Enhancing access to public information

Magnus Merkel, Michael Petterstedt & Arne Jönsson

Department of Computer and Information Science,
Linköping University, Sweden
magme@ida.liu.se, g-micpe@ida.liu.se, arnjo@ida.liu.se

Abstract. We are gradually entering a society where multiple electronic information sources, located both globally and locally, are available to support the performance of everyday activities, and are accessible through a wide range of devices, both stationary and mobile. Being able to access the information does, however, not necessarily mean that it is easy to find what is relevant for a certain task in a given situation. We present a system exploring the prospects of using a simple ontology in order to enrich public electronic documents with domain-specific information to allow for simple question-answering.

1 Introduction

Public information from authorities and public organizations is in most cases found in text or in semi-structured documents, and not as structured data, such as databases. Given the increasing need for information, it is of utmost importance to promote efficient means and techniques for easy access to such information and also techniques for pinpointing specific and relevant information. This involves new issues in the area of interaction design where documents must be analysed and structured before allowing the information system to act appropriately.

As a result of many public authorities' endeavours to make public information more readily accessible, the Swedish Tax authorities (Riksskatteverket, RSV) have been very active to publish regulations, tax forms, instructions and other documents on their web site (http://www.rsv.se). The web site holds thousands of documents in various formats (html, text, pdf, etc.) and the document collection is searchable through standard IR facilities (i.e. a Search box). However, when dealing with vast amounts of information, a simple Boolean search approach does not assist the user either in expressing the appropriate request nor does it aid the system sufficiently in pinpointing the required information in the document collection.

2 Simple Q&A with ontology in the tax domain

ASK-RSV is a Q&A system to tax documents describing taxes and the tax account for citizens in Sweden. ASK-RSV utilised several standard language processing techniques, from document conversions to morphosyntactic analyses. The documents were converted from PDF and Word to XML via html, and domain-specific lexicons were automatically constructed using Conexor's Swedish FDG parser[1].

The lexicon mentioned above was also extended with multi-word units, extracted from the document collection using a phrase extractor. The XML documents were then marked-up syntactically using information from the parser and domain-specific lexicon, including multi-

[1] FDG stands for Functional Dependency Grammar. See http://www.conexoroy.com for further references about Conexor's linguistic analyzers.

word terms. However, it was apparent that only syntactic information would not significantly increase the possibilities of retrieving the appropriate answers; semantic information was indeed necessary in some form.

A question corpus was also compiled, mainly from FAQs from the web site and by manual inspections of the documents. The questions assembled were factoid questions and did not require any explicit subdialogues as the answers can be extracted as snippets of text from the documents. However, the question corpus revealed a widespread use of synonyms and hyponyms that were not present in the documents. Many concepts in the tax domain are expressed differently by experts and laymen, and this was obvious from comparing the actual documents with the question corpus. For example, an expression like ”back tax” (Sw. ”restskatt”) is likely to be used by a layman questioner, whereas the documents only mention the more precise term ”tax account deficit” (Sw. ”underskott på skattekontot”).

This led to the construction of a simple domain-specific ontology containing concepts in the tax domain. Around 100 concepts were defined in the ontology, realising hyponymy and synonymy relations. The lexicon was then updated with the references to the ontological objects. A number of templates that could handle temporal and numerical expressions were added which together with the ontological information were used to mark up the XML document semantically.

In the question interpretation module of ASK-RSV the same ontological representation was used, but only with a bag-of-words approach, meaning that no detailed analysis of the questions was made. Instead the content words were mapped to ontological objects and these were in turn used to find appropriate sentences and sections in the documents that were extracted as answers. The extracted answers were then ranked depending on a weighting scale, based on the ontological distance between the concepts used in the question and the semantic information expressed in the candidate answers.

Even though the approach taken in the ASK-RSV is admittedly a simple one given the restricted ontology design and the restricted question analysis, the system shows impressive results when the initial question corpus were tested. The system is capable of handling questions about interest rates, sums, dates and intervals, definitions of concepts, but will fail on any interaction that would require follow-up questions or connected dialogue.

3 Discussion and ongoing work

ASK-RSV was our first step towards extending access to public information. Through techniques for document analysis, applied to various types of documents, simple Q&A functionalities involving factoid questions and answers to document collections have been developed.

However, the relevant information need can not always be successfully formulated in a single request. Instead the user and system have to take part in a task-oriented dialogue where various pieces of information have to be provided by the user for the system to arrive at a satisfactory formulation of the information problem. Thus, current work involves extending Q&A systems with dialogue capabilities and means for multimodal interaction. In particular, we are focusing on two information systems, BIRDQUEST and ASK-KOMMUNAL. BIRDQUEST deals with information on Swedish birds and is being developed in cooperation with Swedish Television (SVT). ASK-KOMMUNAL lets users interact with documents holding information on trade union agreements. The major differences compared to Ask-RSV lies that both the new systems utilise a proper dialogue manager, which enables collaborative dialogues, a significantly improved module for question interpretation, as well as a richer ontological framework.