Automatic Construction of Myanmar-English Bilingual Machine Readable Dictionary using Parallel Corpora

Hla Hla Htay
Department of Computer and Information Sciences
University of Hyderabad
hla_hla_htay@yahoo.co.uk

Kavi Narayana Murthy
Department of Computer and Information Sciences
University of Hyderabad
kmnuh@yahoo.com

Abstract

Parallel texts are scarce resource and have begun to serve as an appealing source of data for corpora based approaches for Natural Language Processing. This paper describes what we can do with the parallel text and how we can construct a dictionary, which is essential resource (in an automatic way) for most of the natural language processing tasks machine translation, information retrieval system, information extraction, word sense disambiguation, and so on.

1. Introduction

Parallel texts are scarce resource for Natural Language Processing. Many researchers are finding such kind of resources and constructing manually if it is necessary. It is really time consuming when manually building this.

Available resources where we can find parallel text are newspapers, bible, and stories and so on.

Corpus-based analysis is especially needed for corpora from specialized fields for which no electronic dictionaries or thesauri exist.

Researchers have devised methods to mine the internet for large amounts of parallel corpora automatically with relatively minimal manual labor at high accuracy levels [16].

2. Myanmar Language

Myanmar language is like Chinese, Japanese, India, and Thailand and so on in Asian Languages. The words are not separated by the space. Therefore, it is considerable more difficult than for Western Languages.

What a difficult for researcher is we don't have any free available monolingual or bilingual machine readable dictionary for research purpose. The lack of such kind of resources prevents the researchers for further study.

Therefore, we have to find the way how to determine the words in the sentences without using the dictionary.

2.1. Segmenting Myanmar Language into words

After observing the Myanmar sentences, what we found out is that it can be tokenized by using the stop list. The stop list is closed word. In figure 1, the underlined words are showing an example how we can detect the words in the sentences by utilizing the stop words. We performed preliminary observation using the newspapers in order to get the stop words.

We need to look at "stop word". It is also called closed word. The removal of closed word items is necessary from alignments since they are a source of noise [6]. The elimination of stop word like this can get the root or stem of the word, for example, "ပန်းများ" ("များ" is used as postfix for plural) will become "ပန်း"(singular) . On the other hand, it may discard some words, for example, the object particle "ကို" and the noun "အကို" (brother).

For example, "သန့်ရှင်းလတ်ဆတ်" is two words "clean, fresh" in English words. It needs further more processing to determine the words like that because there is no stop word between them.

Another good fact in Myanmar language is that it has simple enough morphological structure.

Let see the following examples: the words which are expressed in parentheses can be treated as the stop words. After removing this we could even get the stem or root form of the word.

Single	**man**	**ယောကျား**
Plural	**men**	**ယောကျား(များ)**
Present	**eat**	**စား(သည်)**
Past	**ate**	**စား(ခဲ့သည်)**

3. Preparing for further processing

Once we have parallel texts, we have to do two tasks: Sentence Alignment and word alignment.

3. 1. Sentence Alignment

As we've mentioned above, if we manually align the parallel text, it is very costly in time and labour. Therefore we need to find how we can align pair of text automatically.

The purpose of sentence alignment is to identify correspondences between sentences in one language and sentences in another language. Note that most of the time, a sentence is matched with one or two sentences in the other language.

[18] described a method based on the words that sentences contain. [3] proposed that relies on the simple model of character lengths. They use the heuristics of finding the mark sentence boundaries and paragraph boundaries with a delimiter character, the use of the sentence length, the number of *words* in sentences, and the number of *characters* in sentences.

3. 2. Word Alignment

Word alignment is to identify word correspondence that are translations of each other based on information found on parallel text.

In word aligning, there may have the problem to detect the mapping of compound word of both languages.

In [7], researcher finds the alignment that maximizes the probability of generating the corpus with this translation model.

