Constructing English-Myanmar Parallel Corpora

Hla Hla Htay, G. Bharadwaja Kumar, Kavi Narayana Murthy
Department of Computer and Information Sciences
University of Hyderabad
knmuh@yahoo.com, hla_hla_htay@yahoo.co.uk, g_vijayabharadwaj@yahoo.com

Abstract

In this paper we describe our work in constructing an aligned English-Myanmar parallel corpus. Corpora are not available for Myanmar language and our work in developing parallel corpus will also hopefully be very useful in many natural language applications. For aligning the sentences, we use Gale & Church method. We have obtained an alignment accuracy of about 90%. While detecting sentence boundaries is trivial in Myanmar, it is not entirely trivial in the case of English. We also present a generic method for sentence boundary detection based on machine learning techniques for English. We have used a general public license tool called Weka, which has a collection of implemented machine learning algorithms. We have used decision tree algorithms for model building and verification tasks.

1. Introduction

Corpora and other lexical resources are not yet widely available in Myanmar. Research in language technologies has therefore not progressed much. In this paper we describe our efforts in building an English-Myanmar aligned parallel corpus.

A parallel corpus is a collection of texts in two languages, one of which is the translation equivalent of the other. Although parallel corpora are very useful resources for many natural languages processing applications such as building machine translation systems, multi-lingual dictionaries and word sense disambiguation, they are not yet available for many languages of the world. Myanmar language is no exception.

Building a parallel corpus manually is a very tedious and time consuming task. A good way to develop such a corpus is to start from available resources containing the translations from the source language to the target language. Researchers have devised methods to mine the internet for large amounts of parallel corpora automatically with relatively minimal manual labor at high accuracy levels [28]. In this work we have downloaded English-Myanmar parallel texts from news websites, on line magazines and so on. Topics covered include news, sports, bible, stories, health, and religion and so on. Currently our corpus has about 75,000 sentences.

A parallel corpus becomes very useful when the texts in the two languages are aligned. Here we have used the Gale and Church algorithm [1] to align the texts at sentence level. Nearly 90% correctness of alignment has been obtained.

By exploiting parallel corpora, we can generate a statistical machine readable bilingual dictionary. This bilingual dictionary, in turn, also can be applied, for example, in rule based automatic Machine Translation. When translating technical or specialized texts, it is very important to have a bilingual dictionary that tells how certain terms are translated. The ability to automatically create such resources from parallel texts would make it possible to quickly extend a fairly generic bilingual lexicon to a more specialized domain [27]. We are developing English-Myanmar electronic dictionary from our corpus.

Many words in natural languages have multiple meanings. It is important to identify the correct sense of a word before we take up

translation, query-based information retrieval, information extraction, question answering, etc. For example, the word 'function' can be a '*mathematical function*' as well as a '*social gathering or party*'. The process of determining the correct sense of a given word in context is called *word sense disambiguation*. Recently, parallel corpora are being employed for detecting the correct sense of a word. At least we can reduce the number of candidate senses for a word. We can study equivalent *Polysemy* as well as contrastive *Polysemy* by taking advantage of parallel corpora. Ng [29] proposed that *if two languages are not closely related, different senses in the source language are likely to be translated differently in the target language*. Parallel corpus based techniques for word sense disambiguation therefore work better when the two languages are dissimilar. It may be noted that English-Myanmar scores well here.

High quality lexical resources are needed both to train and to evaluate WSD systems [30]. Parallel corpora are one of the promising resources and there is a good deal of interest in applying corpus based methods in WSD within a multilingual frame work [21].

Kaji and Morimoto [22] describe an unsupervised method for word sense disambiguation using bilingual comparable corpora. Diab [21] proposed the idea of finding the correct selection of sense for the target word using the multilingual corpora. Melamed [2] has done word sense disambiguation system using parallel texts. We plan to use our parallel corpus for word sense disambiguation.

Information Retrieval (IR) systems search and retrieve relevant documents based on a user query. Mono-lingual IR systems find documents only in the language of the query. For example, a query in English that includes the term *AIDS* in English will not find possibly relevant information in other languages. Thus, there is a need for *cross-language* IR systems which retrieve relevant documents in a language other than the query language [27]. Nowadays, *Google* like such engines have started embedding *cross-language* IR systems. Parallel corpora will be of use for developing such IR systems too.

