Sri Lanka Association for Artificial Intelligence (SLAAI)
Proceedings of second Annual Session
19th February 2005 – Colombo

English to Sinhalese Bilingual Web Client for accessing the Semantic Web

Asoka S. Karunannanda
Department of Information Systems and Computing
Brunel University, Uxbridge, Middlesex, UB8 3PH, UK
asoka.karunananda@brunel.ac.uk

Budditha Hettige
Department of Mathematics and Computer Science
Open university of Sri Lanka, Nawala, Nugegoda
budditha@yahoo.com

Abstract

Bilingualism has been recognised as a means for using the power of mother tongue for comprehending materials in a second language. We have been developing an English to Sinhalese bilingual translator as an Expert system, known as BEES, which can operate on an ordinary Web Client enabling the access to World Wide Web in a bilingual manner. The current system is capable of translating simple web pages appeared in English so as to generate bilingual displays of the web pages containing a certain percentage of English and Sinhalese words, depending on the user profile. BEES has been incrementally tested and currently we are working for improving BEES to handle semantics of bilingual translation towards a novel approach to ontological modelling for the Semantic Web.

1. Introduction

The next generation of World Wide Web is emerging as the Semantic Web, which is expected to enable humans and machines to work cooperatively using the huge repository of data, information and knowledge on the World Wide Web regardless of languages, applications, data-formats and operating systems (Tim Berners-Lee *et. al*., 2001). In particular, machine-machine communication is intended as a major goal of the Semantic Web. However, together with its marvellous ambitions of the Semantic Web, it would be appropriate to question how seriously humans have considered the

facilitation of communication and sharing of knowledge among humans regardless of different natural languages used by various nations. It is a well-known fact that there is a limited access to sources of world knowledge available in English for the nations whose mother tongues are other than English. Although one can suggest that everybody should learn English as a solution to this issue, this is rather impractical and wrong view. This is because languages are strongly based on respective cultures, societies and environments. In this sense, English as a language will

not be capable of expressing entire world knowledge. For example, there cannot be a better language than Japanese to present and explore Japanese system of knowledge. This applies to all the nations including Sri Lankans. In this knowledge-based economy world must be interested in evolution of knowledge within the respective best environments rather than attempting to promote the use of a single language world wide. However, since world knowledge is already available in English, all nations must also be knowledgeable in English language so as to use the available knowledge. Yet this solution fails to exploit the power of mother tongues for the benefits of the world in general and individuals in particular. What is the solution then?

The emerging solution for breaking the language barrier can be identified as the bilingual approach to communication and knowledge sharing. Bilingual approach encourages the use of a particular mother tongue together with a second language in such way that one language helps the comprehension when the other language fails to do so (Williams, 2000). It appears that almost all leading counties including France, Germany, Japan and Scandinavian countries have given prime importance for thinking in mother tongue and use of bilingual approach for education. Further, major English speaking countries such as United Kingdom and United States have also researching into power of bilingual approach for effective communication in multilingual societies. Many people believe that thinking also requires a language and it may be the mother tongue. This is a debatable idea, yet power of the bilingual approach has already been realised.

This paper presents our approach to use bilingual approach for making the huge repository of World Wide Web accessible by persons using Sinhalese as the mother tongue. We have been working on a research project to develop an Expert system that runs on an ordinary web browser and generates English to Sinhalese bilingual translation depending on the user profile. The system is named as BEES, an acronym for Bilingual Expert system for English to Sinhalese translation. The rest of the paper is organised as follows.

Section 2 describes the philosophy behind the bilingual approach. Section 3 reports on relationship between bilingualism and natural language processing. Section 4 describes the design and implementation of the bilingual translation system, BEES. Section 5 provides a discussion with a note on further work.

2. How Bilingualism works?

Language is an artifact for mediating between our perceptions and conceptions. Since humans are living in various societies, cultures and environments, different nations and societies perceive differently. This leads to emergence of various natural languages. One language for all is rather impractical idea and misguided by the ignorance of the importance of mother tongue. Although Sri Lankans think of English as a second language, people in many countries including European and Scandinavian countries really need more than English as their second language (Andersson & Andersson, 1999). For them learning only English does not solve their multi-language requirements. In this sense, it is quit natural for many people to think of strategies for learning and use of more than one language. According to literature, one solution would be the use of bilingual approach. The philosophy behind bilingualism can be presented as follows.

