Combining Case-Based Reasoning and Semantic Indexing in a Question-Answer Service

Antti Vehviläinen, Olli Alm and Eero Hyvönen
Semantic Computing Research Group (SeCo)
Helsinki University of Technology (TKK), Laboratory of Media Technology
University of Helsinki, Department of Computer Science
firstname.lastname@tkk.fi, http://www.seco.tkk.fi/

Abstract

This paper argues that knowledge technologies can be utilized in creating question-answer services on the semantic web. To ease the content indexer's work, we propose semi-automatic semantic indexing for annotating question-answer pairs and case-based reasoning techniques for finding similar questions. To provide answers matching with the indexer's and end-user's information needs, methods for combining case-based reasoning with semantic search and browsing are proposed. A real life ontology-based question-answer application OPAS is presented as a proof of concept.

1 Introduction

We consider question-answer (QA) services where human experts provide answers to written questions. Examples of such services include help-desks and the popular Frequently Asked Questions (FAQ) lists. QA services share some characteristics: 1) An archive of existing questions and answers is usually available for browsing and searching. 2) The QA pairs are often annotated with *index terms* for information retrieval purposes. 3) Questions are frequently repeated.

This paper shows how semantic web technologies can be utilized in QA systems from 1) the content indexer's and 2) the end-user's viewpoints. The ideas are presented by describing ongoing work of a semantic QA service OPAS[1]. OPAS is based on the existing *Ask a librarian* service[2] offered by the Helsinki City Library. Here the clients can send questions to a virtual librarian via email, and a librarian of the service provides an answer within three working days. All QA pairs are indexed by the librarians using the YSA thesaurus[3] of some 23,000 common Finnish terms. The dataset at the moment consists of over 20,000 QA pairs. A keyword-based search service is available on the web for both end-users and indexers to use. In this paper, we focus on the content indexer's viewpoint and the two main features of OPAS: 1) how to use semi-automatic semantic indexing to help in choosing appropriate index terms for QA pairs and 2) how to apply case-based reasoning (CBR) for finding existing similar questions for a new submitted question.

[1] http://www.seco.tkk.fi/applications/opas/

[2] http://www.kirjastot.fi/tietopalvelu

[3] http://vesa.lib.helsinki.fi

2 Semi-Automatic Semantic Indexing

Two major problems of the current service were identified from the indexer's viewpoint: 1) Choosing the appropriate index terms for a question-answer pairs is often time consuming and difficult. 2) There are different conventions used in indexing by different people, which makes the content unbalanced. For example, one librarian may use a few general terms to describe an answer, whereas another uses a large number of more detailed terms.

To address these problems semiautomatic indexing is employed in OPAS. We created an ontology-based information extraction tool POKA[4] for textual data, and integrated it with OPAS. POKA provides the QA indexer with a list of possible index terms as ontological references, and the indexer chooses which terms she wants to use. The basis for indexing with *common noun terms*, such as *dog*, *astronomy*, or *child*, is the General Finnish Upper Ontology, YSO[5] that is an ontologized version of the YSA thesaurus used originally in the service. YSO contains over 20,000 Finnish concepts organized into 10 major subsumption hierarchies. POKA also supports semantic indexing with *free indexing terms*, such as places (*Beijing*), person names (*John Lennon*), or other specific terms (*Lassie*) not found in the ontology. For each specification, an instance of the corresponding class (e.g., *city*, *person*, *dog*) can be created. Free indexing terms with the same name can be distinguished with different URIs and with an additional comment.

If the input text is long, POKA yields a considerable number of possible index terms. For that reason it is useful to order the terms according to their likely relevance. In our case we use the idea [Luit Gazendam and Brugman, 2006] of searching for *semantic cluster(s)* from the term set and conclude that these terms are more relevant than semantically isolated terms. For example terms *doctor*, *sickness* and *medication* form a semantic cluster. For common noun terms we use the concept relations defined in YSO to identify these clusters. In [Holi *et al.*, 2006], an ontological extension of the classic tf-idf (term frequency - inverse document frequency)

