A first step design in integrating an English-Malay Translation Memory System into the Semantic Web

Suhaimi Ab. Rahman
Department of Software Engineering, College of Information Technology
UNIVERSITI TENAGA NASIONAL
KM7, Jalan Kajang-Puchong, 43009 Kajang, Selangor
MALAYSIA
smie@uniten.edu.my

ABSTRACT

This paper discusses on the first step designing integration Semantic Web into our English-Malay Translation Memory (TM) prototype system. The main research activity that we need to perform is on how we can design and construct the ontology database in the architecture of the semantic web environment. Research also need to be carried out by understanding precisely the layer-cake architecture of semantic web, mainly in designing and implementation interaction between English-Malay TM system via a semantic web.

Keywords
Semantic Web, Ontology, Translation Memory (TM), Database.

1.0 INTRODUCTION

The Malay language is spoken by over 300 million people and considered the fourth largest language group as described in (Arena Online, 2007). According to Dewan Bahasa dan Pustaka (DBP)[1], there are 118 institutions throughout the world that teaching and pursuing research on Malay language. There is a need of electronic language tools to carry out research on Malay to match increased research activity in universities in Asean countries, France, Britain, the Netherlands and Australia where Malay is taught and researched. Several established Language Technology (LT) tools are available in the market for other major languages of which some are generic. These tools can be tailored for Malay language, but such customisation is only a short-term solution. This customisation is not only costly in terms of both time and effort, but also the issue of intellectual property and rights (IPR) associated with the product may hinder further development.

There is also high demand, especially from translators, for a better tool that can automatically translate English to Malay text efficiently. To meet this demand and to provide a better solution to the above problems, a prototype of the English-Malay Translation Memory (TM) system has been developed. This tool can be used to translate documents such as company annual reports, manuals, brouchers and others in a shorter time. Currently, there is no dedicated English to Malay TM system available in the market. There are TM systems such as Trados' Translator's Workbench, SDl, IBM Translation Manager/2 and Transit but they need to be customised. Customization of tools has its own constraints as described earlier.

The rest of this paper is organised as follows: Section 2 discussions about the semantic web. Section 3 provides an overview of the English-Malay TM system. Section 4 describes the prototype project in more detail. In Section 5, this paper briefly describes the integration between English-Malay TM system with a semantic web. Finally, conclusion and future work are presented in Section 6.

2.0 THE SEMANTIC WEB

The next generation of the Web is often characterized as the semantic web. The information will no longer only be intended for human readers, but also for processing by machines, enabling intelligent

[1] Dewan Bahasa dan Pustaka (Malay for The Institute of Language and Literature) (abbreviated DBP) is the government body responsible for coordinating the use of the Malay language in Malaysia and Brunei.

information services, personalized Web-sites, and semantically empowered search-engines. Following the definition of Tim-Berners-Lee, “The Semantic Web will bring structure to the meaningful content of the web pages, creating an environment where software agents roaming from page to page can readily carry out sophisticated tasks for users” (Berners-Lee and Hendler and Lasilla, 1999)

The current Web can be characterized as the second Web generation: the first generation Web started with handwritten HTML pages; the second generation made the step to machine generated and often active HTML pages. These generations of the Web were meant for direct human processing (reading, browsing, form-filling). The third generation Web, which one could call the "Semantic Web". The main aim of Semantic Web is to enrich documents with semantic information about the content and to develop powerful mechanisms capable of interpreting this information. These goals are achieved through implementation of models, standards as well as annotation of resources at the following layers (Berners-Lee, 2003) presented in Figure 1:

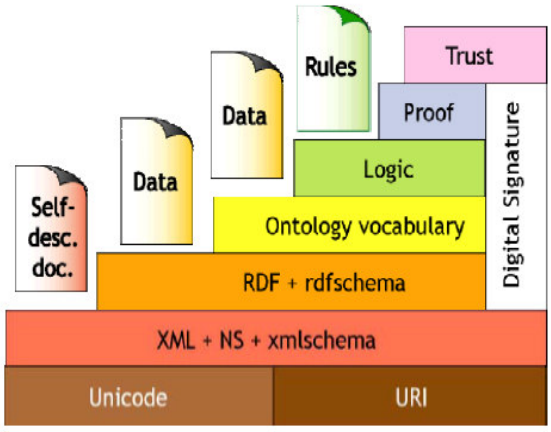

Figure 1: Semantic Web Architecture

We describe briefly these layers.

