Automatic Bilingual Lexicon Extraction for a Minority Target Language[*]

Eileen Pamela Tiu[a], and Rachel Edita O. Roxas[a]

[a]College of Computer Studies, De La Salle University-Manila
2401 Taft Avenue, Malate, Manila, Philippines
roxasr@dlsu.edu.ph

Abstract. An automated approach of extracting bilingual lexicon from comparable, non-parallel corpora was developed for a target language with limited linguistic resources. We combined approaches from previous researches which only concentrated on context extraction, clustering techniques, or usage of part of speech tags for defining the different senses of a word. The domain-specific corpora for the source language contain 381,553 English words, while the target language with minimal language resources contain 92,610 Tagalog word, with 4,817 and 3,421 distinct root words, respectively. Despite the use of limited amount of corpora (400k vs Sadat's (2003) 39M word corpora) and seed lexicon (9,026 entries vs Rapp's (1999) 16,380 entries), the evaluation yielded promising results. The 50 high and 50 low frequency words yielded 50.29% and 31.37% recall values, and 56.12% and 21.98% precision values, respectively, which are within the range of values from previous studies, 39 - 84.45% (Koehn et al., 2002 and Zhou et al., 2001). Ranking showed an improvement to overall F-measure from 7.32% to 10.65%.

Keywords: Automatic lexicon extraction.

1. Introduction

Automatic bilingual lexicon extraction is the automated process of acquiring bilingual word pairs from corpora in order to construct a lexicon. The product of this process, a bilingual lexicon (or a dictionary), is commonly used for Machine Translation, Natural Language Processing, and other linguistic usages. Since acquiring lexicon from parallel corpora already yields 99% accuracy (Rapp, 1999), this study focused into processing non-parallel corpora, specifically on comparable corpora.

In addition, a lexicon is more helpful if it has the feature of grouping similar senses together. This feature is called word sense discrimination. In this study, the syntax and senses of the word are the contributing factors to discriminate words that have several meanings.

Throughout the years, linguists and translators compile different types of lexicons manually from experts and automatically from corpora, which are collections of texts and other forms of writings. *Parallel corpora* refer to a source text and its translation into one or more target languages (Ahrenberg, et al., 1999). Parallel corpora are used for lexicon extraction due to the following characteristics (Fung, 1998): (1) Words have one sense per corpus; (2) Words have single translation per corpus; (3) No missing translations in the target document; (4) Frequencies of bilingual word occurrences are comparable; and (5) Positions of bilingual word occurrences are comparable.

However, acquiring parallel corpora is labor intensive and time consuming. It is also unlikely that one can find parallel corpora in any given domain in electronic form especially for minority languages. This problem can be solved by using non-parallel corpora. Non-parallel corpora

[*]

come in two forms: comparable and non-comparable corpora. *Comparable* corpora are collections of documents that have common domains or topics but different authors and published dates, while *non-comparable* corpora have totally different domains, authors, and published dates. This factor is crucial for languages with minimal language resources, and the utilization of existing language resource (such as non-comparable corpora) should be maximized to automatically generate other language resources (such as a bilingual lexicon).

Several methods have been developed to utilize these types of corpora to automatically construct a lexicon such as the co-occurrence model (Rapp, 1999), the Convec algorithm (Fung, 1998), the exploration of other clues (Koehn and Knight, 2002), the dependency grammar approach (Zhou, 2001), and the word filtering approach (Sadat et al., 2003). Others utilize these types of resources for word sense disambiguation (WSD) (Kikui, 1999 and Kaji, H et al., 2005). Rapp (1999) yielded 65% accuracy when the first word in the ranked list was considered and 72% when other senses were considered. Convec (Fung, 1998) yielded a 30-76% precision when the top 20 ranks of the translations were considered. Zhou et al. (2001) achieved a 70.8% accuracy for high frequency verbs and 81.48% for low frequency words. Koehn et al. (2002) achieved 39% noun translation accuracy. The average precision of Sadat et al. (2003) was 41.7%.

