INTERNATIONAL JOURNAL OF ADVANCE RESEARCH AND DEVELOPMENT

*(Volume 4, Issue 3)*

Available online at: www.ijarnd.com

# An effective approach of semantic analysis to natural language translation system (English-Myanmar language)

**Tin Htar Nwe[1], Phyo Phyo Wai[2]**

[1]*Professor, University of Computer Studies, Magway, Myanmar*

[2]*Lecturer, University of Computer Studies, Magway, Myanmar*

**ABSTRACT**

***English is one of the most widely spoken languages in the world these days. Most of the commercial websites are also being designed in the English language. So, we need a language translation system to understand all websites and all information from the internet. This paper is one part of the English-Myanmar Machine Translation system. In our system, the tagged English text is accepted as input. Firstly, these tags are tokenized. Each phrase is tokenized. The structure is basically a list of tokens. The linguistics clues are normalized. And then the dictionary lookup is performed. The guessing process is applied. Since the files are sorted according to alphabetically, required information can be searched quickly with a binary search. The dictionary-based translation technique produces one or more translation terms in the target language for each term in the source language. We propose a method called word sense disambiguation to solve the ambiguity of words. By using this technique and bilingual lexicon, the system may retrieve correct translated words.***

***Keywords***— *Language translation, Machine translation, Word sense disambiguation, Bilingual lexicon*

## 1. INTRODUCTION

Processing of human languages is actually one of the arguments of greater interest in the field of Artificial Intelligence. Major concerns our understanding of the natural languages and perform multi-dimensional processing as information retrieval, text summarization, scenario understanding, machine translation, grammar inference, document classification, language translation and so many other applications [18].

Through the past centuries, a huge amount of documents have been written and translated manually in different languages. Some parallel texts are stored electronically and most of the new multilingual corpora will be stored using computers [17]. Recently, deep linguistic processing, which aims to provide a useful semantic representation, has become the focus of more research, as parsing technologies improve in both speed and robustness.

The lexicon can be used to automatically translate on a word by word basis and in some cases phrase by phrase. There are instances in all languages where a phrase may not make sense when broken down word by word into another language.

There are many related works on lexicon modelling in recent years. We require the availability of the examples from which to learn. The types of examples these methods require are parallel texts that have been manually and decide which words are a translation of each other and then indicate this somehow in a form that a computer program can utilize.

An important requirement for machine translation is the existence of a bilingual lexicon containing large sets source-language/ target-language correspondences. A translation system or online dictionary can be used to identify good documents for translation. For the purpose of multilingual document retrieval, such a search engine must have access to a (bilingual or multilingual) dictionary to translate queries (or indexes). In this paper, we construct a semantic analysis architecture for English to Myanmar language translation system.

## 2. RELATED WORK

A statistical Translation Model (TM) is a mathematical model in which the process of human language translation is statistically modelled [11]. Model parameters are automatically estimated using a corpus of translation pairs. TMs have been used for statistical machine translation (Berger et al., 1996), word alignment of a translation corpus (Melamed, 2000), multilingual document retrieval (Franz et al., 1999), automatic dictionary construction (Resnik and Melamed, 1997), and data preparation for word sense disambiguation programs (Brown et al., 1991). Developing a better TM is a fundamental issue for those applications [10].

Researchers at IBM first described such a statistical TM (Brown et al., 1988). Their models are based on a string-to-string noisy channel model. The channel converts a sequence of words in one language (such as English) into another (such as French). The channel operations are movements, duplications, and translations, applied to each word independently. The movement is conditioned only on word classes and positions in the string, and the duplication and translation are conditioned only on the word identity.

Mathematical details are fully described in (Brown et al., 1993). One criticism of the IBM-style TM is that it does not model structural or syntactic aspects of the language. The TM was only demonstrated for a structurally similar language pair (English and French). It has been suspected that a language pair with very

 

different word order such as English and Japanese would not be modelled well by these TMs [11].

The reorder operation is intended to model translation between languages with different word orders, such as SVO (Subject, Verb, Object) -languages (English or Chinese) and SOV (Subject, Object, Verb) -languages (Japanese or Myanmar or Turkish). The word-insertion operation is intended to capture linguistic differences in specifying syntactic cases. E.g., English and French use structural position to specify case, while Japanese and Korean use case-marker particles. Wang (1998) enhanced the IBM models by introducing phrases, and Och *et al.* (1999) used templates to capture phrasal sequences in a sentence. Both also tried to incorporate structural aspects of the language, however, neither handles nested structures [14].

