SUiS – cross-language ontology-driven information retrieval in a restricted domain

Kristina Nilsson[*†], **Hans Hjelm**[*†], **Henrik Oxhammar**[*]
[*]CL Group, Department of Linguistics
Stockholm University, SE-106 91 Stockholm, Sweden
[†]Graduate School of Language Technology (GSLT)
{kristina.nilsson,hans.hjelm,henrik.oxhammar}@ling.su.se

Abstract

In this paper we present SUiS - Stockholm University Information System. SUiS tries to find answers to a predefined set of question types: who, what, when and where. A domain-specific ontology is used for query expansion and translation, for answer generation, and for document analysis together with Named Entity Recognition modules for recognizing people, locations, organizations, and date expressions. Qualitative user evaluations show satisfactory results as well as indications of areas where the system could be improved.

1 Introduction

SUiS, Stockholm University Information System, functions as an automatic answering service, which tries to find answers to a restricted set of question types: who, what, when and where. The question set is predefined in order to increase the precision of the results: by restricting the questions we can predict what kind of analysis is needed to find the answer, e.g., what type of named entities must be recognized in the retrieved documents in order to find the answer to the posed question. The answers are presented as text excerpts, where the relevant entities are marked up. SUiS is not a prototypical question answering system as described in e.g., (Harabagiu and Moldovan, 2003) in that it is restricted to a specific domain, and does not allow for free-form questions.

SUiS uses the Stockholm University Ontology[1], a domain-specific ontology covering the university domain, for query expansion and translation. Concepts in the ontology are referred to by words and expressions in both English and Swedish. Thus, SUiS can be used for cross-language information retrieval of Swedish and English documents. SUiS is currently restricted to the Stockholm University web domain[2].

Research on using semantic information stored in e.g., thesauri or ontologies for query expansion shows that general-purpose resources such as WordNet (Fellbaum, 1998) seem to have very little, or even negative, impact on the results. If queries are expanded with synonyms, hyponyms, and hypernyms to increase recall, this can lead to a too broad interpretation with a resulting decrease in precision (Tzoukermann et al., 2003; Vossen, 2003). Domain-specific ontologies however usually work well for the domain they were created (ibid.), and by using a conservative form of query expansion where only synonyms and/or more specific terms (i.e., hyponyms) are added, this problem can be avoided.

1.1 Related work

Within the MOSES project, a cross-language QA system for free-form querying of knowledge bases is used. The facts in the knowledge bases are extracted from two separate web sites in the university domain. NLP techniques such as part-of-speech tagging, lemmatization, and query analysis are used to produce a semantic representation of each query, which is then matched against the knowledge bases (Atzeni et al., 2004). The ontologies are each created specifically for the web sites of the two universities, and thus there are differences between the ontologies: they are in different languages (Italian with concepts labels in English, and Dan-

[1] Henceforth referred to as the SU Ontology.

[2] Stockholm University, URL: http://www.su.se

S. Werner (ed.), *Proceedings of the 15th NODALIDA conference, Joensuu 2005*, Ling@JoY 1, 2006, pp. 139–145
ISBN 952-458-771-8, ISSN 1796-1114. http://ling.joensuu.fi/lingjoy/