[4] http://www.seco.tkk.fi/applications/poka/

[5] see http://www.seco.tkk.fi/ontologies/yso/

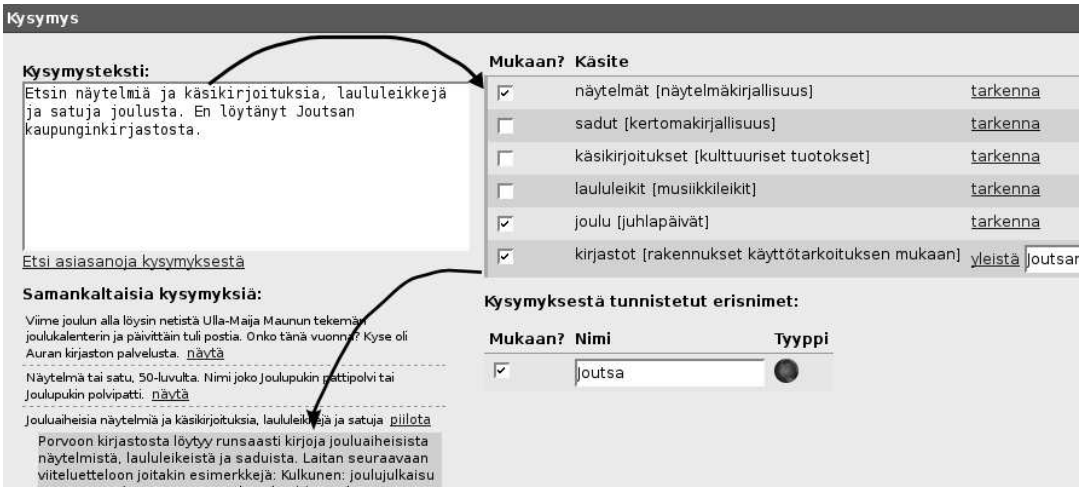

Figure 1: An example of the semi-automatic indexing and similar question finding in the OPASUI

method was developed, which enables us to identify synonyms and to utilize the concept hierarchies of the ontology. We apply this work so that more weight is given to terms that appear frequently in the text but haven't been used often as index terms in previous questions. In addition, OPAS can suggest index terms that are usually used together. For example, if a question has the term *aviation* extracted, and there are lots of questions indexed with both *aviation* and *airplane*, the term *airplane* can be suggested for indexing, even though it is not explicitly present in the question text.

3 Utilizing Case-Based Reasoning

Case-based reasoning (CBR) [Aamodt and Plaza, 1994] is a problem solving paradigm in artificial intelligence where new problems are solved based on previously experienced similar problems. Since similar QA pairs recur in QA services, we decided to investigate the usefulness of CBR in QA indexing and information retrieval. OPAS contains a CBR component that automatically searches for similar questions based on the terms that POKA has extracted from the question text. The weighted index term list discussed above is used as the basis for the search with the following modifications: 1) The terms that the indexer has selected are given a substantially higher weight since their relevance has been confirmed by the indexer. 2) The extracted places, names and specified terms are given a higher weight due to their specificity.

Figure 1 depicts a screenshot of OPAS. The end-user has submitted a question about Christmas fairy tales and plays (on the left, in the box "Kysymysteksti"). The index terms suggested by POKA are on the right, and the indexer has selected index terms *näytelmät* (plays), *joulu* (Christmas) and *kirjastot* (libraries). Below, the place name *Joutsa* is shown separately, because the question concerned this place name that is not present in YSO ontology and that is considered a free indexing term. Using the button "Tyyppi", the right type (class) of the instance can be given. Based on the found terms and the term selections, OPAS has found a set of similar QA pairs. On the left bottom, the indexer has opened one of them in order to see whether it provides useful information for answering the question.

4 Evaluation, Discussion and Further Work

First experiments with combining semi-automatic semantic indexing with the ideas of case-based reasoning seem promising. However, the work is still ongoing, and empirical evaluations of the application with real librarians and end-users are yet to be done. Currently OPAS is focused on the indexers' role in QA applications but OPAS will include the end-users' side, too. Here we work on questions such as: how to classify the QA pairs for semantic view-based search, how to do semantic recommending in order to show other interesting answers, and how to integrate the system with semantic content and services at other locations on the web related to the end-user's information needs.

Our work is funded mainly by the Finnish Funding Agency for Technology and Innovation (Tekes).

References

[Aamodt and Plaza, 1994] Agnar Aamodt and Enric Plaza. Case-based reasoning: foundational issues, methodological variations, and system approaches. *AI Commun.*, 7(1):39–59, 1994.

[Holi *et al.*, 2006] Markus Holi, Eero Hyvönen, and Petri Lindgren. Integrating tf-idf weighting with fuzzy view-based search. In *Proceedings of the ECAI Workshop on Text-Based Information Retrieval (TIR-06)*, Aug 2006. To be published.

[Luit Gazendam and Brugman, 2006] Guus Schreiber Luit Gazendam, Veronique Malaisé and Hennie Brugman. Deriving semantic annotations of an audiovisual program from contextual texts. In *Semantic Web Annotation of Multimedia (SWAMM'06) workshop*, 2006. http://www.cs.vu.nl/ guus/papers/Gazendam06a.pdf.