Figure 1 shows the word correspondence of Myanmar and English sentence.

4. Automatic Bilingual Dictionary Construction

Once we have the alignment words, we can construct the bilingual dictionary. It is promising fact we should follow. There are many reasons difficult to get the up-to-date dictionary.

Dictionary can not contain all the words in real world. It is note that the numbers of the words included in dictionary are not the same with each other. The reason is most of the published dictionary can not be updated by month or year. And newly words are appeared in real world nearly everyday. And many of the technical terms would be missing in a great dictionary.

[3] also claimed that "aligning sentences is just a first step toward constructing a probabilistic dictionary for use in aligning words in machine translation or for constructing bilingual concordance for use in lexicography".

For example, if the company produces a new product. They will give a new name to that product. That can become a new word in later dictionary update. And also new words can appear in today world because of new weather condition, new products, new company, new disease, and so many reasons.

Nowadays, researchers are trying the way how we can construct up-to-date dictionary in an automatic way.

[8] proposed to build a probabilistic lexicon which assigns to each possible transition of an entry a probability measure to indicates how likely the translation is.

4.1 Overview of system

Figure 3 shows the proposed system.

Step 1: Accept pair of Myanmar and English files in HTML or SGML format.

Step 2: These files are passed to Sentence Alignment tool. The file output is Sentence Aligned Corpora.

Step 3: English is well-developed, and there are many freely available resources for that language. English text files are passed to Parser and it will produced Part-of-speech tagged output. Figure 2 describes that the output from the parser. After that, using English Stop word list file, remove the stop words and non-POS tagged word from these files. Depending on the parser, we need to pass the POS tagged word to the morphological analyzer.

Step 4: Segment the words in Myanmar files using Myanmar Stop word list file, and remove the stop words. In this step, human intervention is necessary to add single words to the single word list file. There are still some more problems for detecting the compound word as well as two or more words which is combined and no stop words between them. Using Single word file and split the combined word and again update the single word file.

Step 5: The output from Step 3 and Step 4 are passed to Word Alignment tool which is the statically alignment tool. The result from this step is the aligned words. The high frequency words are taken to insert to Bilingual Machine Readable Dictionary. It is better if the bilingual lexicon created after smoothing the frequency counts would improve its quality.

4.2 The applicability of bilingual dictionary

Bilingual dictionary are an important resource for humans and automated natural language processing systems alike. When translating technical or specialized text, it can be very important to have a bilingual dictionary that tells if and how certain terms are translated. The ability to automatically create such resources from parallel text would make

it possible to quickly extend a fairly generic bilingual lexicon to a more specialized domain [5].

4.2.1 Machine Translation

Systems that perform Machine Translation (MT) are most probable users of bilingual lexicons as these are one of the most important components of such systems. The quality of the resulting translation greatly depends on the quality of the lexicons. A MT system does not have an entry for each word of the text to be translated, and then even a word to word translation is impossible [5].

4.2.2 Information Retrieval

Nowadays, *Google* like such engine embed a cross language *cross-language* IR systems. Information Retrieval (IR) systems search and retrieve relevant documents based on a query. Mono-lingual IR systems find the 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 [5].

4.2.3 Word Sense Disambiguation

Word Sense Disambiguation (WSD) is still open problem. Word sense Disambiguation is the process of distinguishing between different senses of an ambiguous word given in the context.

For example, the word 'function' can be '*mathematical function*' as well as '*social gathering party*'.

As WSD is an intermediate task in many NLP tasks, resolving ambiguity in NLP can provide a lot of promises in solving these NLP tasks such as machine translation, query-based information retrieval and information extraction, and Question and Answering system, etc.

High quality lexical resources are needed to both train and evaluate WSD system [15]. Parallel corpora are one of the promising resources to do that and much interest is achieved to apply in WSD within multilingual frame work [6].