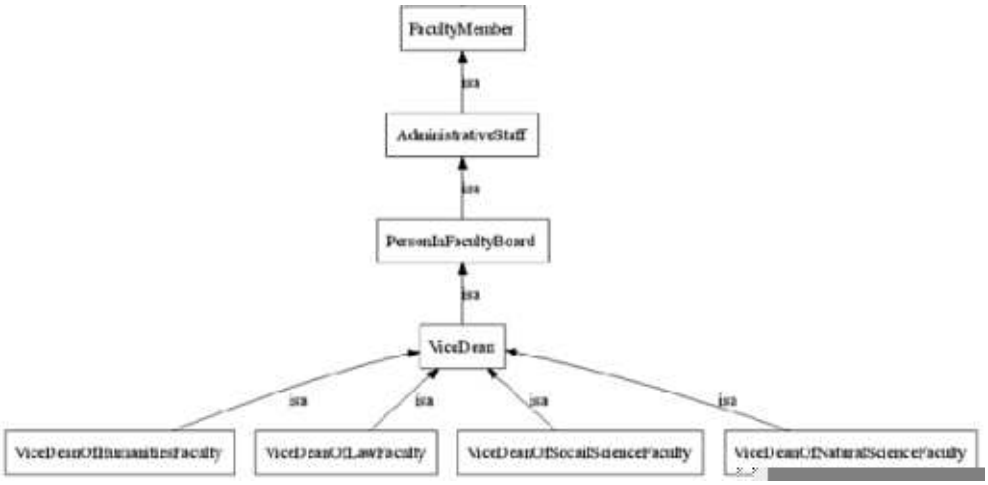

Figure 1: Super- and sub concepts of the concept 'ViceDean' in the SU Ontology.

ish respectively), they are structurally different and only partially overlapping, and thus require mapping as described in (Pazienza et al., 2005).

In the MUCHMORE project, ontologies are used as an interlingua for cross-language information retrieval of German and English documents in the medical domain. The results show that the use of ontologies for query translation improves the performance of the system compared to machine translation (Volk et al., 2003).

2 Stockholm University Ontology

An ontology can be defined as an inventory of the objects and their properties within a specific domain as well as a specification of some of the relations that hold among them (Mitkov, 2003). The SU Ontology contains concepts that refer to university entities in general, e.g., lectures and undergraduate students, as well as entities specifically associated to Stockholm University, e.g., the campus area Frescati.

The SU Ontology describes concepts in four sub-domains, namely: *occupations*, *educational programs and degrees*, *places* and *events*. It includes concepts such as 'ViceDean' and 'ResearchAssistant' in the *occupations* sub-domain, 'DepartmentOfLinguistics' in *educational programs and degrees*, 'AulaMagna' in *places*, and 'Enrollment' in the sub-domain *events*.

In the ontology, such concepts are described in relation to other concepts, e.g., 'ViceDean' is a type of 'PersonInFacultyBoard', similar to the concept 'Dean'. This relationship describes the concepts as a hierarchy of classes, with classes ordered in a super/sub-class structure. This type of structure provides information on super/sub class relationships (hyponymy) and relatedness (sibling relationship); the super- and sub-concepts of the concept 'ViceDean' are shown in figure 1, above.

Each concept is referred to by words and expressions in both English and Swedish in the ontology, e.g., the concept 'ViceDean' is referred to by the English phrase 'vice dean' and the Swedish equivalent 'prodekan', thus allowing for the translation between the two languages that is required in Cross-Language Information Retrieval. The inclusion of synonyms in both languages also allows for query expansion.

The open source ontology editor Protégé[3] (Noy et al., 2001) has been used for building and maintaining the ontology.

3 SUiS system architecture

The system makes use of two kinds of knowledge sources: a domain specific ontology (see section 2), and domain independent date and named entity recognition modules for the name types *person names*, *organizations*, and *locations*. The named entity recognition method has been developed by Volk and Clematide (2001).

3.1 Components

SUiS consists of the following components:

[3] Available at http://protege.stanford.edu/index.html (Last checked Oct. 20, 2005.)