Myanmar language is one of the Southeast Asian languages written in Myanmar script. Myanmar is written without necessarily adding spaces between words. However, spaces are usually added between phrases and wherever convenient [25]. Word level alignment is therefore not straight forward. In this paper we describe our work in sentence level alignment only.

2. Preprocessing

In this section, we discuss preprocessing of the downloaded texts. This involves cleaning up the HTML tags. As the texts come from a variety of sources and in a range of formats, it is essential to do this carefully. We have developed the necessary scripts in Perl to handle a wide variety of web pages.

3. Sentence Segmentation

In *Myanmar* script, we have "။"as a unique sentence boundary marker. Therefore segmenting paragraphs into sentences is trivial. In case of *English language*, however, detecting sentence boundary is not entirely trivial. Even though there are explicit sentence boundary markers such as the period, the question mark and the exclamation mark, the same symbols can be used for other purposes. For example a period can also be used as a decimal point in numbers, in ellipses, in abbreviations and in email-addresses. The exclamation mark in the name of a web site *Yahoo!* may not signify a sentence boundary and so is the question mark in *Which?* - the name of a magazine. Although the first letter of the first word in a sentence is capitalized by convention, segmenting into sentences is still not so straight forward as the following examples show.

- She needs her car by 5 p.m. on Saturday evening.
- At 5 p.m. I had to go to the bank.
- It was due Friday by 5 p.m. Saturday would be too late.
- She has an appointment at 5 p.m. Saturday to get her car fixed.

We have developed a generic corpus based approach for sentence segmentation. Our approach does not make use of meta-information, annotations, POS tagging, or pre-compiled lists except for a list of common abbreviations. We describe our approach below after a brief survey of sentence segmentation work.

3.1. Sentence segmentation – a survey

There have been primarily two kinds of approaches to sentence boundary detection. On the one hand there are knowledge based approaches where manually generated linguistic rules and heuristics are encoded in programs that are optimized to perform sentence segmentation accurately for the specific language on hand. There are also generic approaches where sentence segmentation is based on a set of features which are obtained from a training corpus. Within the latter class, neural networks and maximum entropy techniques have also been reported in the literature.

An example for knowledge based approaches is the UNIX STYLE program [3]. This program applies lists of abbreviations and proper names preceding and following potential sentence boundary markers. The program makes a decision based on a few preened heuristics. The error rate of this system was reported to be about 6.3%.

The Alembic workbench [3] proposed a sentence splitting module, which employs over a 100 regular-expression rules written in Flex. These rules made use of an extensive list of abbreviations, which are classified into different groups according to their semantics and the following word expectation. The error rate of this system has been shown to be 0.9% on a portion of Wall Street Journal corpus.

Riley [3] in 1989 reported a decision tree classifier which was trained on 25 million words and it achieved 0.2% error rate on the Brown Corpus. This system used a context of one word to the left and one word to the right of a potential sentence splitting punctuation. The system used features such as length of a word, capitalization, abbreviation and probabilities for a word to appear at the beginning or at the end of a sentence.

Palmer and Hearst [4] proposed in 1997 a system called SATZ, which used part-of-speech distribution for words in the local context of a potential sentence splitting punctuation. These parts of speech were estimated from the unigram frequency of words vs. their syntactic classes with no regard to their capitalization. Then these estimates were fed into a back-propagation neural network and a decision tree induction algorithm called C4.5. The decision tree variant obtained 1.6% error ate on mixed-case texts and 1.9% on single case texts of a subset of Wall Street Journal corpus. The neural net variant achieved 1.5% error rate on mixed-case texts and 3.3% error ate on single case texts. The abbreviation list has a definite effect on the system's performance. When the system operated without its abbreviation list of 206 entries, its error rate jumped to 4.9% on mixed-case texts. This system was then successfully ported to German and French with similar performance.

Reynar and Ratnaparkhi built a system called MAXTERMINATOR [5], which considered suffixes, prefixes and the abbreviation class of words in the immediate local context of a period. It achieved 2.1% error rate on the Brown Corpus and 1.2% error rate on the subset of Wall Street Journal.