The main idea behind these techniques is to collect words that co-occur with the source word in the corpora and to establish the correlation between the patterns of word co-occurrences in the corpora of another language. Words that occur in a certain context window in one language have translations that are likely to occur in a similar context window in another language (Koehn et al., 2002).

Nevertheless, relying on context alone is not sufficient to handle ambiguity. For example, the word “find” in the sentence “How did you find the Philippines?” may be interpreted automatically as either “discovery” or “observe”. The process of assigning correct senses to an ambiguous word, called *disambiguation,* is addressed by the following word sense disambiguation approaches (Kikui, 1999 and Kaji, H et al., 2005).

On the other hand, discrimination is a subtask of disambiguation that clusters (or groups) similar senses of a word into a number of classes (Schutze, 1998). Part-of-speech (POS) provides syntactic information and serves as a useful clue to sense disambiguation (Stevenson et al., 2001), and imply its usefulness to sense discrimination. However, current studies on discrimination are applied to monolingual corpora and syntactic information is not included in classifying word senses.

This study used the basic concepts introduced by Rapp (1999) to build a lexicon from two comparable, non-parallel corpora of the same domain. Thus, it assumes the existence of a bilingual lexicon as its initial seed lexicon. The application domain of this study is on tagged English and Tagalog corpora. Also, a small scale set of tagged English and Tagalog corpora is used to simulate the presented approach.

The research did not support phrasal translations, and bi-directional translations such as Tagalog to English.

The evaluation based on Rapp (1999) and Zhou et al. (2001) used 50 high and 50 low word frequency occurrences from the corpora and manually validated by human experts. The criteria of their judgment were based on the acceptability of the outputs.

Section 2 discusses the approach taken in this study, and the various modules of the developed system. Section 3 presents the evaluation results. And finally, Section 4 presents some recommendations and conclusions.

2. The Modules of the Lexicon Extraction System

The modules of the lexicon extraction algorithm are presented in Figure 1. Before processing, the initial linguistics resources were first collected and the corpora were pre-processed where POS tagging and stemming were performed. Both source and target corpora underwent context extraction where co-occurrence analysis of each of the words is performed. Alignment is performed on the contexts of the words to be translated with the target corpora through the aid of an initial bilingual lexicon. Initial clustering is performed wherein similar senses of the target translations are grouped together. Feature vector computation and word ranking computes for the ranks of candidate translations.

The initial resources involved in this process include the corpora and the lexicon. The types of corpora used in this research involve comparable types. The corpora contain 381,553 English and 92,610 Tagalog terms, with corresponding 4,817 and 3,421 distinct root words, respectively.

An initial bilingual lexicon, taken from a previous project IsaWika! (Roxas, 1997) served as a seed lexicon and contains 9,026 English elements, and unique meanings and parts of speech. In addition, both English and Tagalog entries were stemmed with the stemming algorithms.

The corpora underwent preprocessing to remove punctuation marks, convert the words to lowercase, stemming, removal of function words, and part of speech assignment. Function words were removed from the source corpus by removing words that were found in the function word lists obtained (Mitton, 1987). Function words in the target language were removed manually by computing for the highest word frequencies and removing the undesired terms. The Porter's Algorithm (Anu, 2003) was used to stem the English documents, while TagSA, a Tagalog stemmer (Bonus, 2003), was used to stem the Tagalog documents. An online English Memory Based Tagger (Zavrel et al., 1999) was used to tag the source documents.