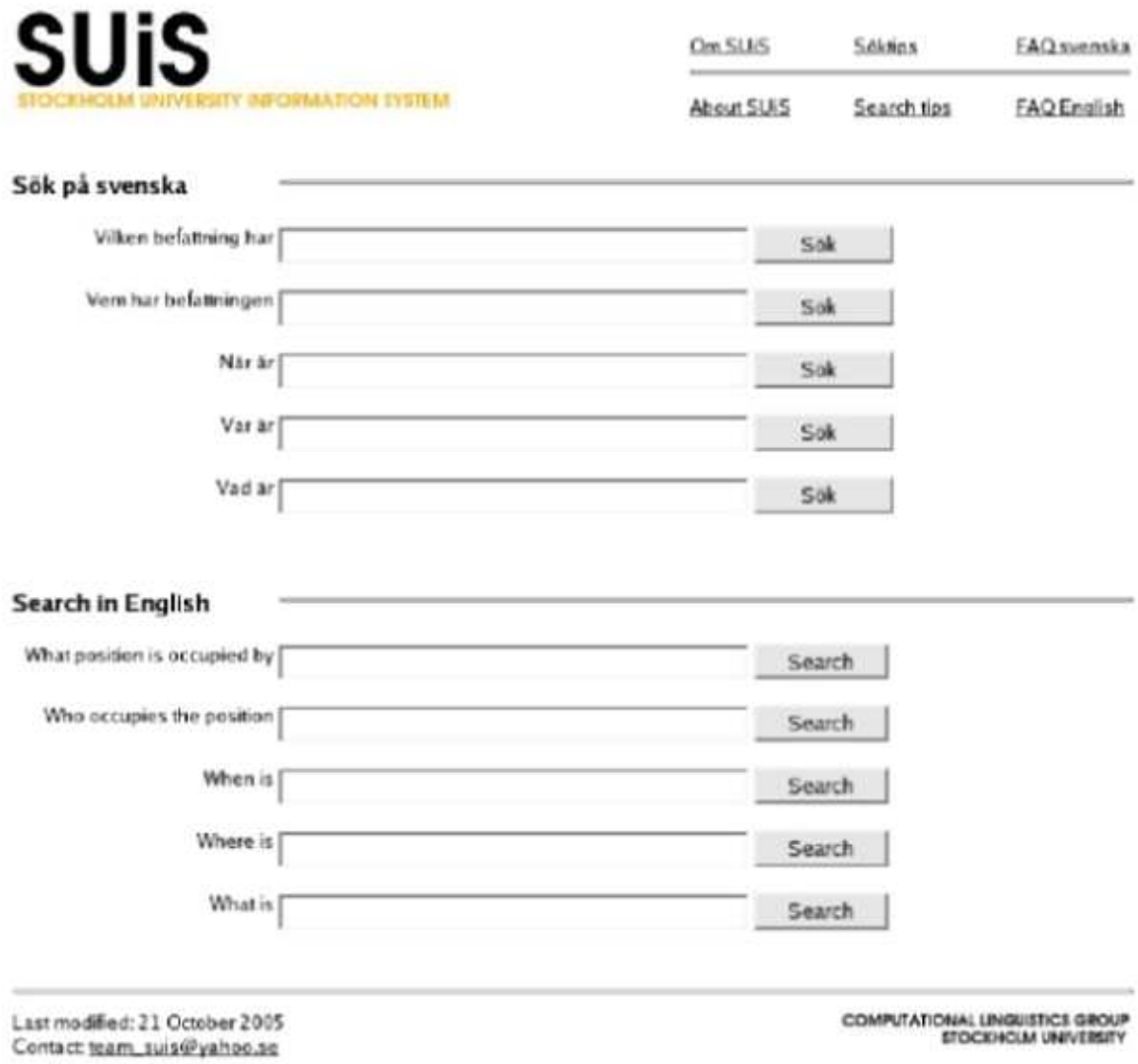

Figure 2: SUiS Interface.

Interface The interface is an HTML-page, communicating with the rest of the system over a CGI-connection (see figure 2). The interface constricts the user to a specific question format. SUiS can handle the following types of factual questions:

- What position is occupied by `person`?
- Who occupies the position `position`?
- When is `event`?
- Where is `location/organization`?
- What is `entity`?

Information on what type of question has been posed is passed along to the system. This information determines which modules are applied (e.g., the what-question, as opposed to other question types, bypasses the IR-part and the answer is generated directly from the ontology).

Query Parser The query parser removes stop words and extracts query parts relevant for expansion. E.g., from the query string 'Who occupies the position of Vice Chancellor of Stockholm University?', the title 'Vice Chancellor' is extracted for expansion in the Query Expansion Module. The preposition 'of' is removed and the remainder of the string ('Stockholm University') is kept as it is.