- **Unicode and URI:** Unicode, the standard for computer character representation, and URIs, the standard for identifying and locating resources (such as pages on the Web), provide a baseline for representing characters used in most of the languages in the world, and for identifying resources.
- **XML:** XML and its related standards, such as Namespaces, and Schemas, form a common means for structuring data on the Web but without communicating the meaning of the data. These are well established within the Web already. However XML tags cannot describe contents of documents. Therefore RDF (Resource Description Framework) model has to be used, and the concepts used for semantic description have to be organised in ontologies.
- **Resource Description Framework:** RDF is the first layer of the Semantic Web proper. RDF is a simple metadata representation framework, using URIs to identify Web-based resources and a graph model for describing relationships between resources. Several syntactic representations are available, including a standard XML format.
- **RDF Schema:** a simple type modelling language for describing classes of resources and properties between them in the basic RDF model. It provides a simple reasoning framework for inferring types of resources.
- **Ontologies:** a richer language for providing more complex constraints on the types of resources and their properties.
- **Logic and Proof:** an (automatic) reasoning system provided on top of the ontology structure to make new inferences. Thus, using such a system, a software agent can make deductions as to whether a particular resource satisfies its requirements (and vice versa).
- **Trust:** The final layer of the stack addresses issues of trust that the Semantic Web can support. This component has not progressed far beyond a vision of allowing people to ask questions of the trustworthiness of the information on the Web, in order to provide an assurance of its quality.

3.0 OVERVIEW OF ENGLISH-MALAY TM SYSTEM

Translation memory, as the name implies, “memories” the translation previously made. Most translation memory systems (often also called “TM systems”), consist of a database that stores the original text along with its translation – a database of segment pairs.

There are three basic components in our English-Malay TM System: input, process and output. For the input process, the user is allowed to input sentence or paragraph into the text-box area provided by the system. After inputting process is done, TM engine will filter the input sentence or paragraph by eliminating unrecognised characters, symbols or punctuations. During translation, the tool will

automatically look up the source language segment to be translated in the existing bilingual translation memory. The first step for the searching process is to identify matching text using exact text match. Referred to by (Bowker, 2002), the exact match is 100 percent identical to the segment that the translator is currently translating, both linguistically and in terms of formatting. This means that the two strings must be identical in every way, including spelling, punctuation, inflection, numbers, and even formatting (e.g., italics, bold). If the exact same segment is found, the system will suggest it to the user. Otherwise, the system will start looking for closely similar segments through an approach that we refer to as "fuzzy-matching". Using this approach, the system try to calculate a word-based "edit distance", that is, the widely used measure of string similarity which counts the minimum number of substitutions, insertions and deletions needed to change one string into another as discussed in (Suhaimi et al., 2006).

When the system is suggested with a list of possible translation from the TM database, the translator can decide which of those that can be reused for the current translation. They can also set the level of "fuzziness", i.e. the similarity percentage, so that the system will only offer translations that can reused without having to make too many changes to the suggested translation. The overall architecture of our English-Malay TM system is depicts in Figure 2.