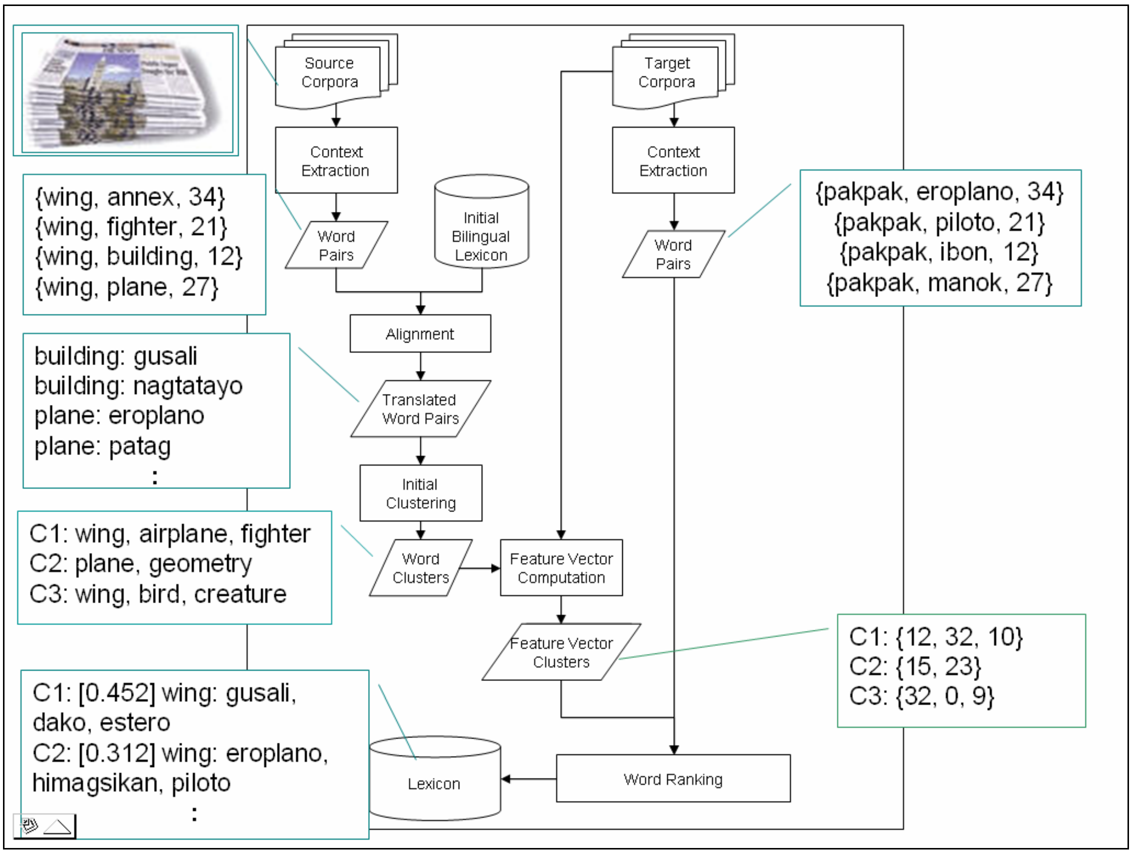

Figure 1: Automatic Bilingual Lexicon Extraction for a Minority Target Language.

2.1.Context Extraction

In the source corpora, context extraction was performed on every word found in the corpora to generate *source word pairs* that represent a set of words that are gathered in a window size of 25 in conjunction with the work of (Kaji et al., 2005). Word pairs are represented in the form:

{SourceWord, Source POS Tags, ContextWord, Context POS Tags, Co-occurrence Frequency} which indicates the number of times a word co-occurs with its neighboring words. For example, as shown in Table 1, we can say that "aaron" co-occurs with "abihu" 20 times in the entire corpora.

Source words are then collected and selected from the set of source word pairs. Similarly, in the target corpora, context extraction was performed on every word found in the corpora to generate *target word pairs*.

Table 1: Representation of Word Pairs.