We all can remember how we learn a second language. It is not a secret that we used our mother tongues to step into the second language. Whenever, we cannot understand something entirely in the second language (say English), we are explained or supported with a little help from mother tongue. This is because mother tongue is readily capable of associating of process of our comprehension with what we already know. Mother tongue prompts the mind to appropriate context for thinking. In other words, we can exploit the power of mother tongue to learn a second language. We have used this approach not only as children but also as adults. This is the underlying philosophy behind the bilingual approach. In simple terms bilingual approach means the use of mother tongue whenever the knowledge including vocabulary of the second language is inadequate for further comprehension. The amount of supports required from the mother tongue depends on the person. During the bilingual process, at the beginning one may use the mother tongue more and end up with using the second language more recursively.

We argue that the bilingual concept with the above features can be used to implement semantic manipulation on the huge repository of world knowledge such as World Wide Web. Despite bilingual approach has been used a means for education; no one has used bilingualism as a means for semantic manipulation on the Semantic Web. Our BEES is an initiative for semantic manipulation between any two languages using the bilingual concept. We are confident that the experienced gained through the BEES project will contribute to devise a new approach for semantic handling on the Semantic web. However, this paper presents our findings

about the importance of bilingual approach to natural language handling. We argue that the following advantages can be readily obtained through the use of bilingual approach for handling natural languages.

- Use of power of mother tongue to comprehend the material in second language
- Improve the vocabulary of second language during bilingual process
- Ability to contribute mutual improvement to ontologies in both languages
- Ability to preserve thinking in mother tongue even after using a second language

Among other advantages, we wish to highlight the importance of the last point. Since bilingual approach proceeds with a reference to mother tongue, at the end of a learning session also, one will be capable of comprehending the material in terms of mother tongue. This point is relevance to Sri Lanka too, since some of our persons who are educated in English are unable to comprehend in mother tongue. This issue has also resulted in devising meaningful Sinhalese terms for the concepts in modern areas such as computer science. We argue that if bilingual approach is promoted, society will devise more meaningful terms through a natural process of their understanding. It is an utter failure to devise technical terms by using groups of subject experts and language experts. This is because, most of the time the subject experts cannot comprehend in mother tongue, while the language experts cannot comprehend the subject matter. We believe that BEES will be able to use by languages experts to devise more sensible Sinhalese terms for the concepts in modern subject areas.

3. Natural language processing & Bilingualism

Bilingual translation of natural languages can be considered under the area of natural language processing in Artificial Intelligence. However, bilingual translation is philosophically different from traditional natural langue processing, which is mainly based on manipulation of parsers of respective languages. Nevertheless, recent attempts to natural language processing on the basis ontological modelling can be closely compared with bilingual approach to natural language manipulation. This is because ontologies are more than parsers, yet provide semantic description of languages, associated relationships within a language and between two languages. Obviously, bilingual translation must also consider comprehensive ontologies for describing semantics of two languages. Next we present some of the recent ontology-based approaches to natural language processing and show their relevance and limitation with a reference to our BEES project.

Perhaps Chat-80 is one of the first major Prolog-based natural langue processing systems. It operates on a geographical database that comprises of relatively small ontology containing about 20 categories (Warren & Pereira, 1982). In contrast, Cyc is a huge expert system, which has a quite comprehensive set of ontology for enabling a variety of knowledge-intensive products and services to work together (Lenat, 1995). Ontology of Cyc includes many categories and millions of axioms associated with those categories. Cyc is not intended to do serious natural language processing, yet support the communication among various knowledge bases. In the mean time, WordNet is yet another system, which is merely developed for natural language processing pertaining to English language (Miller, 1995). It comprises of ontology of most general concepts in English language; *nouns, verbs, adjectives* and *adverb*s together with set of synonyms related to each. None of these Chat-80, Cyc or WordNet involves natural language translation between any two languages. However, EDR system can be identified as an electronic dictionary linking two languages of English

and Japanese (Yokoi 1995). EDR ontology consists of over 400,000 concepts, with their mapping to word in both English and Japanese languages. Although EDR deals with two languages, it does not exploit power of Japanese language to comprehend English language. The power of EDR is dependent on the expressiveness of mappings between the two language and this is not the philosophy behind bilingualism. Therefore, EDR cannot be treated as a bilingual approach to natural language processing.

