Automatic Question Generation from Swedish Documents as a Tool for Information Extraction

Kenneth Wilhelmsson
Swedish School of Library and Information Science
University of Borås
kenneth.wilhelmsson@hb.se

Abstract

An implementation of automatic question generation (QG) from raw Swedish text is presented. QG is here chosen as an alternative to natural query systems where any query can be posed and no indication is given of whether the current text database includes the information sought for. The program builds on parsing with grammatical functions from which corresponding questions are generated and it incorporates the article database of Swedish Wikipedia. The pilot system is meant to work with a text shown in the GUI and autocompletes user input to help find available questions. The act of question generation is here described together with early test results regarding the current produced questions.

1 Introduction

Question generation has been the focus of several recent international workshops where the field has been defined as including sub-fields like tutorial dialogue and FAQ generation. In this paper, the focus is on the *Text-to-Question* subtask. Rus and Graesser (2009) define the task as follows: “given a text, the goal of a QG system performing the Text-to-Question Question Generation task would be to exhaustively create a set of Text-Question pairs, such that every possible question that could be generated would be included in the set”, see Table 1.

The formulation thus includes the notion of *all* possible questions to which a text can be said to provide answers. This can for example mean all the questions from the explicit propositions but also facts deduced using various algorithms for inference, anaphora resolution etc. This is a complicating factor as this set is hard to estimate and will make it impossible to compute the relative coverage of the set of questions produced.

It is not clear what counts as one unique question, and whether producing various formulations of the same question is advantageous. In a practical user scenario, there can be benefits to generating variations of the same question (using e.g. substitution of synonymous words) to help the user find at least one way of expressing the query in a large question set produced by a natural query system.

Given:

- Text ***T***

Create:

- Text-Questions pairs $\{P_1 \ldots P_n\}$ each represented as a (K_i , Q_i) pair, where K_i , the target text, indicates which text segment from ***T*** represents the answer and the Q_i represents a question that would elicit K_i

Table 1: The Text-to-Question task as characterized by Rus and Graesser (2009)

The situation described in this paper is the use of a natural language query system which explicitly generates a set of questions per text as an alternative to the functionality of several systems which permits a user to pose queries in question form freely, but which never guarantee that these are answered by the current database. If the system uses a black-box algorithm for finding the answers and/or uses a database that is unknown or vast (like the entire Internet), this can be particularly striking. An example is *PowerSet* (Converse et al 2008) which will rank text segments of all of English *Wikipedia* using a collection of different techniques, when a question is formulated in natural language. The proposed answers (text segments) will be presented to the user according to best match ranking given the question. That approach, like that of Harabagiu et al (2000), mixes the task of information retrieval (search for documents) with that of information extraction. From the user perspective, it may be unknown whether a (formulation of a)

Bolette Sandford Pedersen, Gunta Nešpore and Inguna Skadiņa (Eds.)
NODALIDA 2011 Conference Proceedings, pp. 323–326

question is in fact answered at all by the database in these types of systems where any question string can be formulated.

This paper deals with an implementation of automated question generation from raw text in Swedish. The focus here is on the actual question generation task by syntactic means, the user interface and some preliminary tests of the current state of the implementation. The system incorporates the Swedish *Wikipedia* article data-base and generates questions for one text article, or other input text, at a time. This means that the current text subject (the available information) is somewhat known to the user. In fact, the text source is visible in the user interface, shown in Figure 1, and the questions produced will mark and scroll the corresponding answer into view when a question is selected.