Mikheev [5] reported a similar system, called LT FS, but with a sophisticated feature selection mechanism on top of maximum entropy parameter estimation. This system achieved 0.7% error rate on the same subset of Wall Street Journal used by Reynar and Ratnaparkhi and 1.3% error rate on the entire Brown Corpus.

3.2. Our Approach for Sentence Segmentation

Machine Learning algorithms rely on selection of features that adequately characterize the patterns of interest. The task of identifying the features that perform well in a classification algorithm is a difficult one, and the optimal choice can be non-intuitive.

The following were the possible features that we have identified for English sentence segmentation:

- *Delimiter*: We considered the [.!?:;] as potential sentence delimiters in English language.
- *Prefix*: We considered 'prefix' as feature because if a prefix of a sentence boundary delimiter is an abbreviation or (initials etc. in a) proper name, it may or may not indicate a sentence boundary.
- *Suffix*: We considered 'suffix' as feature because if, for example, a suffix of a sentence boundary delimiter is a digit, it may or may not indicate a sentence boundary.
- *After word*: We considered 'after word' as a feature because if the word which comes after a sentence boundary delimiter starts with a lower case letter, it may or may not indicate a sentence boundary.

We have used the British National Corpus (BNC) [24] to identify feature instances for sentence segmentation. BNC Corpus is a 100 million word collection of samples from a wide range of sources. BNC is sentence tagged and contains a total number of s-units of just over 6 million.

At each potential delimiter, we have observed the instances of features that can potentially indicate whether it is a sentence boundary or not. We have used the general public license tool called "Weka" [7] for our training and testing purposes.

Weka is a tool having a collection of machine learning algorithms for data mining tasks. The algorithms can either be applied directly to a given data set or called from our own Java code. Weka contains tools for data preprocessing, classification, regression, clustering, association rules, and visualization. The main features of Weka are: 1) Comprehensive set of data pre-processing tools, learning algorithms and evaluation methods 2) Graphical user interfaces (including data visualization) 3) Environment for comparing learning algorithms.

Since our data is nominal data, decision trees are very much suitable for the task. A decision tree takes as input an object or situation described by a set of properties, and outputs a classification decision. We have employed ID3 and C4.5 Decision tree algorithms for our purpose. The Weka classifier package has its own version of C4.5 known as J48.

ID3 is a simple decision tree learning algorithm developed by Ross Quinlan in 1983 [8, 9]. The basic idea of ID3 algorithm is to construct a decision tree by employing a top-down, greedy search through the given sets to test each attribute at every tree node. Information gain is used as the metric in order to select the attribute that is most useful for classifying a given set.

C4.5 [10, 11] is an extension of the basic ID3 algorithm to address the following issues not dealt with by ID3: 1) Avoid over fitting the data 2) Determine how deeply to grow a decision tree. 3) Reduced error pruning 4) Rule post-pruning 5) Handle continuous attributes 6) Choose an appropriate attribute selection measure 7) Handle training data with missing attribute values 8) Handle attributes with differing costs 9) Improve computational efficiency.

We found that Weka is unstable and crashes for large input data under constraints of memory available. We have therefore divided instances of features collected from BNC corpus into 20 different random sets where each set contains 50000 samples. We have performed 10 fold cross validation. On 10 test samples we used ID3 decision tree and on the remaining 10 samples we have used J48 decision tree method. The results are given in Table 1. It may be observed that an overall performance of about 99% has been obtained. This is comparable to other published results [16].

Table 1. F-measure from the 20 test samples

Methods			
ID3		J48	
sample	F-measure (%)	sample	F-measure (%)
1	98.7	11	99.0
2	99.0	12	98.7
3	98.5	13	99.0
4	99.1	14	98.7
5	98.7	15	98.8
6	99.2	16	99.0

7	98.7	17	99.1
8	98.6	18	98.3
9	98.9	19	98.8
10	99.1	20	99.2
Average	98.85	Average	98.86

4. Aligning English –Myanmar Text

Alignment is a central issue in the construction and exploitation of parallel corpora. In this section, we will discuss the various aligning methods and why we have chosen Gale and Church method in aligning English-Myanmar text. We will show our alignment results in the next section.