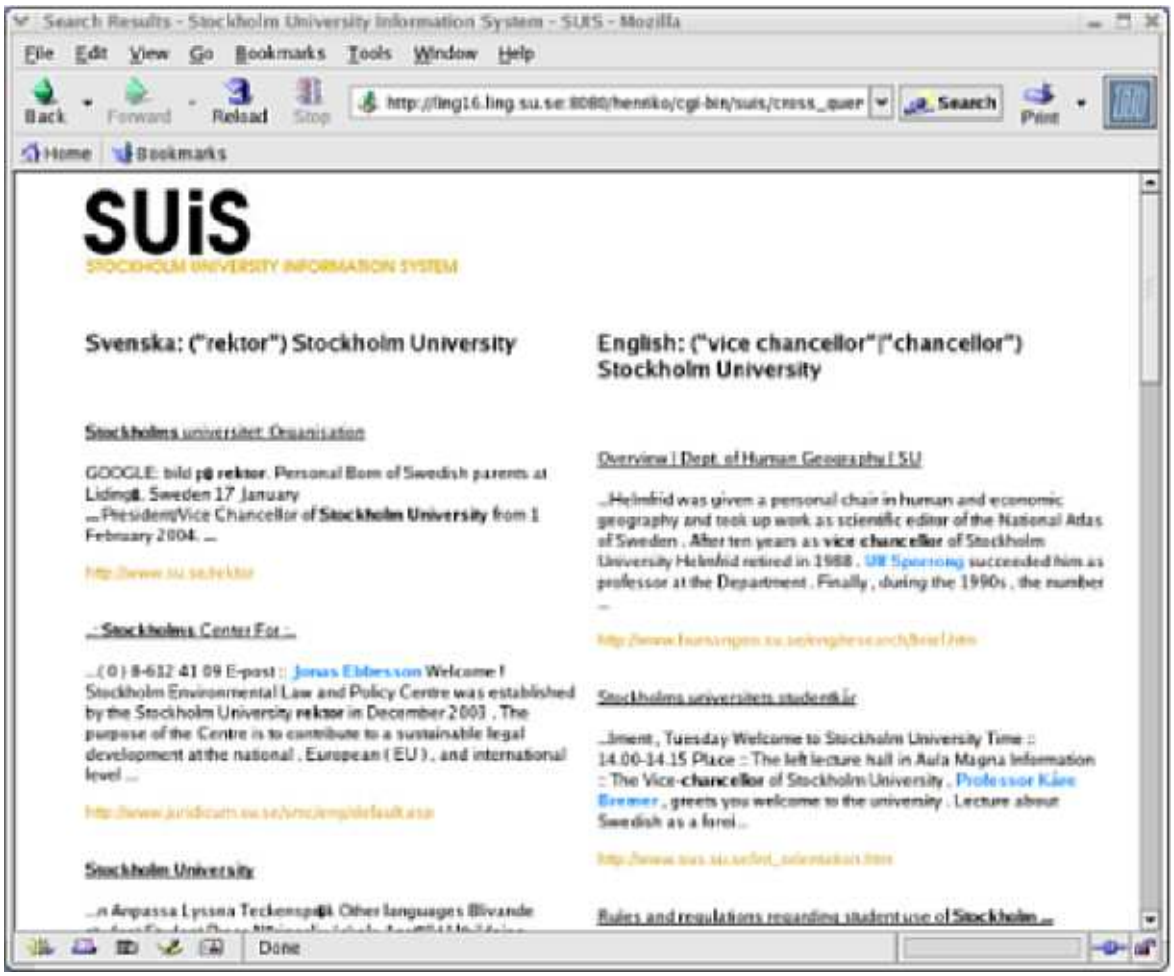

Figure 3: SUiS output for the query 'Who is Vice Chancellor of Stockholm University?'.

Query Expansion and Translation Module Continuing with our previous example, this module looks up 'vice chancellor' in the ontology and, having found the corresponding concept 'ViceChancellor', looks for hyponyms (sub concepts) and synonyms to expand the query. Query translation for cross-language querying is performed by looking up translation equivalents for that concept. The translation equivalents are in turn expanded with hyponyms and synonyms, if possible. We end up with one expanded English query, 'vice chancellor OR chancellor' and its Swedish translation 'rektor' (see figure 3).

Google Web API For searching the web pages in the Stockholm University domain (presently about 495.000 pages) the Google Web API is used. Google's language restrict option enables SUiS to search a predefined subset of the index, and thus to retrieve pages in English and in Swedish separately.

Document Retrieval For each query, we receive a ranked list of URLs from Google. The corresponding documents are retrieved and stripped from HTML.