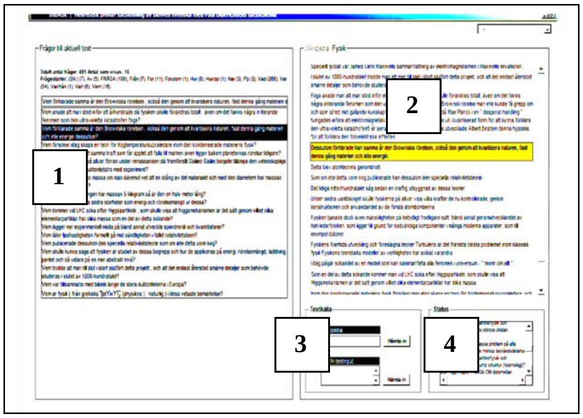


Figure 1: The GUI of the program

1) Autocompleting input form for choice of question
2) The text source in which the suggested questions will mark and scroll to the corresponding section with answers
3) Forms for choice of *Wikipedia* article or arbitrary text input
4) Statusbox displaying various information during a run

2 The Initial Steps: Text Pre-processing and Syntactic Parsing

The first steps of the text-to-question task includes sentence splitting, tokenization, POS tagging and syntactic parsing with mark-up of grammatical functions on the main clause level. In the process, the text is tagged, whereafter it is parsed and questions are finally extracted.

A trigram-based Hidden Markov Model POS tagger is used to provide input for the syntactic parsing. The parsing of Swedish free text is carried out using a heuristic algorithm based on the sentence schema for Nordic languages, originally introduced for Danish by Diderichsen (1946). The parser, which is described by Wilhelmsson (2010), makes use of the sentence schema by avoiding identification of multi-word constituents (unbounded constituents) by explicit matching, resulting in a format shown in Example 1.

```
<subjekt>Ni som frågar</subjekt>
<pfv>hade</pfv>
<adverbial>nog</adverbial>
<adverbial>ändå</adverbial>
<piv>kunnat</piv>
<piv>köpa</piv>
<objekt>en vän</objekt>
<objekt>en present</objekt>
<tom>.</tom>
```

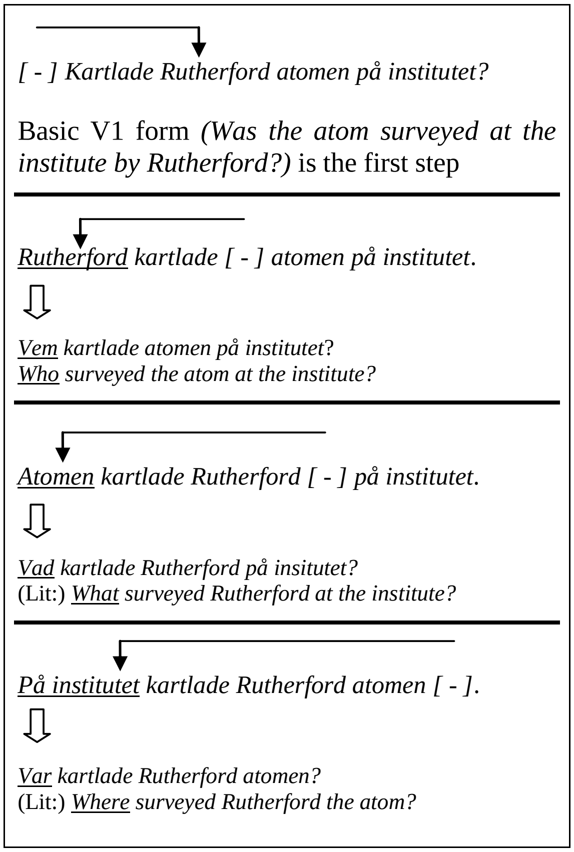
Example 1: The XML output format from the parser for the Swedish sentence '*You, who ask, would anyway probably have been able to buy a friend a present*' includes labels *pfv* (primary finite verb), *piv* (primary non-finite verb) and *tom* (empty).

3 Swedish Question Generation from Parses with Grammatical Functions

The question generation of this project primarily involves questions corresponding to the unbounded constituents which fall into two main groups. The *nominal* ones are subjects, objects/predicatives and the rest are the various types of *adverbials*, of which certain kinds like sentence adverbials, are not considered here. The approach here particularly aims at a high precision value, i.e. the share of correct answers for the generated questions. On the other hand, the system presented does not attempt to make an exhaustive coverage of all questions (recall). The input to the question generation is a separate main clause. A construction with coordinated finite VPs on the main clause level similarly will produce a main clause of the second VP by inheriting the most recent main clause level subject in the sentence (*Halley's Comet is the best-known of the short-period comets, and is visible from Earth every 75 to 76 years. → Halley's Comet is the best-known of the short-period comets, Halley's Comet is visible from Earth every 75 to 76 years*).