There are also varieties of browser-based products for natural language translations on the World Wide Web. For example, Google has facility for translation of web pages in English to Sinhalese. However, such approaches generally use English alphabet to write the Sinhalese words. This leads to an issue due to the nature of Sinhalese language. For example, the Sinhalese translation of the English term "father" can be written using the English alphabet as either TATTA or THATHTHA. Further, English term "I" will be translated to Sinhalese and written in English alphabet as MAMA, which has the similar meaning as "uncle" according to English language. Therefore, the use of the English alphabet to write Sinhalese words is an unsuccessful approach and we must devise a mechanism to use the Sinhalese alphabet itself to write Sinhalese words. We have addressed this issue in BEES.

Sinhalese language has an additional feature which can be exploited to increase the power of bilingual approach. That means, Sinhalese language has a written language and a spoken language, which are considerably different. It is commonly agreed that the verbal languages are easier to understand than written languages. Further, grammatical constraints in Spoken Sinhalese can also be relaxed to a large extend. Therefore, we argue that spoken Sinhalese would be more appropriate for the bilingual approach than the written Sinhalese, especially when the material in the second language is difficult to understand or too technical. The current version of BEES has not exploited this dimension, yet we are interested in doing more research in this line. It should be reemphasised that bilingualism is more interested in supporting the users' comprehension of materials in a second language than doing a perfect grammatical translation between two languages. In other words, bilingualism permits to exploit any form of mother tongue for supporting the comprehension of materials in a second language. Therefore, research into spoken languages would be of great importance for bilingualism.

4. BEES – An approach to bilingual translation

Bilingual Expert system for English to Sinhalese (BEES) translation has been developed as software plug-in for an ordinary web browser such as Microsoft Internet Explorer. The overall system consists of five major modules, namely, an HTML Parser, an English parser, a Sinhalese parser, Bilingual Translator and an English-Sinhalese dictionary. The Expert system functionality in BEES are mainly implemented by the Bilingual translator module. Figure 1 shows the top level architecture of BEES. Next we briefly describe the functionalities of three parsers and the bilingual translator of BEES.