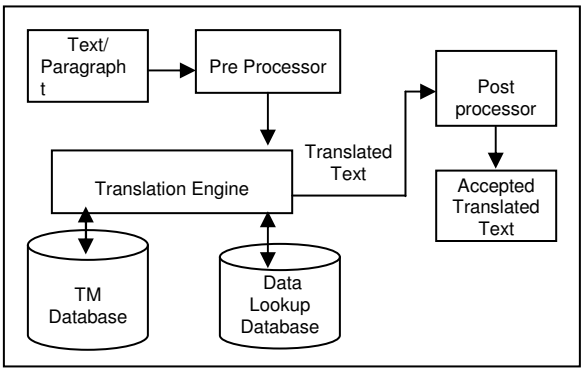

Figure2: Architecture of English-Malay TM system

4.0 ARCHITECTURAL DESIGN

Our architectural design as in Figure 2 shows how the data flow of English-Malay TM system is executed from one component to others. The functional independence method in (R. S. Pressman, 2005) is used to design our system. The purpose of this method is to reduce coupling and increase cohesion between connections of modules. Since, the module is designed modularly, it is easier to develop because function may be compartmentalized and interfaces are simplified. Moreover, it allows easier maintenance and testing, reduces error propagation, and allows reusable of codes.
The detail of each module in Figure 2 is explained in the following sub-sections.

4.0.1 Pre-processing Module

The purpose of this module is to read out the text/paragraph entered by user. In general, this pre-processing module consists of three main modules: text recognizing, tokenization and word filtering. The function of each module is as follows:

4.0.1.0 Language Identifier

The text/paragraph to translate should be clear and accurate so that further processing on the text can be carried out accordingly. To some degree, a bad original text will result in a bad translation. In order to recognize the source text is in English, one module is applied to check the number of word similarity between English sentence and Malay sentence from TM database. If a return value for the input sentence is greater than 50%, the system probably identified the source text entered as English text and is now ready for further processing.

4.0.1.1 Tokenising

The pre-processed text is tokenised. Thus, the multiword units of text are tokenised into single-word units. The single-word units will be assigned into a vector list or array. Once the word is in the array lists, it can be manipulated easily, using various *java string* operations such as java.substring(), java.countTokens(), java.nextToken(), and a few others as in (Sun Microsystems, Inc. 1995-2001), to find word meaning from our TM database.

4.0.2 TM Database

The English-Malay sentence pairs, which have been edited and verified are stored in TM database. The system will examine the new input sentence added into database. If it is similar to any of the existing text in the database, the system will reject it. This is to avoid duplication records.

In our system, the TM database itself can be constructed by the following methods. The simplest method, though the most time-consuming one, called "interactive translation" as mentioned in (Bowker, 2002) is to build a TM from scratch, that is, to store in the memory each sentence as it is translated. A second method, "post-translation alignment", and much heralded by manufacturers, is to extract a TM from an already translated text by *aligning* the source and target texts. Post-translation alignment allows translator to "bulk-up" their TMs and thereby increase the probability of getting a match. Once information has been entered into the TM database, either through interactive translation or by post-translation alignment, the translator has a TM available for use when undertaking new translations. As described by (O'Brien, 1998), a Translation Memory is always more accurate when it has been created by interactive translation as opposed to automatic alignment, but alignment can produce a reasonably accurate translation memory which can be used as a start-up.

Our TM database also contains dictionary table, terminology table and phrases table to store all the relational logical data required by the system. To populate the translation memory, we used an English-Malay parallel corpus of financial domain, economic domain and health domain consisting 30,000 pairs of sentences. We enriched the corpus by adding alignment annotation through our word alignment tool[2] as discussed in (Suhaimi et al., 2004). The aligned parallel text was verified by a group of linguist for better quality of parallel corpus as described in (Rogayah et al., 2007). Some information on the corpus is as in Table 1.

Table 1: Number of sentences and tokens in the TM database

Language	English	Malay
No. Sentences	30,000	30,000
No. Tokens	480,000	630,000
Average word/sentence	16	21

[2] This is one of our tools to auto- generate word alignment output in our Bilingual Knowledge Base (BKB) construction process.

4.0.3 Data Lookup Database

Our data lookup database has a set of logical database such as terminology, phrase and dictionary that will be used by our TM engine to retrieve only the needed words from the unknown word or phrase retrieved from the TM database. The translation memory database and the terminology database work together during translation. The translator will not only be suggested of the translation of whole segments but also a list of all the terms within that segment that were found in the terminology database.

