefan Schulz for contributing important information to this paper.

8. References

Andersen, U., Braasch, A., Henriksen, L., Huszka, C., Johannsen, A., Kayser, L., Maegaard, B., Norgaard, O., Schulz, S., and Wedekind, J. (2012). Creation and use of language resources in a question-answering eHealth system. In *Proceedings of the 8th International Conference on Language Resources and Evaluation*, pages 2536–2542, Istanbul.

Athenikos, S. J. and Han, H. (2010). Biomedical question answering: A survey. *Computer Methods and Programs in Biomedicine*, 99(1):1–24.

Bernhard, D., de Viron, L., Moriceau, V., and Tannier, X. (2012). Question generation for French: Collating parsers and paraphrasing questions. *Dialogue and Discourse*, 3(2):43–74.

Carbonell, J. and Goldstein, J. (1998). The use of MMR, diversity-based reranking for reordering documents and producing summaries. In *Proceedings of the 21st Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*, pages 335–336, Melbourne.

Croce, D., Moschitti, A., and Basili, R. (2011). Semantic convolution kernels over dependency trees: Smoothed partial tree kernel. In *Proceedings of the 20th ACM International Conference on Information and Knowledge Management*, pages 2013–2016, New York, NY.

Demner-Fushman, D. and Lin, J. (2006). Answer extraction, semantic clustering, and extractive summarization for clinical question answering. In *Proceedings of the 21st International Conference on Computational Linguistics and the 44th Annual meeting of the Association for Computational Linguistics*, pages 841–848, Sydney.

Duclaye, F., Yvon, F., and Collin, O. (2003). Learning paraphrases to improve a question-answering system. In *Proceedings of the 11th Conference of the European Chapter of the Association for Computational Linguistics, Workshop in NLP for QA*, pages 35–41, Budapest.

[8] A point also raised by Udo Hahn during a panel discussion at Medinfo 2013.

Erkan, G. and Radev, D. R. (2004). LexRank: Graph-based lexical centrality as salience in text summarization. *Journal of Artificial Intelligence Research*, 22(1):457–479.

Gillick, D. and Favre, B. (2009). A scalable global model for summarization. In *Proceedings of the NAACL HLT Workshop on Integer Linear Programming for Natural Language Processing*, pages 10–18, Boulder, CO.

Guo, Y., Wang, H., and van Genabith, J. (2011). Dependency-based n-gram models for general purpose sentence realisation. *Natural Language Engineering*, 17(4):455–483.

Heilman, M. and Smith, N. A. (2010). Good question? Statistical ranking for question generation. In *Human Language Technologies: The 2010 Annual Conference of the North American Chapter of the Association for Computational Linguistics, Proceedings*, pages 609–617, Los Angeles, CA.

Jongejan, B. and Dalianis, H. (2009). Automatic training of lemmatization rules that handle morphological changes in pre-, in- and suffixes alike. In *Proceedings of the Joint Conference of the 47th Annual Meeting of the ACL and the 4th International Joint Conference on Natural Language Processing of the AFNLP*, pages 145–153, Suntec.

Kaisser, M., Hearst, M., and Lowe, J. (2008). Improving search results quality by customizing summary lengths. In *Proceedings of the 46th Annual Meeting of the Association for Computational Linguistics*, pages 701–709, Columbus, OH.

Kilicoglu, H., Fiszman, M., and Demner-Fushman, D. (2013). Interpreting consumer health questions: The role of anaphora and ellipsis. In *Proceedings of the 2013 Workshop on Biomedical Natural Language Processing*, pages 54–62, Sofia.

Kuhn, R., Chen, B., Foster, G., and Stratford, E. (2010). Phrase clustering for smoothing TM probabilities – or, how to extract paraphrases from phrase tables. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 608–616, Beijing.

Lee, M., Cimino, J., Zhu, H., Sable, C., Shanker, V., Ely, J., and Yu, H. (2006). Beyond information retrieval – Medical question answering. *AMIA Annual Symposium Proceedings*, 2006:469–473.

Levy, R. and Andrew, G. (2006). Tregex and Tsurgeon: Tools for querying and manipulating tree data structures. In *Proceedings of the 5th International Conference on Language Resources and Evaluation*, pages 2231–2234, Genoa.

McDonald, R. and Pereira, F. (2006). Online learning of approximate dependency parsing algorithms. In *Proceedings of 11th Conference of the European Chapter of the Association for Computational Linguistics*, pages 81–88, Trento.

Mikolov, T., Chen, K., Corrado, G., and Dean, J. (2013). Efficient estimation of word representations in vector space. *ArXiv e-prints:1301.3781*, pages 1–12.

Niu, Y., Zhu, X., and Hirst, G. (2006). Using outcome polarity in sentence extraction for medical question-answering. *AMIA Annual Symposium Proceedings*, 2006:599–603.

Prager, J. M., Chu-Carroll, J., Brown, E. W., and Czuba, K. (2006). Question answering by predictive annotation. In Strzalkowski, T. and Harabagiu, S., editors, *Advances in Open Domain Question Answering*, pages 307–347. Springer, Dordrecht.

Søgaard, A. and Rishøj, C. (2010). Semi-supervised dependency parsing using generalized tri-training. In *Proceedings of the 23rd International Conference on Computational Linguistics*, pages 1065–1073, Beijing.

Zhang, Y. (2010). Contextualizing consumer health information searching: An analysis of questions in a social Q & A community. In *Proceedings of the 1st ACM International Health Informatics Symposium*, pages 210–219, Arlington, VA.

Zhekova, D. and Kübler, S. (2010). A language-independent system for coreference resolution. In *Proceedings of the 5th International Workshop on Semantic Evaluation*, pages 96–99, Uppsala.