Named Entity Recognition For questions of the type 'Who occupies the position *position*?', each retrieved document is processed by a named entity recognition module. This module marks up the person names in the text. The named entity recognition module for locations is used for questions of the type 'Where is *location/organization*?'. The named entity recognition modules are not used for the other question types.

Date Recognizer When answering questions of the type 'When is *event*?', the date recog-

nizer marks up any date and time expressions found in the text.

Ontology The purpose of the ontology is threefold:

1. It is used during *query expansion and translation*, as described above. The added concepts and synonyms are connected by the Boolean `OR` operator.

2. When answering the question type 'What position is occupied by *person*?', the part of the ontology describing the sub-domain *occupations* is used for marking up occupations in the text.

3. When answering the question type 'What is *entity*?', e.g., 'What is a teaching assistant?', 'teaching assistant' is looked up in the ontology and the answer is generated from its super concept and siblings.

Output Generator The results are presented to the user as a ranked list of URLs and a brief excerpt of the document, following the ranking from Google. In this excerpt, the query term(s) are highlighted in bold and named entities and dates are highlighted in color codes, using different colors for different types of names (see figure 3).

4 Future work

A qualitative user evaluation has been performed, indicating a number of areas where SUiS can be improved. Firstly, the ontology needs to be expanded and enhanced. Currently the SU Ontology covers around 140 occupations, 70 educational programs and degrees, 80 locations and 80 types of events. But many important concepts are not yet modeled in the ontology, and must be added. Further, inconsistencies, incorrect relationships and incomplete translation information need to be corrected or added to improve on the quality of the ontology.

The question is how one can tackle the problem of expanding the ontology. There exist several methods within the area of Ontology Learning for automating the process of populating and enriching ontologies with new concepts as well as maintaining ontologies (see e.g., (Aguirre et al., 2000; Faatz and Steinmetz, 2002; Hahn and Schnattinger, 1998)).

However, the university domain is believed to be rather static, in comparison to other domains (e.g., IT, medicine), and does not undergo frequent changes. Organizations (e.g., departments), locations (e.g., seminar rooms), occupations (e.g., dean, research assistant), and degrees (e.g., master/bachelor degree) are unlikely to change names or even less so to disappear. There are of course exceptions, for instance concepts denoting educational programs and courses. Having said that, a manual approach, provided a good editing- and visualization tool such as Protégé (Noy et al., 2001), should be sufficient for keeping the ontology current and of good quality.

Aside from enhancement of the ontology, future efforts will be focused on two areas in particular:

1. Query translation A query such as 'Who is the assistant professor in Computational Linguistics?' is translated through look-up of the concepts in the ontology (see section 3), but as there is no concept 'ComputationalLinguistics' in the ontology, this phrase cannot be translated by the current system. The present solution is to translate as much of the query as possible, and add any remaining query terms untranslated. Thus, the example query is translated into a string consisting of the Swedish translation 'Vem är lektor i'(*Who is the assistant professor in*) and the original English query string 'Computational Linguistics'. A possible solution would be to add a lexicon, with both general and domain specific terminology, or to add a machine translation system.

2. Re-ranking and grouping of search results The current version presents the analysis of the retrieved documents in the same order as the documents are returned from Google. However, after the system has analyzed the documents, it has more information about the relevancy of the documents at its disposal than was present during Google's ranking. E.g., a document returned for the search 'Who is NN?', where the system finds no matching occupation or title, should be given a low ranking, no matter how relevant Google's ranking algorithm judged the document to be. Conversely, a document where an occupation appears in close proximity to the name entered in the query, should be given a high ranking. Another way to re-rank the results would be

to try to find groups of documents which all give the same answers. These documents could be grouped together in the presentation of the results, to give the user a better overview of which possible answers the system has found. Within these groups, the results could be ordered according to relevancy as described previously. The groups themselves could also be ordered, with the most relevant group presented first.

5 Concluding remarks

In this article, we have described a prototype for ontology-driven cross-language information retrieval in a restricted domain. The system uses a domain-specific ontology for query expansion and translation, together with modules for recognizing named entities and temporal expressions.