3.1 The Process of Question Generation

The question types considered are similar in that all these questions are built up using a three-step procedure of syntactic fronting of the unbounded constituents and substitution of suitable question elements with *wh*-words or similar. The procedure is shown in Figure 2. First, the currently fronted element is placed in the canonical (non-fronted) position. This V1 form will in general be the corresponding *yes/no*-question. V1 questions are considered to be of slightly less interest than the others since they generally just confirm facts (the existence of such a question – *'Is Halley's Comet the best-known of the short-period comets?'* – just indicates the validity of that fact). The second step is fronting of each unbounded constituent from this arrangement, producing that number of paraphrases which are grammatical in Swedish. Finally, each fronted element is replaced by e.g. the corresponding *wh*-word to form a question which is collected.



[-] Kartlade Rutherford atomen på institutet?

Basic V1 form *(Was the atom surveyed at the institute by Rutherford?)* is the first step

Rutherford kartlade [-] atomen på institutet.

⇩

Vem kartlade atomen på institutet?
Who surveyed the atom at the institute?

Atomen kartlade Rutherford [-] på institutet.

⇩

Vad kartlade Rutherford på insitutet?
(Lit:) *What surveyed Rutherford at the institute?*

På institutet kartlade Rutherford atomen [-].

⇩

Var kartlade Rutherford atomen?
(Lit:) *Where surveyed Rutherford the atom?*

Figure 2: The basic procedure for question generation from declarative sentences

The number of questions produced can be lower than the number of unbounded constituents present in the sentence due to incomplete parses, as a general result of the method's focus on producing correct questions safely and rule-based with this transformation-like technique.

3.2 Questions Regarding Nominal Grammatical Categories

The nominal constituents correspond to a small set of Swedish *wh*-words corresponding to *what, which* and *who/whom*. The system currently works by determining the head word if it is an NP. If this is a personal pronoun or otherwise corresponds to an animate reference, e.g. a personal name or animate noun, *who* is used, whereas *vad/what* is the default. *What*-questions are currently the most common type of question and typically constitute half of the generated questions to a text.

3.3 Questions Regarding Adverbials

Question generation for adverbials is interesting as the choice of corresponding fronted initial part is more complicated than for nominals. Adverbials are structurally prepositional phrases, adverb phrases, noun phrases with a head from a particular group of nouns *(denna gång/this time)* and a subset of sub clauses. Whereas many of the members of the groups have clearly corresponding question words, the major PP type is particularly large in Swedish (133 different prepositions are currently covered) and have correspondences that often are determined by the head of the prepositional complement, rather than the preposition. This is particularly the case, as in English, for some of the most common prepositions: *i/in på/on* etc. Current adverbial questions considered are:

- NP adverbials (*denna gång/this time*), which predominantly refer to time and are replaced by *när/when*.
- PP adverbials. Swedish is particularly rich in prepositions since adverb + preposition compounds (*inifrån/"from-within"*) are frequent.
- The subset of sub clause types which corresponds to adverbials (*eftersom/since*).

Particularly in the case of prepositional objects a pied piping question (*Till vad/To what*) is mostly preferred as a question form.

4 Results and Possible Improvements

Aspects examined in evaluation of QG systems have e.g. been represented by the following categories of errors from Heilman and Smith (2009), which can be overlapping: “Ungrammatical”, “Does not make sense”, “Vague”, “Obvious answer”, “Missing answer”, “Wrong *wh*-word”, “Formatting” and “Other”. The lack of formal definitions of these terms has not encouraged such evaluations at this early point.