Wu (1997) and Alshawi *et al.* (2000) showed statistical models based on syntactic structure [13]. The way we handle syntactic parse trees is inspired by their work, although their approach is not to model the translation process, but to formalize a model that generates two languages at the same time. Following (Brown et al., 1993) and the other literature in TM, this paper only focuses only on one part of the details of TM. This is Source Target transfer part. Applications of our TM, such as machine translation or dictionary construction, will be described in a separate paper.

## 3. DESCRIPTION OF THE MACHINE TRANSLATION ARCHITECTURE

Firstly, I would like to introduce the whole translation process. This is called the architecture of Machine Translation. This architecture is shown in figure 1.

In this architecture, source language text (English) is accepted as input. These texts are analyzed by some parser (example link grammar, tree tagger or Gate) in source text analyzing phase. The output of this phase is tagged in English texts. They are then transferred into corresponding Myanmar texts by using English-Myanmar Bilingual Lexicon and Word Sense Disambiguation algorithm. This second phase is called Source-Target Transfer phase. This phase is the major description of this paper. Then Myanmar texts are generated as Myanmar Language texts using Myanmar Lexicon. This phase is the final one and is known as Target text generation. The final output is the target language text (Myanmar). With this architecture, our proposed system concerns with only source target transfer process.

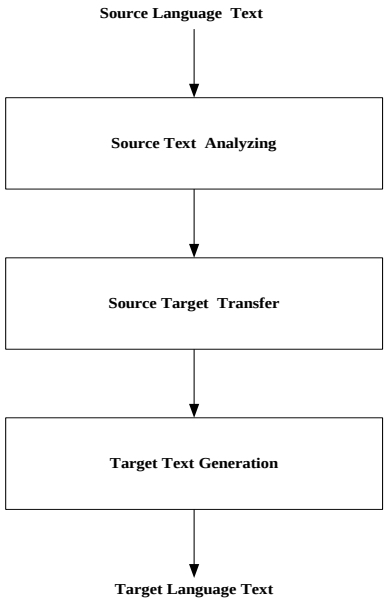


**Fig. 1: Main components in machine translation**

## 4. COMPONENTS IN PROPOSE ARCHITECTURE

Figure 2 illustrates the major steps in my propose architecture. The detail description of each component is as follows:

### 4.1 Tokenization process

In this process, the input is English (tagged) sentence. We need to tokenize input tags to find easily in appropriate tables from the bilingual lexicon. The structure is basically a list of tokens. The linguistics clues are normalized. To more clearly, I would like to describe with an example.

Example: I want to read the best book.

[ I ] [ want ][ to read ][ the best book ]
1 2 3 4
(These are chunk numbers)

The sentence is tagged and numbered the chunk during the Source text analyzing process. Each chunk is separated by a square bracket. After tokenizing, we get the following information for a chunk.

PRED I/NC/STRT-STP/PP/I ( chunk type)
IDX 1 ( Index number )
FUNC SUBJ ( Function of the word )
ARG NULL(functional relation )
This information is for 'I'.

PRED want/VC/STRT-STP/VBP/want
(root word)
IDX 2
FUNC ACTIVE
ARG 1, 3(functional relation to chunk 1 and 3)

This information is for 'want'.
And other information for each chunk may be obtained.
After tokenization, we get part of speech, root word and their arguments for each chunk.

### 4.2 Performing dictionary lookup

To do this process, first of all, we need to construct Bilingual lexicon. So, I would like to describe the creation of lexicon briefly.

**4.2.1 Creation of Bilingual Lexicon:** This lexicon contains information about nouns, verbs, adjectives, adverbs and some phrases in English-Myanmar languages. All dictionary information for entries is structured in attributes:

- Head-word as in English
- Part of speech
- Myanmar meaning
- Definition
- Morpheme: ~ suffix, ~prefix
- Frequency of word
- Compound Nouns
- Example sentence

In developing our lexicon, it has been convenient to divide the work into two interdependence tasks. One task has to construct the source files that contain the data. The second task was to create a set of computer programs that would accept the source files and do all the work leading ultimately to the generation of a display for the user. We construct Noun, Verb, Adjective and Adverb tables. The database is in a Unicode format that is human and machine readable. Each file is in an alphabetized list of all of the word forms in the dictionary. The files are sorted alphabetically.