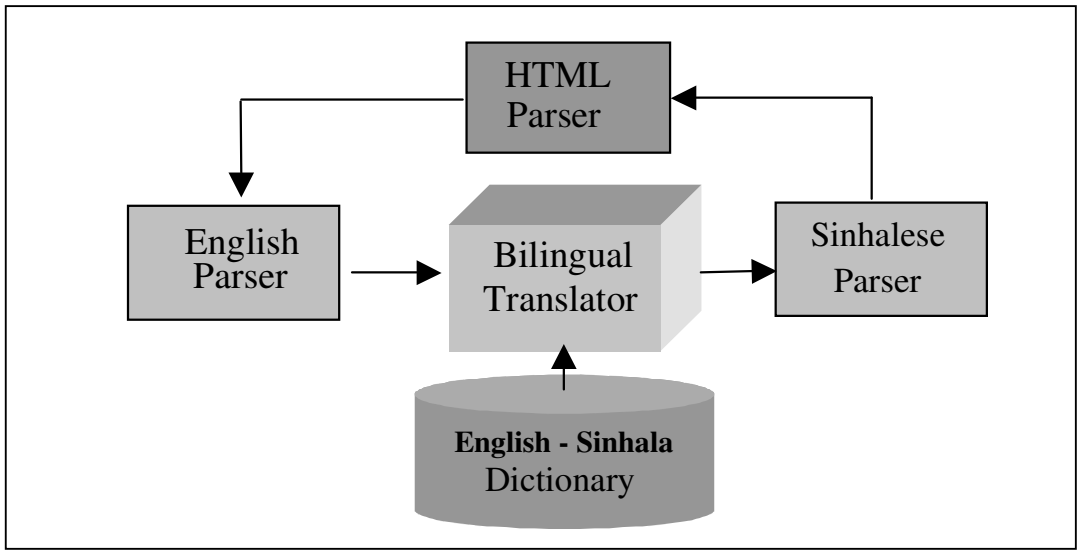

Figure 1: Top level architecture of BEES

4.1 HTML Parser

HTML parser has been developed for two purposes. Firstly, it reads web pages in English and decodes them so as to suitable for feeding into the English parser and also preserves the web page structure with HTML tags. The second purpose of the HTML parser is to read the bilingual output generated by the Sinhalese parser and construct the bilingual web page using the preserved structure of the original web page. In this sense, HTML parser is associated with both English and Sinhalese parsers from the input and the output ends of the system. Note that BEES can also be used as a bilingual translator, without the HTML parser module, for translating English texts into English-Sinhalese bilingual texts. This mode will be useful if we want to translate a text document such as a word file into bilingual format. This is why we have decided to developed HTML parser as a separate module in BEES, enabling bypassing of it from the main system when necessary. Since the Semantic Web is more than web pages, we argue that it is appropriate to design the BEES in the above manner.

A simple parser has been developed in BEES for English language. We have purposely designed a simple parser rather than using a Standard English parser due to the following reason. Currently available English parsers are too complex and too general. However, we wanted an English parser, which is more specific and comparable with our Sinhalese parser so that bilingual translation will not be affected by the complexity of parsers.

4.2 English Parser

We argue that this philosophy is very appropriate since English and Sinhalese languages have some fundamental differences in the way that they handle concepts such as prepositions. We discuss such issues under the description on Sinhalese parser in Section 4.3 below. SWI-Prolog has been used for developing both Sinhalese and English parsers. These parsers work on the basis of paragraphs rather than one to one translations at the word level. As the first step, parsers have been developed to support syntax-based translations of simple grammatical structures. The current ontology of two languages must be extended to incorporate semantic manipulation as the next step. However, our current system itself is adequate to show how the bilingual concept works in practice.

4.3 Sinhalese Parser

Development of a Sinhalese parser is a relatively new project in Sri Lanka. Although some attempts have been made to developed simple and specific parser for the Sinhalese

language, a reusable product has not been developed yet (Vithanage, 2003). Our parser has addressed Sinhalese-language specific concepts such as *vibakthi* for handling the concept of preposition in English languages. Grammar rules for construction of Sinhalese sentences are dependent on the noun forms derived using the *vibakthi* rules in Sinhalese. There are nine *vibakthis* in Sinhalese and we have developed *vibackthi* rules for manipulation of Sinhalese words. Otherwise we have to store all the associated forms of each noun in a table, which will be impractical to implement and also causes to reduce the efficiency of the entire bilingual translator. With this approach the dictionary contains only the base form of each noun.

The Sinhalese parser reads the words, which are tokenised by the English parser, and finds the equivalent Sinhalese base word from the dictionary, and then derives the appropriate noun form required by the relevant grammar rule in the Sinhalese parser. In general noun cannot be treated in isolation, but with the prepositions attached to the noun. For example, English phrase "cut with a knife" has the Sinhalese meaning similar to "cut from a knife". This is why we argue that bilingual parsers should be developed with the understanding of both languages considered. However, with the use of bilingual concept we are free to use certain amount of English words as they are, in the translated document. This allows the reader to comprehend the material with the help of one of the two languages, which brings more meaning to the user. In this sense, we argue that bilingual approach also allows relaxing some inherent constraints associated with natural language translations. Note that the philosophy behind the bilingual approach is to support the better comprehension of a material by the user rather than generating a hundred percent technically accurate translation. It is pointless to talk of technical correctness, if the comprehension cannot be supported. Although both aspects should go together, we are more interest in machine support for comprehension of a particular material in a second language. We believe that the technical correctness will emerge as a result of proper comprehension.

Sinhalese parser has also been written in SWI-Prolog. The parser internally uses Sinhalese words written in English alphabet as a means of avoiding font related issues within the Prolog compiler. Manipulation of Sinhalese font within Prolog is very complicated since Sinhalese fonts use some restricted characters for Prolog language. Since this internal representation is not visible to users of BEES, it is possible to maintain a standard within the system to avoid issues with writing Sinhalese words using the English alphabet. HTML parser undertakes the Sinhalese font manipulation at the time of displaying the bilingual texts on the web browser. It is essential to use Sinhalese fonts to write Sinhalese words for generating the bilingual display. The Sinhalese words using the English alphabet do not give any bilingual feeling to the user.