[9] describes that an unsupervised method for word sense disambiguation using a bilingual comparable corpora. [6] proposed the idea which is finding the evidence that support in determining the correct selection of sense for the targeted word using the multilingual corpora. [2] have done word sense disambiguation system using parallel texts.

5. Conclusion and future work

In this paper, we have presented a method to construct a machine readable dictionary. Although this method merely show an idea, we want to shed a light to Myanmar NLP which is needed much work to be done. We conduct a manual experiment to investigate the feasibility of using parallel corpora for identifying such a system.

By construction a lexicon like that may be affected by factors such as the quality of the translation of parallel corpora, the accuracy of sentence boundary marker, and the accuracy of word segmentation, incompleteness and inconsistency in translation, misspellings of words in both language, and these factors even degrade the quality of the resulted lexicon.

In the future work, we intend to modify this method by mapping existing popular resource, lexical resource like WordNet Ontology (which has been used extensively for its wide coverage, and large network of semantic relations), in order to get the senses, synonym like lexical relation, and hypernym like conceptual relation and so on for disambiguation word sense [14].

References

[1] Thomas Emerson, "Segmenting Chinese in Unicode",*16th International Unicode conference, Amsterdam, The Netherlands, March 2000.*

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

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

[4] Yang Muyun. et. al, "Auto Word Alignment Based Chinese-English EBMT".

[5] Nitin Varma, "Identifying Word Translations in Parallel Corpora Using Measures of Association", December 2002

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

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

[8] Multilingual domain modeling in Twenty-One: Automatic creation of a bi-directional translation lexicon from a parallel corpus"

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

[10] Hang Li, Cong Li, "Word Translation Disambiguation using bilingual bootstrapping".

[11] Katherine Deibel, "Current Approaches for Unrestricted Word Sense Disambiguation".

[12] Hang Li, Cong Li , "Word Translation Disambiguation using Bilingual Bootstrapping"

[13] Nicola Cancedda, Herve Dejean, Eric Gaussier, Jean-Michel Renders, and Alexei Vinokourov,"Report on CLEF-2003 experiments: two ways of extracting multilingual resources from corpora".

[14] Geroge Miller, "WordNet: an on-line lexical database", *International Journal of Lexicography, 3(4): 235-312.*

[15] Helen Langone, Benjamin R. Haskell, Geroge A.Miller, "Annotating WordNet".

[16] 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.*

[17] Aito Chen, Kazuaki Kishida, Hailing Jiang and Qun Liang, "Automatic construction of a Japanese-English lexicon and its application in cross-language information retrieval".

[18] Brown P. Lai J. and Mercer R, "Aligning sentences in parallel corpora", *Proceeding of ACL 1991.*

Showing Example *Stop Words* in the sentence:

Myanmar	(Meaning in English)
၏	of
ကို	object particle
သည်	subject particle
ဖြစ်သည်	is

မူးယစ်ဆေးဝါး$_1$ သည် နိုင်ငံအားလုံး$_2$ ၏ လူထုလူတန်းစား$_3$ အလွှာအသီးသီး$_4$ ကို ထိုးဖောက်ဝင်ရောက်$_5$ နေသည့် ပြဿနာ$_6$ ဆိုးကြီး$_7$ တစ်ခု ဖြစ်သည်။

The devastating$_1$ problem$_2$ of narcotic$_3$ drugs$_4$ had penetrated$_5$ all$_6$ walks$_7$ of$_8$ life$_9$ throughout the international$_{10}$ community$_{11}$.