4.1. A survey of techniques for alignment

Brown et al. [26] performed bitext alignment with a statistical technique with their translations in two parallel corpora. They designed an algorithm based on sentence length in terms of the number of words contained in sentence. No other lexical details of the sentence are utilized and they have achieved accuracy in excess of 99% in a randomly selected set of 1000 sentence pairs.

Gale and Church's approach [1] has been widely adopted for the alignment of European languages and has subsequently been improved with complementary techniques. Their method is based on a simple statistical model of sentence length measured in terms of characters. This method uses the fact that longer sentences tend to be translated into longer sentences in the target language and shorter sentences tend to be translated into shorter sentences.

K-vec algorithm [19] makes use of the word position and frequency feature to find word correspondences using Euclidean distance.

Extended K-vec developed by Nitin [27] uses tests of association as a similarity measure. In his thesis, he investigated the use of measures of association for finding translations in parallel text-score, the Log-likelihood Ratio, Fisher's Exact Test, the Odds Ratio, the Dice Coefficient, and Pointwise Mutual Information.

Fung and Ye [31] describe an approach for finding translations from non-parallel yet comparable texts. They make use of word frequency to find the translation pair. This approach makes use of the observation that words that appear in the context of words that are translations of each other are similar. It makes use of an already existing bilingual lexicon to find the meaning of these context words.

Although the above-mentioned methods are language independent and work for different language pairs, some use the lexical features of the words in the text. *Myanmar texts do not always include spaces between words.* Word segmentation tools, monolingual or bilingual dictionaries are not yet available to date in Myanmar. Hence some of these methods cannot be used as of now. As Gale & Church technique is based only on the character count in the sentences, we have chosen it to align our English and Myanmar parallel corpus.

4.2. Aligning English-Myanmar texts

The purpose of sentence alignment is to identify correspondences between sentences in one language and sentences in another language. Note that alignment may not be one to one, although one to one alignments are more frequent.

Figure 1 plots sentence length correlation in terms of number of characters. We plot this chart using a random sample of 10 files. In the chart horizontal and vertical axes represent English and Myanmar sentence lengths respectively. The regression line indicates that a strong length correlation exists between English and Myanmar. Similar correlation has been observed between other pairs of languages as well [15].

Plain English and Myanmar texts containing one sentence per line are input to the alignment program. The output will be two separate aligned files with line to line correspondence.

The web pages we have downloaded are mostly from Myanmar Chronicle Online Magazine[1] which takes good articles from the other already published magazines and monthly

[1] http://www.mchronicle.com.mm

news of Myanmar. Fortunately, the translation of English into Myanmar is very good and mostly we got one to one sentence correlation in these texts.

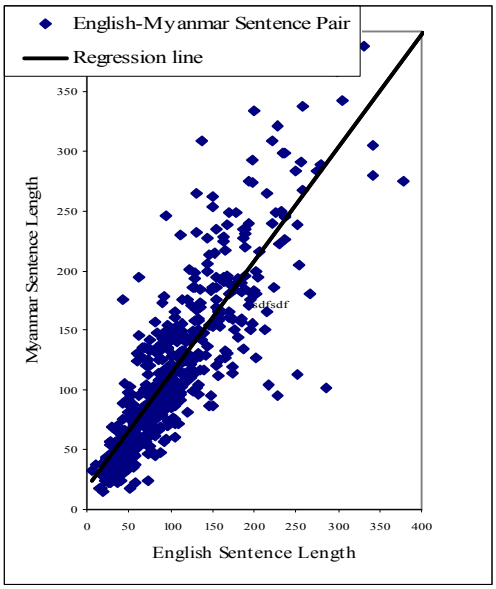

Figure 1. The English-Myanmar sentence length correlation in terms of characters

As an example, we have taken texts from v2n1.pdf (not HTML version) just to show how the Gale and Church approach works for the languages on hand. The example consists of English text includes 8 sentences and Myanmar Text includes only 5 sentences and here we have not taken advantage of paragraph alignment. Aligned output can be seen in Table 4. The alignment of sentences in English to Myanmar is: (E1-M1), (E2, 3-M2), (E4, 5-M3), (E6-M4) and (E7, 8-M5).