4.0.4 Post processor

When translation is completed, a post-process is done against its original format. The purpose of this module is to allow the translator to edit, add or reject the translated results. With such a feature, any newly translated sentence together with its translation, can be learned and added into the TM database. This allows the TM database to grow dynamically during the translation process. As the number of entry grows, the system is able to produce better translation suggestions to its user.

5.0 PROPOSED A NEW ARCHITECTURAL DESIGN OF ENGLISH - MALAY TM SYSTEM BASED ON THE SEMANTIC WEB ENVIRONMENT

Now Semantic Web applications are calling for different infrastructure software to support essential ontology operations, such as ontology persistence, ontology consistency, ontology query, ontology management, reasoning and so on. As an ontology platform using the standard of RDF, the key characteristic of the Semantic Web Platform is that the platform has a "total solution" for the semantic web applications built on an expansible, flexible, scalable and open architecture.

5.0.1 Semantic Web Architecture of English-Malay System

The prototype version of English-Malay system is already developed. The methods and techniques used for handling a translation memory is still remain. The different on this version is integrating with the ontology database to store related information requires for the system. Ontologies (often also referred as Domain Model) can play a crucial role in

enabling the processing and sharing of knowledge between programs on the Web. The first design of our ontology is only focuses more on developing vocabulary that describes the meaning and relationships of terms. The overall of the new integration architecture for this system depicts in Figure 3:

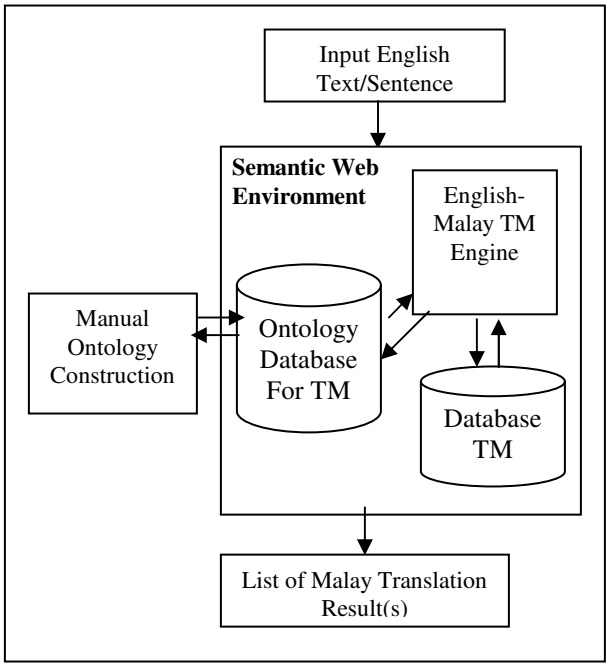

Figure 3: Integration English-Malay TM System with new Semantic Web environment

Referring to the above Figure 3. The translation process is begin by determine what type of the sentences in English that need to be translated. The ontology database is used as the corresponding fields or terms to search related information used for the translation. After the related information retrieved from the databases (Ontology database for TM and Database TM), then it will goes to the other next process in English-Malay TM engine to produce a list of Malay translation results. The accuracy of the Malay translation memory results is depending on the total numbers of records stored. The way that we can use to create and populate our new record into ontology database is using Proteje 2000 tool.

Based on this new integration process, generally we can divide the process into four categories:

Category one: Alignment Process

The alignment process is done manually by collecting bilingual parallel corpus of English-Malay pair of sentences. As a result we have an aligned database of examples, by linking the Malay sequence of words with their equivalents in English.

Below are some of examples of sentence alignment between English(E) and Malay(M) pairs of sentences:

E:The pill is generally safer for healthy women.
M:Pil itu secara umumnya adalah selamat bagi wanita-wanita yang sihat.

E:The pill is suitable for non-smoking women over the age of 35.
M:Pil itu adalah sesuai untuk wanita-wanita yang tidak merokok melebihi umur 35.

E:The recommended fat intake is 65gm a day.
M:Pengambilan lemak yang disyorkan ialah 65gm sehari.