The ontology contains concepts from the university domain in general as well as concepts specific to Stockholm University. It contains concepts from four sub-domains: *occupations*, *educational programs and degrees*, *places* and *events*. Synonyms and hyponyms are used for query expansion and the corresponding terms in the target language are used for the cross-language search. The named entity recognition modules also contain knowledge resources, but these have a flat structure (i.e., they are kept as lists) as opposed to the ontology resources.

The SUiS system consists of six major components: the web interface, the query parser, the query expansion and translation module, the named entity recognition module, the ontology and the output generator.

A user evaluation was carried out, partly confirming our theories of the effectiveness of the system but also pointing towards a number of areas where the system could be improved or extended. The areas where improvements or extensions would have the greatest impact are the ontology (correcting and extending it), query translation (increasing the robustness) and re-ranking and grouping the search results.

Although covering the university domain in the present version, we believe that the system is highly portable to other domains (e.g., the financial domain) and languages. When porting the system to another language, it would be necessary to update the knowledge sources used in the named entity recognition modules. The content of the ontology would also have to be adapted to the new domain. However, updating the knowledge resources represents a comparatively minor effort, considering that the rest of the system can be reused as it is.

Acknowledgments

We would like to thank professor Martin Volk of Stockholm University, who initiated this project, and the students who participated in the work of compiling the SU Ontology, as well as in the first evaluation of the system.

References

Eneko Aguirre, Olatz Ansa, Edward Hovy, and David Martinez. 2000. Enriching very large ontologies using the WWW. In *Proceedings of the Ontology Learning Workshop, organized by ECAI*, Berlin, Germany.

Paolo Atzeni, Roberto Basili, Dorte H. Hansen, Paolo Missier, Patricia Paggio, Maria Teresa Pazienza, and Fabio Massimo Zanzotto. 2004. Ontology-based question answering in a federation of university sites: the MOSES case study. In *Proceedings of the 9th International Conference on Applications of Natural Language to Information Systems (NLDB'04)*, Manchester, United Kingdom.

Andreas Faatz and Ralf Steinmetz. 2002. Ontology Enrichment with Texts from the WWW. In *Proceedings of ECML-Semantic Web Mining 2002*, Helsinki, Finland, August.

Christiane D. Fellbaum, editor. 1998. *WordNet: An Electronic Lexical Database*. MIT Press.

Udo Hahn and Klemens Schnattinger. 1998. Towards text knowledge engineering. In *AAAI'98/IAAI'98 Proceedings of the 15th National Conference on Artificial Intelligence and the 10th Conference on Innovative Applications of Artificial Intelligence*, pages 524–531, Madison, Wisconsin, July. MIT Press.

Sanda Harabagiu and Dan Moldovan. 2003. Question Answering. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, chapter 31, pages 560–582. Oxford University Press.

Ruslan Mitkov, editor. 2003. *The Oxford Handbook of Computational Linguistics*. Oxford University Press.

Natalya F. Noy, Michael Sintek, Stefan Decker, Monica Crubézy, Ray W. Fergerson, and Mark A. Musen. 2001. Creating semantic

web contents with protege-2000. *IEEE Intelligent Systems*, 16(2):60–71.

Maria Teresa Pazienza, Armando Stellato, Lina Henriksen, Patrizia Paggio, and Fabio Massimo Zanzotto. 2005. Ontology mapping to support ontology-based question answering. In *Proceedings of the second MEANING workshop*, Trento, Italy, February.

Evelyne Tzoukermann, Judith L. Klavans, and Tomek Strzalkowski. 2003. Information Retrieval. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, chapter 29, pages 529–544. Oxford University Press.

Martin Volk and Simon Clematide. 2001. Learn-Filter-Apply-Forget. Mixed Approaches to Named Entity Recognition. In *Proc. of 6th International Workshop on Applications of Natural Language for Information Systems*, Madrid, Spain.

Martin Volk, Spela Vintar, and Paul Buitelaar. 2003. Ontologies in Cross-Language Information Retrieval. In *Proceedings of the 2nd Conference on Professional Knowledge Management*, Lucerne.

Piek Vossen. 2003. Ontologies. In Ruslan Mitkov, editor, *The Oxford Handbook of Computational Linguistics*, chapter 25, pages 464–482. Oxford University Press.