We have also extracted examples sentences from English-English-Myanmar student dictionary which is produced by the Ministry of Commerce and transformed to CD version by Inforithm-Maze [13]. In these files, English and Myanmar texts are mixed together and we can not use it directly as a machine readable dictionary. As of now, we have collected 75,000 plus pairs of sentences in all.

5. Experiments and Results

We have aligned the entire corpus of more than 75,000 sentences and we have manually checked and corrected alignment errors. Here we have taken a random set of 600 sentence pairs to show the performance of the alignment system. See table 3. It can be seen that an overall alignment accuracy of about 90% has been obtained. We show the number of one to one alignments in table 2.

Table 2. 1-1 Category in 10 sample files

File Name	No of correctly aligned pairs	1-1 mapping	(%)
Actor	19	14	73.68
Ambassador	27	20	74.07
Lupyo	13	12	92.31
Swim	58	49	84.48
Donation	88	85	96.59
Afta	86	79	91.86
First	7	5	71.43
Story	70	61	87.14
Cover	91	87	95.60
Proverb	53	52	98.11
Total	**512**	**464**	**90.63**

6. Conclusions and Further work

In this paper, we have described our efforts in developing an English-Myanmar parallel aligned corpus. The corpus has been developed by downloading parallel texts and aligning at sentence level using the Gale and Church algorithm. Good performance has been obtained in automatic alignment.

We have also described our work in sentence boundary determination for English. We have presented a generic method for sentence segmentation. We have employed decision tree algorithms for this. We have obtained an average of 98.8% F-measure which is comparable to the earlier work in literature.

Table 3. Percentage of correctness for 10 sample files

File Name	No of English Sentences	No of Myanmar Sentences	No of aligned pairs produced	Correct	Wrong	Correctness percentage
Actor	30	21	23	19	4	82.61
Ambassador	28	38	29	27	2	93.10
Lupyo	16	17	15	13	2	86.67
Swim	81	73	70	58	12	82.86
Donation	112	110	106	88	18	83.02
Afta	95	91	94	86	8	91.49
First	9	7	7	7	0	100.00
Story	77	79	75	70	5	93.33
Cover	103	96	100	91	9	91.00
Proverb	64	64	61	53	8	86.89
Total	**615**	**596**	**580**	**512**	**68**	**88.28**