 

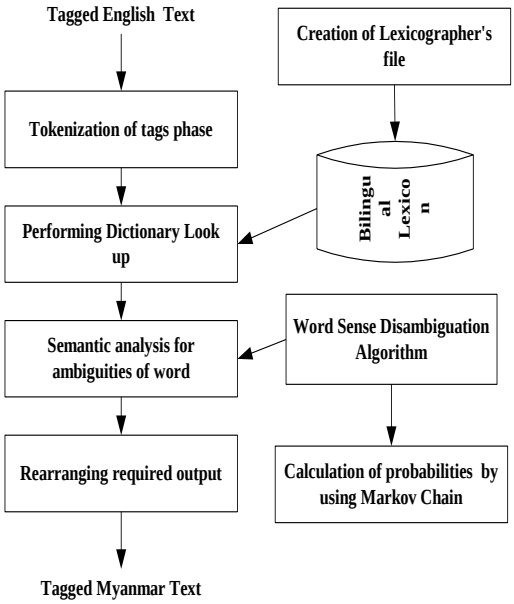


**Fig. 2: Architecture of the proposed system**

After creation of the lexicon, the guessing process is applied. If the tags are CD,CC,FW, IN,NN, NNS, NP, NPS, POS, PP, PP$, SYM, UH, WDT, WP, WP$ and DT, we will search in noun table. If the tags are JJ.JJR, JJS and PDT, we will search in adjective tables. If the tags are RB, RBR, RBS and WRB, we will search in Adverb table. If the tags are MD, RP, VB, VBD, VBG, VBN, VBP and VBZ, we will search in the verb table. The meanings of abbreviation are shown in the appendix. Since the files are sorted according to alphabetically, required information can be searched quickly with a binary search.

### 4.3 Semantic analysis for the ambiguity of words

In any application where a computer has to process natural language, ambiguity is a problem. Many words have several meanings or senses. For such words given out of context, there is this ambiguity about how they are to be interpreted. A Word Sense Disambiguation technique is used. Word sense disambiguation is the process of identifying the correct meanings of words in particular contexts. Word sense disambiguation has been a research area in Natural Language Processing for almost the beginning of this field. It is known that whenever a system's actions depend on the meaning of the text being processed, disambiguation is beneficial or even necessary. The most important robust methods in word sense disambiguation are machine learning methods and dictionary-based methods. For our application, we used the dictionary based approach. Selecting the right word translation among several options in the lexicon is a core problem for machine translation.

The dictionary-based translation technique produces one or more translation terms in the target language for each term in the source language. The sense disambiguation process as follows:

1. Obtain the translation terms of the given English term from the dictionary.
2. Calculate probabilities of the word by using Markov chain for the same English term.

An approach to estimate word translation probabilities is to use the frequencies of the translation word.
We may obtain the counts as follows:

| count | translation |
|---|---|
| c1 | E1 |
| . | . |
| . | . |
| . | . |
| cn | En |

We now calculate the translation probabilities.

$$Pw = c1/( c1+c2+\ldots..+cn) \quad (1)$$

3. Select the translation term from terms obtained in step 1 that has the highest value in the entries obtained in step 2.
4. If the English term is not found in the dictionary then it is taken without translation

Some words are more probably before or after some other words. We need to find these probabilities. We may guess the word if we know the nearby ones.

Let's a sequence of states be $x_i$.
The sequence $\{x_n\}$ is said to be a Markov chain.

$$P(x_n,x_{n-1}\ldots..,x_1)=P(x_n /x_{n-1}) \quad (2)$$

Since the number of states is finite, we use a finite state Markov chain.

$$P = P(X_1) \quad \prod_{i=1}^{n-1} P(Xi+1/ Xi) \quad (3)$$

### 4.4 Rearranging required output

Since the tags are separated one by one, we need to collect these tags. By rearranging these tags, it is more convenient to process in the next step. We rearrange the tags according to input tags. We labelled the chunk numbers and rearranging process is done according to this order.

## 5. CONCLUSION

We have described the text translation method for structurally different languages. We describe the overview of semantic analysis to the translation system. We proposed a worse sense disambiguation algorithm for our system. By using this algorithm and lexicon, may retrieve the correct translated word.