4.4 Bilingual translator

Bilingual translator is basically the expert system of BEES. This component reads the English sentences analysed by the English parser and try to find the corresponding words from an English-Sinhalese dictionary. The words in both languages are written using the same English alphabet and this is only for the internal use. As we mentioned this avoids internal issues pertaining to font manipulation. Bilingual translator handles the key features of the system including the use of English words as they are, when corresponding Sinhalese words are not available or inappropriate depending on the user profile. User access to the BEES has also been implemented as an interface to bilingual translator.

Using the expert system technology we expect to improve the usability and customisation of BEES. Our BEES is necessarily an interactive program. Among other features, expert system handles the percentage of the use of Sinhalese words in the final display depending on the user profile. If many translation sessions use lots of English words as they are, the expert system may recommend language experts to improve the dictionary by adding more Sinhalese words. In this manner, BEES evolves as we use it. Inherent features of expert systems such as explanation ability, provision of alternative answers and handling certainty of answers will also be incorporated in BEES. For example, if BEES is unable to

translate a sentence, it can tell why the translate process has been unsuccessful. The current version of BEES has limited set of expert system features.

5. Discussion

This papers has presented our approach to make world is more accessible for majority of Sri Lankans through bilingual strategy to development of an English to Sinhalese bilingual Expert systems as a web client. We explained the emerging trend of the use of power of mother tongue for comprehension of material in a second language. We presented the design and implementations of BEES system as a means of showing how the bilingual approach can work in practice. Although, the current version of BESS has limited capabilities, it can prove the power of bilingualism. We are currently working on the further improvements to BEES in such way that it will be capable of handling more semantics during the process of bilingual translation. It should also be noted that BEES can also be readily used as a bilingual systems for translating of English text files such as Word documents, PDF files and downloaded ftp files. In fact, BEES has become little more complicated due to the involvement of HTML and Web technology within it. Obviously, BEES can be used as a general purpose English to Sinhalese bilingual translator. As such BEES opens more opportunity for reading and comprehending English documents through the power of mother tongue.

While improving the BEES as a better system for majority of Sri Lankan to access the world knowledge, we intend to propose bilingualism as a novel approach for ontology modelling for Semantic web. In this sense, as analogous to human beings, we postulate the idea that machines can also communicate in bilingual manner. As recent researches have shown, bilingualism can be extended leading to multilingualism too. As such, we argue that bilingualism will be able to use for providing a new means for communication and sharing of knowledge in systems such as heterogeneous Multi-Agent Systems on the Semantic Web. In line with this, we believe that research into bilingualism through our BEES, will reveal some finding for supporting Semantic Web research. Since the researches into Semantic Web are currently very intuitive-based, on the same groundings such as extensions to XML and object-orientation, without a novel direction, we believe that bilingual concept will provide yet another theoretical basis for Semantic web research.

6. References

[1] Tim Berners-Lee, James Hendler, Ora Lassila (2001), *The Semantic Web*, Scientific American

[2] Lenat Douglas B. (1995), *Cyc: A large-scale investment in knowledge infrastructure,* Communications of the ACM, 38(11), 31-40

[3] Miller George (1995), *WordNet: A Lexical database for English*, Communication of the ACM, 38 (11), 39-41

[4] Yokoi Toshio (1995), *The EDR electronics dictionary,* Communications of the ACM, 38(11), 42-44

[5] Vithanage N.V.C.T. (2003), English to Sinhala Intelligent Translator for Weather forecasting domain, Thesis submitted BIT degree, University of Colombo.

[6] Williams, C. (2000): "Bilingual teaching and language distribution at 16+" . International Journal of Bilingual Education and Bilingualism 3(2). Multilingual Matters Ltd. Clevedon: pp 129-148.

[7] Una Cunningham-Andersson & Staffan Andersson (1999), Growing up with two languages: A practical Guide, Routledge