Table 4. After Aligned Text

Myanmar's hotel industry was started in a large scale only after 1988 when marketoriented economy was introduced.	မြန်မာဟိုတယ်လုပ်ငန်းတွေကိုကြည့်ရင် ၁၉၈၈ နောက်ပိုင်း ဈေးကွက်စီးပွားရေးစနစ် စတင်ကျင့်သုံးပြီး တဲ့နောက်မှ ကြီးကြီးမားမား စတင်ခဲ့ကြတာလို့ ဆိုရမှာဖြစ်ပါတယ်။
Earlier Myanmar had passed many years having hotels under the Government management providing poor service. These hotels are: long-lasting Strand Hotel, Inya Lake Hotel with exceptionally good natural environment on the bank of the lake, an outstanding Thamada Hotel in downrown, Mandalay Hotel beside the Mandalay Moat, veteran Tun Hla Hotel, Thiripyitsaya Hotel at Bagan, a hotel at Ngapali Beach, three hotels of Pyin-oo-lwin: Nann Myaing Hotel, Thiri Myaing Hotel and Gandamar Myaing Hotel, Shwe Inlay Hotel at Nyaung-shwe, Taunggyi Hotel at Taunggyi.	အရင်က တော့ သက်တမ်းရှည် ကမ်းနား ဟိုတယ်၊ မြေအနေအထား အထူးသာယာပြီး အင်းလျားကန်နဲ့ပါ တွဲစပ်နေလို့ သဘာဝကပေးတဲ့ အခြေအနေကောင်းတွေနဲ့ အင်လျားလိတ်ဟိုတယ်၊ ရန်ကုန်မြို့ အလယ်က အထင်ကရ သမ္မတဟိုတယ်၊ မန္တလေးကျွန်းဆေးက မန္တလေးဟိုတယ်၊ မန္တလေးရဲ့ သက် တမ်းရှည် ထွန်းလှဟိုတယ်၊ ပုဂံက သီရိပစ္စယာ ဟိုတယ်၊ ငပလီကမ်းခြေကဟိုတယ်၊ ပြင်ဦးလွင်က နန်းမြိုင်ဟိုတယ်နဲ့ သီရိမြိုင်–ဂန္ဓမာမြိုင်ဟိုတယ်များ၊ ညောင်ရွှေက ရွှေအင်းလေးဟိုတယ်၊ တောင်ကြီးက တောင်ကြီးဟိုတယ်စတဲ့ အစိုးရစီမံခန့်ခွဲမှုအောက်က ဟိုတယ်ပီသတဲ့ ဝန်ဆောင် မှုအပြည့်မဟုတ်တဲ့ နေရာတွေနဲ့ နှစ်အကြာကြီး ဖြတ်သန်းလာခဲ့ကြတာပါ။
In 1989 when market economy was introduced in Myanmar, global tourism had already passed its golden age globally. The word: Smokeless Industry started to became an obsolete word.	၁၉၈၉ မြန်မာနိုင်ငံမှာ ဈေးကွက်စီးပွားရေး စတင်ချိန်ဟာ ကမ္ဘာမှာ ကမ္ဘာလှည့် ခရီး သွားလုပ်ငန်းဟာ အရှိန်ကောင်းချိန်ကိုပင် ကျော်လွန်ခဲ့ပြီး ဟိုတယ်နဲ့. ကမ္ဘာလှည့် ခရီးသွား လုပ်ငန်းဆိုတာ 'မီးခိုးမထွက်တဲ့ စက်ရုံ' ဆိုတဲ့ဝေါဟာရတောင် ဟောင်းစပြုစချိန်ဖြစ်ပါတယ်။
That is why Myanmar business entrepreneurs tried very hard to reach the international status.	ဒါကြောင့် မြန်မာ့စီးပွားရေးလုပ်ငန်းရှင်တွေဟာ အလျင်အမြန်ပဲ အဓိလိုက်နိုင်ဖို့. ကြိုးစားခဲ့ကြပါ တယ်။
However building hotels and training tourism-related staff cannot be realized overnight. Accordingly Myanmar's hotel industry and tourism could run smoothly in 1993-94.	ဟိုတယ်တည်ဆောက်တာ၊ ကမ္ဘာလှည့်ခရီးသွား တွေကို ဝန်ဆောင်မှု ပေးနိုင်ဖို့. ကျွမ်းကျင် တဲ့သူတွေကို လေ့ကျင့်မွေးထုတ်တာတွေဟာ နေ့ချင်းညချင်း ပြီး နိုင်တာတွေ မဟုတ်တာကြောင့် ၁၉၉၃–၉၄ လောက်မှာ မြန်မာဟိုတယ်လုပ်ငန်းနဲ့. ခရီးသွားလုပ်ငန်း ကုမ္ပဏီများ အရှိန်အရ လုပ်ငန်းကောင်းစွာ လည်ပတ်နိုင်ခဲ့ကြပါတယ်။

References

[1] Gale, W.A. and K.W. Church, "A program for aligning sentences in bilingual corpora". In *Proceedings of the 29th Annual Meeting of the Association for Computational Linguistics* (ACL), Berkley, pp. 177–184, 1991.

[2] Dan Melamed, "Empirical Methods for Exploiting Parallel Texts", MIT press, Cambridge, New York City, 2001.

[3] A. Mikheev, "Periods, capitalized words, etc.," Computational Linguistics, Vol. 28, Issue 3, pp: 289 – 318, 2002.

[4] D. D. Palmer, "SATZ - an adaptive sentence segmentation system", *Master's thesis*, 1994.

[5] J. Reynar and A. Ratnaparkhi, "A maximum

entropy approach to identifying sentence boundaries", In *Proceedings of the Fifth Conference on Applied Natural Language Processing*, Washington D.C., pp. 16-19, 1997.

[6] A. Mikheev, "Feature lattices for maximum entropy modeling", In COLING-ACL, pp. 848-854, 1998.