## 6. ACKNOWLEDGEMENT

This research is supported by UCSY (University of Computer Studies, Yangon). UCSY is one of the higher institutions in the Ministry of Science and Technology. It is to conduct teaching and research in various branches of computer science and technology.

## 7. REFERENCES

[1] M. S.C. Thomas and K. Plunkett "Representing the bilingual's two lexicons", University of Oxford, UK, URL www.psyc.bbk.ac.uk/people/academic/thomas_m/Thomas_cogsci95.pdf

[2] S. Fujita, F. Bond," A Method of Creating New Bilingual Valency Entries using Alternations", NTT Machine Translation Research Group, NTT Communication Science Laboratories, Nippon Telephone and Telegraph Corporation, URL www.kecl.ntt.co.jp/icl/mtg/members/bond/pubs/2004-MLR-valency.pdf

[3] D. Hiemstra, "Using statistical methods to create a bilingual dictionary", Master's thesis Parlevink Group Section Software Engineering and Theoretical Informatics Department of Computer Science, August 1996

[4] C. D. Manning and H. Schutze (Stanford University and Xerox PARC)," Foundations of Statistical Natural Language Processing", Cambridge, MA: The MIT Press, Volume 26, Number 2, 1999.

[5] J. Carroll, T. Briscoe, A. Sanfilippo, "Parser Evaluation: a Survey and a New Proposal", Cognitive and Computing Sciences, University of Sussex, UK, URL www.cogs.susx.ac.uk/lab/nlp/carroll/papers/lre98.pdf.

 

[6] V. Claveau, P. S´ebillot, **"**Learning Semantic Lexicons from a Part-of-Speech and Semantically Tagged Corpus Using Inductive Logic Programming", Journal of Machine Learning Research, *France*, 4 (2003), pp.493-525

[7] J.Yun Nie, M. Simard, "Using Statistical Translation Models for Bilingual IR", URL clef.isti.cnr.it/DELOS/CLEF/nie-wkshp01.pdf

[8] A. Kilgarriff, "Foreground and Background Lexicons and Word Sense Disambiguation for Information Extraction", Workshop on Lexicon Driven Information Extraction, Frascati, Italy, July 1997.

[9] J. P. G. Mahedero, A. Martinez, P. Cano, " Natural Language Processing of Lyrics", URL www.iua.upf.edu/mtg/publications/ 9d0455-ACM-MM-2005, November 6-11, 2005, Singapore.

[10] J. Pinkham, M. Corston-Oliver, M. Smets, and M. Pettenaro, "Rapid assembly of a large-scale French-English MT system", URL research.microsoft.com/nlp/ publications/MTSummit01-FE system-final.pdf

[11] S.Kwon Choi, H.Min Jung, C.Min Sim, T. Kim, D.In Park, J.Sik Park, K.Sun Choi, "Hybrid Approaches to Improvement of Translation Quality in Web-based English-Korean Machine Translation", URL acl.ldc.upenn.edu/P/P98/P98-1039.pdf

[12] G. Serban, D. Tatar, " UBB system at Senseval3", Third International Workshop on the Evaluation of Systems for the Semantic Analysis of Text, Barcelona, Spain, Association for Computational Linguistics, July 2004.

[13] A. Roberts, "Machine Learning in Natural Language Processing", URL www.andy-roberts.net/misc/ latex/sessions/bibtex/ bib_example_nat.pdf, 16th, October 2003.

[14] S. Young Jung, S. Lim Hong, E. Paek, "An English to Korean Transliteration Model of Extended Markov Window", 18th International Conference on Computational Linguistics (COLING2000), URL acl.ldc.upenn.edu/C/C00/C00-1056.pdf

[15] S.Lawrence, C. L. Giles and S.Fong, "Natural language grammatical inference with recurrent neural networks. IEEE Transactions on Knowledge and Data Engineering, 12(2000): pp 126.140.

[16] W. Daelemans, A.Van den Bosch, J. Zavrel, J. Veenstra, S. Buchholz, and G. Busser, "Rapid development of NLP modules with memory-based learning". In Proceedings of ELSNET in Wonderland (1998), pp 105.113.

[17] "Methods of bilingual lexicon extraction": http://stp.ling.uu.se/~joerg/diplom/node4.html

[18] Veronique Benzaken, Giuseppe Castagna, and Alain Frisch Cduce: "An XML- centric general-purpose language". In Proceedings of the 8th ACM SIGPLAN International Conference on Functional Programming.

 