(English) Source Word Pairs				
Source Word	**Source POS Tags**	**Context Word**	**Context POS Tags**	**Frequency**
aaron	nnp	abiasaph	nnp	1
aaron	nnp	abihu	nnp	20
aaron	nnp	abiram	nnp	3
aaron	nnp	abishua	nnp	5
aaron	nnp	abl	jj	7
aaron	nnp	about	in	10
aaron	nnp	account	nn	2
aaron	nnp	across	in	2
aaron	nnp	act	vbg	3
aaron	nnp	addit	nn	1

2.2.Alignment Process

The alignment process functions as a bridge between the source and target languages based on the word pairs from the context extraction module and the initial bilingual lexicon. This process finds all translations to the contexts of the *source words*. This was done at an early stage to reduce the dimensionality of the data and thus increase the speed of the processes to be performed later. An issue in this stage is the existence of multiple translations per context. It was decided, however, to consider all translations since the lexicon was built in such a way that the translations were all referring to a similar sense. Thus, it should enhance the algorithm since it provided other possible translations for a single term projecting one sense. Context words derived from the *source word pairs* are translated to the target language with the help of the *initial bilingual lexicon*. Words not found in the initial bilingual lexicon are discarded from the word pair list. Table 2 shows a sample output of the alignment algorithm.

2.3.Sense Grouping

The usual way of clustering starts by collecting random points from the collection that will serve as its initial points in clustering. This proved detrimental to the results thus, later algorithms, like CBC (Pantel, 2003), were modeled in such a way to solve the deficiency by first using agglomerative algorithms to compute for initial points that will serve as their starting points in clustering. The sense grouping algorithm is based on this idea, and is composed of three stages. The first stage locates clusters from the source corpora and assigns different ClusterIDs to different clusters. The second stage computes for the *feature vectors* (or the average) of each cluster collected in the target corpora. The third stage involves ranking the elements computed within these *feature vectors* by comparing the *feature vector* to the set of word pairs extracted from the target corpora. Note that the word pairs were extracted from the target corpora through the *context extraction* module.

Table 2: Bridging the Two Languages.

Word to be Translated		Contexts of the Word to be Translated		Tagalog Translations of the Contexts Words from the Lexicon
Source Words	Source POS Tags	Context Words	Contexts POS Tags	
said	nnp	agreement	nn	sundu
		altar	nn	dambana
		angel	nn	anghel
		ark	nn	arka
		captain	nn	pit
		come	vbp	daan
		delight	nn	lugod
		ey	nns	butas
		face	nns	mukha
		faith	nn	anampalataya
		fear	nn	takot

The *translated word pairs* contain the source word pairs that were translated using the *initial bilingual lexicon* so as to reduce the dimensionality of the data. The elements within the *translated word pairs* are clustered into their most similar senses with the clustering algorithm. And the output of the process is the *word clusters* wherein words with similar senses are grouped into a cluster.

The clustering technique used to group similar terms together is called *single link clustering*. The aligned words are first grouped into their most similar senses. This aims to merge terms whose similarity is beyond a certain threshold into the cluster table. Cluster tables are used to store word clusters. The table will begin to fill up as elements are being assigned and merged inside the table. The threshold was set experimentally to 0.8, 0.85 and 0.9. The clusters generated by the 0.90 threshold were found to be favorable since it generated clusters with fewer number of elements than the other thresholds. It will only stop merging the clusters once all the elements are not similar with any of the other clusters.

The contexts of the source word having translations in the lexicon were clustered with their most similar senses with neighboring words using the clustering algorithm. After which, words having a common sense were assigned a single ClusterID. A partial output consisting of one cluster, ClusterID 759, is shown in Table 3. As observed, the word "sai", with Part of Speech *vbz*, consisting of cosine similarity 0.93129767, occurs more than once with the same Part of Speech but different Cosine Similarity of 0.87990112. This happens as the cosine similarities of the words are computed based on their common neighboring contextual words. Each source word has many contextual neighbors, thus generating different cosine similarities. Words with higher cosine similarity indicate the "nearness" of the context to the source word. Thus, words with higher similarity score, 0.9, will be clustered together. The contexts were clustered from the source and the *feature vectors* were computed based on the target corpora in order to determine the existence and the frequency of the translated words in the target corpora.