[7] I. H. Witten and E. Frank, "Data Mining: Practical Machine Learning Tools and Techniques with Java Implementations". Morgan Kaufmann publishers, 1999.

[8] A. Colin, "Building decision trees with the id3 algorithm", in Dr. Dobbs Journal, 1996.

[9] P. E. Utgoff, "Incremental Induction of Decision Trees". Kluwer Academic Publishers, 1989.

[10] J. R. Quinlan, "C4.5 Programs for Machine Learning", Morgan Kaufmann publishers, 1993.

[11] T. Mitchell, "Machine Learning", Kluwer Academic Publishers, 1997.

[12] A. Mikheev, "A knowledge-free method for capitalized word disambiguation", In *Proceedings of the 37th Annual Meeting of the ACL*, pp. 159-166, 1999.

[13] English-English-Myanmar dictionary, Ministry of Commerce, CD version.

[14] Piao, Scott Songlin, "Sentence and Word Alignment Between Chinese and English", Ph.D thesis, 2000.

[15] G. Bharadwaja Kumar, Kavi Narayana Murthy, "A Generic approach to Sentence Boundary Disambiguation in English Text", Technical Report - LERC/UOH/2003/2 , University of Hyderabad, India, 2003.

[16] H. Mochizuki, T. Honda, and M. Okumura, "Text segmentation with multiple surface linguistic cues", in COLING-ACL, pp. 881-885, 1998.

[17] Riley and D.Michael, "Some applications of tree-based modeling to speech and language indexing", in *Proceedings of the DARPA Speech and Natural Language Workshop*, Oxford, pp. 339-352, Morgan Kaufmann publishers, 1989.

[18] P. Fung and K. Church. "Kvec: A new approach for aligning parallel texts", In *Proceedings of 15th International Conference on Computational Linguistics* (COLING-94), pp 1096–1102, Tokyo, Japan, 1994.

[19] T. Humphrey and F. q. Zhou, "Period disambiguation using a neural network", in *Proceedings of IJCNN: International Joint Conference on Neural Networks*, 1989.

[20] Mona Talat Diab, "Word Sense Disambiguation Within A Multilingual Framework", *Ph.D Thesis*, 2003.

[21] Hiroyuki Kaji, Yasutsugu Morimoto, "Unsupervised Word Sense Disambiguation using bilingual comparable corpora", In *Proceeding of International Conference on Computational Linguistics* (COLING'02), 2002.

[22] Hang Li, Cong Li, "Word Translation Disambiguation using bilingual bootstrapping", Computational Linguistics, Volume 30, Number 1, March, 2004.

[23] L. Burnard, "The users reference guide for the British National Corpus", Oxford University Computing Services, Oxford, 1995.

[24] Unicode Standard, Chapter 10, http://www.unicode.org/standard/versions/enumeratedversions.html

[25] Peter F. Brown, Jennifer C. Lai, Robert L. Mercer, "Aligning sentences in parallel corpora", Association for Computational Linguistics Morristown, NJ, USA, California, Pp: 169 - 176, 1991.

[26] Nitin Varma, "Identifying Word Translations in Parallel Corpora Using Measures of Association", Master Thesis, University of Minnesota, Duluth, Minnesota, U.S.A, December 2002.

[27] Philip Resnik, "Mining the Web for Bilingual Text", In *Proceedings of the 37th Annual Meeting of the Association for Computational Linguistics* (ACL'99), University of Maryland, College Park, Maryland, June 1999.

[28] Hwee Tou Ng, Bin Wang, and Yee Seng Chan. "Exploiting Parallel Texts for Word Sense Disambiguation: An Empirical Study", In ACL-2003, 455-462, 2003.

[29] Helen Langone, Benjamin R. Haskell, Geroge, A.Miller, "Annotating WordNet", In *Proceedings of the Workshop Frontiers in Corpus Annotation* at HLT-NAACL, 2004.

[30] P. Fung and L. Yee. "An IR approach for translating new words from nonparallel, comparable texts", In *Proceedings of the Thirty-Sixth Annual Meeting of the Association for Computational Linguistics and Seventeenth International Conference on Computational Linguistics*, pp 414–420, 1998.