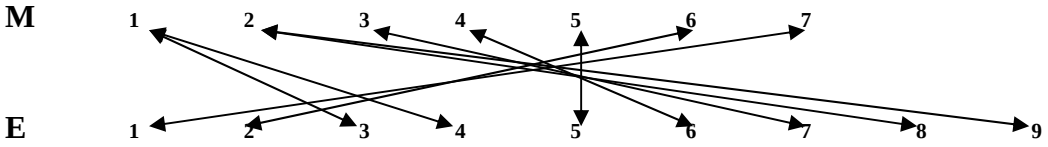

Figure 1. Showing Annotation of words

The devastating.g problem.n [of] narcotic.n drugs.n had.v penetrated.v all walks.n of life.n throughout the international.a community.n

Figure 2. The POS information obtained from a parser

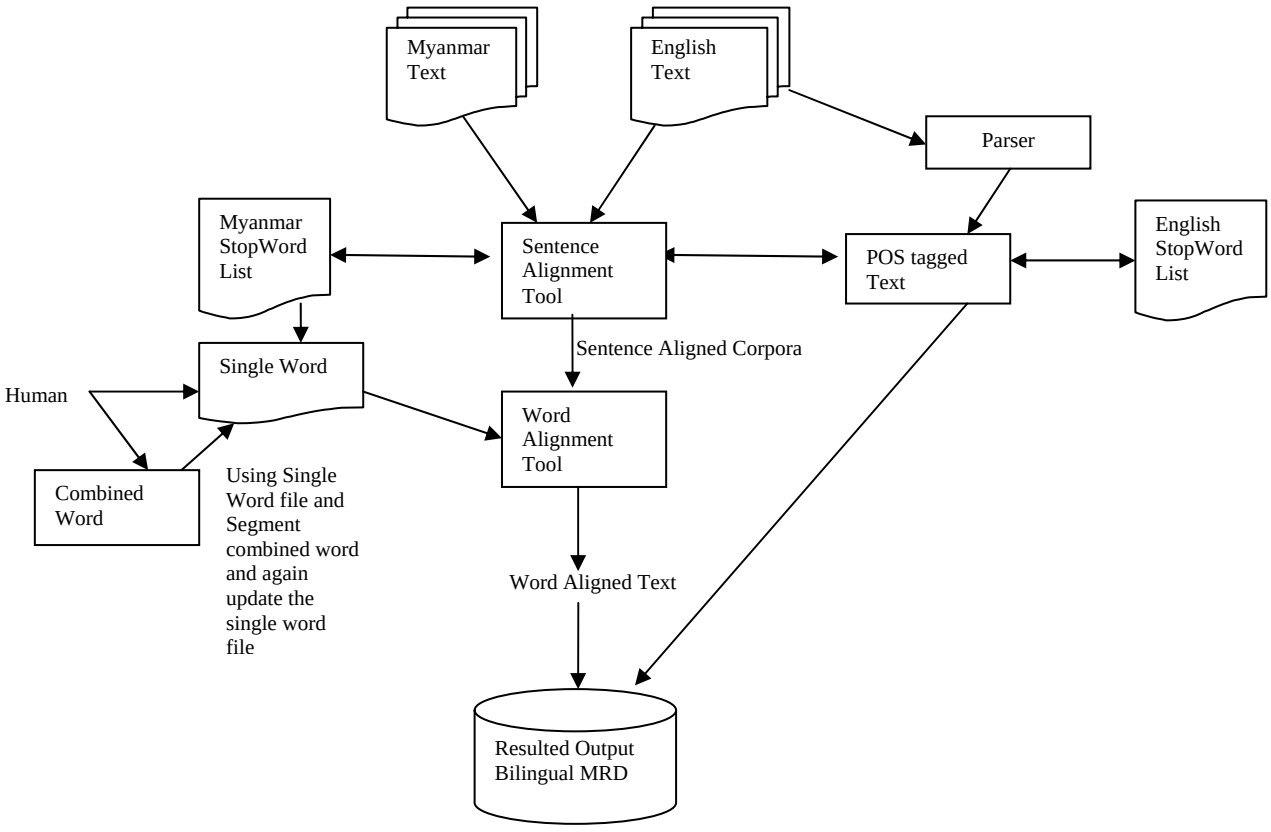

Figure 3. Proposed System Flow