E:This project is a classical attempt to merge both ICT and biotechnology together.
M:Projek ini adalah suatu usaha klasik bagi menggabungkan kedua-dua ICT dan bioteknologi bersama-bersama.
:

Figure 4: Examples of alignment in English-Malay sentences

Category Two: Translation

The translation of the fields is made automatically by first consulting the ontology and then searching in the database of examples for the appropriate translation of the field. The speed of translation process is depending on the number of sentences from the database that need to be referred. If the is more than similarity of word matching from the searching process found. The system need to collect more results and store it as a temporary data into array of results list before display to the users.

Category Three: Matching

The process of finding the right translation is considered to be the matching algorithm. In this system, the matching algorithm includes matching the

Malay field with the English field from database examples.

For the matching process, we have also face problems in determine the corresponding of ONE English word to MANY Malay word, MANY English word to ONE Malay word and MANY English word to MANY Malay word, as discussed in (Suhaimi et al., 2004).

Category four: Recombination

This is where the last step comes in: the target-language fragments suggested by the examples then have to be reassembled or *recombined* to form the target text. First of all because this kind of system includes whole sentences or phrases to be translated. These sentences are divided into words that are later searched and having been translated previously, must be recombined into a new sentence in target language. This is made because different languages have different morphology, different word order.

6.0 CONCLUSION AND FUTURE WORK

In this paper we presented the first step made by us to prepare an English-Malay TM system to be integrated into the Semantic Web. We also described briefly the development of English-Malay TM system.

The application has its limitations and these are the start point for further work. First of all, we would extend the database of examples, by adding more examples for each domain. Some further work should be done for the further integration of the system into the Semantic Web. So far, we have designed the ontology of vocabularies structure, and we plan to use RDF as a knowledge representation for annotate all the documents in the corpus. As a further work too, we want also to implement vice versa of English-Malay TM system, such as Malay-English TM system.

REFERENCES

Arena Online, (2007). Malay/Indonesian for an Official Language of the East Asian Community, http:// www. arenaonline. Org / content/ view / 305/ 151.

Bowker, L. (2002). Computer-Aided Translation Technology. A Practical Introduction, Ottawa: University of Ottawa Press.

O'Brien, S. (1998). *Practical Experience of Computer-Aided Translation Tools in the Software Localization Industry, in L. Bowker, M. Cronin, D. Kenny and J. Pearson (eds) Unity in Diversity? Current Trends in Translation Studies*, Manchester: St Jerome, pp. 115–122.

Rogayah A. Razak et.al. (Nov 2007). *Corresponding Grammatical Analysis of Parallel Text English-Malay in a Machine-aided Translation Software, In Proceedings of the 11th International Conference on Translation – Enriching the Repository of Knowledge in Malay*, Kuala Lumpur.

R. S. Pressman, (2005). *Software Engineering A Practitioner's Approach,Sixth Edition*, McGraw Hill Higher Learning, Boston, 2005.

Suhaimi Ab. Rahman, Normaziah Abd Aziz, (2004) Improving Word Alignment in an English-Malay Paralell Corpus for Machine Translation, *In Pre-LREC 2004 Workhop on Amazing Utility of Parallel and Comparable Corpora*, Lisbon, Portugal.

Suhaimi Ab. Rahman, Normaziah Abdul Aziz, Abd. Wahab Dahalan, (2006). Searching Method for English-Malay Translation Memory based on combination and reusing word alignment information, *In Language, Artificial Intelligence and Computer Science for Natural Language Processing Applications (LAICS-NLP),* Bangkok.

Sun Microsystems, Inc. (1995-2001). *JavaTM 2 SDK, Standard Edition Documentation*, http://java.sun.com/j2se/1.3/docs/.

T. Berners-Lee, (2003). Foreword to *Spinning the Semantic Web: Bringing the World Wide Web to Its Full Potential,* in D. Fensel, D., J. Hendler,, H.Lieberman, and W. Wahlster, (eds.), MIT Press.

T. Berners-Lee, and J. Hendler, and O. Lasilla, (1999). *The Semantic Web, Scientific American*.