4.1 Preliminary Tests

In a minor test with the current system, ten random Wikipedia articles were used, including 78 sentences. The system produced 309 questions (in average about 4.0 per sentence) in 6.3 seconds. Grammatical correctness of the questions is currently not very high according to manual tests. Higher correctness is however likely to be achieved after further work with rules for choice of question words. Since the approach is essentially manual and few sentences are deemed as impossible to analyze, or to generate questions from, the potential correctness of the approach is seen as high. At present, no similar system seems to exist for Swedish text that could be used for comparisons.

The idea of producing all possible questions for an input text is far from realized here. Future work may concern other “safe” conclusions, yielding new questions, such as propositions of on sub-clause levels in constructions with factual verbs (*She knows it will work → it will work*) or questions regarding grammatical modifiers *(They sold the new boat → Which boat did they sell?*).

4.2 Expanding the Set of Formulations of Questions

Ideally, the question set produced consists entirely of questions that are correctly answered by the text. The user of this type of system however faces a different task: finding a formulation of a question she has in mind that corresponds to the text. To help the user find information, it has been assumed that creating additional alternative formulations of questions will generally be helpful. The main difficulty with expanding the question set using synonym substitution (*What automobile/What car*) is that few word pairs qualify as true substitutes. Earlier tests have been carried out testing substituting present base form words with synonyms according to the Swedish *WordNet* (Viberg et al 2002) and *Folkets synonymordlista* (Kann and Rosell 2005). The proportion of truly substitutable word pairs in Wikipedia texts was about 50-60 percent for these sources, considering all suggestions without any word sense disambiguation. In *Folkets Synonymordlista*, there is however a great potential advantage in the fact that each pair of suggested synonyms are judged with a numerical scale up to 5.0. Setting a high threshold score, like 4.5, will leave a smaller number of synonym pairs but increase the appropriateness of the substitution.

References

Converse, Tim, Ronald M Kaplan, Barney Pell, Scott Prevost, Lorenzo Thione, and Chad Walters. ”Powerset’s Natural Language Wikipedia Search Engine.” *Wikipedia and Artificial Intelligence: An Evolving Synergy. Papers from the 2008 AAAI Workshop.* Chicago, USA: AAAI Press, 2008. 67.

Diderichsen, Paul. *Elementær Dansk Grammatik.* Köpenhamn: Gyldendahl, 1946.

Ejerhed, Eva, Gunnel Källgren, and Benny Brodda. *Stockholm-Umeå corpus version 2.0.* Institutionen för Lingvistik, Stockholms universitet, Institutionen för Lingvistik, Umeå universitet, 2006.

Harabagiu, Sanda M, Marius A Paşca, and Steven J Maiorano. ”Experiments with Open-Domain Textual Question Answering.” *Proceedings of the 18th conference on Computational linguistics - Volume 1.* Saarbrücken: International Conference On Computational Linguistics, 2000. 292 - 298.

Heilman, Michael, and Noah A Smith. ”Ranking Automatically Generated Questions as a Shared Task.” *Proceedings of the AIED Workshop on Question Generation.* Brighton, UK, 2009. 30-37.

Kann, Viggo, and Magnus Rosell. ”Free Construction of a Free Swedish Dictionary of Synonyms.” *Proceedings of 15th Nordic Conference on Computational Linguistics – (NODALIDA 05).* Joensuu, 2005.

Rus, Vasile, and Arthur C Graesser. *The Question Generation Shared Task and Evaluation Challenge.* Workshop Report, Memphis, USA: The University of Memphis, 2009.

Viberg, Åke, Kerstin Lindmark, Ann Lindvall, and Ingmarie Mellenius. ”The Swedish WordNet Project.” *Proceedings of Euralex 2002.* Köpenhamn, 2002. 407-412.

Wilhelmsson, Kenneth. *Heuristisk analys med Diderichsens satsschema - tillämpningar för svensk text.* Gothenburg, Sweden: Department of Philosophy, Linguistics and Theory of Science, 2010.