From the *ClusterTable* containing the generated *Word Clusters*, the *TargetCorpora*, and the translations of the elements within those clusters generated, a *feature vector* for each cluster of the source word are computed by getting the average co-occurrence frequency of the results generated. The *centroid* or the average of the generated translation is computed such that the summation of the frequency counts of each translation is divided by its number of occurrences. For example, the word "*sabi*" co-occurs 291 times with other words as shown in Table 4, as each word is recounted each time it appears together with the target words, thus possibly generating a higher co-occurrence value than the word_count. This value is then divided by the actual word count of "*sabi*", which is counted individually within the cluster window of the Target Corpora, which is 12 times, yielding 24.25 for the first word. The same procedure is repeated for all the other elements. Note that overstemming has occurred for the first and last words in the context of the words "*sai*" and "*salvat*" which are supposed to be words "*said*" and "*salvation*", respectively.

Table 3: Example of an Output Cluster Table

Source Word			Contexts of the Source Word		
Word	**Part of Speech**	**ClusterID**	**Word**	**Part of Speech**	**Cosine Similarity**
said	vbd	759	sai	vbz	0.93129767
			right	jj	0.92557105
			turn	vbg	0.92209296
			work	nns	0.91110829
			turn	vbg	0.89938562
			love	nn	0.89279824
			turn	vbg	0.89015483
			heaven	nn	0.88574855
			heaven	nns	0.88109349
			salvat	nn	0.88105634
			sai	vbz	0.87990112

Table 4: Feature Vector of the Translated Terms.

Word to be Translated			Context of the Words to be Translated			
Word	**POS Tags**	**ClusterID**	**Words**	**POS Tags**	**Tagalog Translations**	**Centroid**
said	vbd	759	sai	vbz	sabi	24.25
			right	jj	may	11.12
			turn	vbg	gawa	10.51
			work	nns	gawa	10.51
			turn	vbg	tayo	10.30
			love	nn	Ibig	9.39
			turn	vbg	bigay	7.64
			heaven	nn	langit	6.56
			heaven	nns	langit	6.56
			salvat	nn	ligtas	5.81

After computing for all the feature vectors of the clusters, these were then compared to the set of word pairs extracted. The word pairs that correspond to a similarity higher than a given threshold would serve as the possible senses of the *source word.* The inputs required were the cluster feature vectors computed and the word pairs gathered. The feature vectors were compared against the vectors of the word pairs in the target language and their cosine similarity computed. The computed values were arranged in descending order to indicate that those with higher values have higher similarity. The word alignments of the *Source Words* with probable translations are generated. It loops through the elements belonging to a feature vector and compares it against the words in the target corpora and computes its cosine similarity. As illustrated in Table 5, the *Target Translations* are the words that matched with the *feature vector* computed and its cosine similarity was computed based on the *feature vector* and the vector of the target words. Note that among the candidate translations, "*sabi*" and "*tinig*" which ranked 0.47 and 0.46, respectively, are closest semantically to the source word "*said*" where “*sabi*” is the literal translation, while “*tinig*” (with literal meaning “*voice*”) is semantically related.

Table 5. Similarity of the Translations to the Source Words

Words to be Translated			Translations	
Word	**POS Tags**	**ClusterID**	**Word**	**Cosine Similarity**
said	vbd	759	takot	0.47
			anghel	0.47
			sabi	**0.47**
			tinig	**0.46**
			ulan	0.46
			yari	0.45
			may	0.44
			taka	0.44
			labas	0.44
			propeta	0.43

3. Evaluation Results

Various evaluation approaches have been used to assess the performance of the automatic lexicon extraction algorithm, and no existing standard has been developed. Thus, a test plan was designed that is similar to the approaches presented by Rapp (1999) and Zhou (2001). The approach of Rapp performed an evaluation using 100 randomly selected words whereas the approach of Zhou performed an evaluation using a combination of frequently and rarely occurring words. A hybrid approach was used because this study aims to determine the effectiveness of the algorithm on words having different number of occurrences in the corpora. The number of candidate words was limited to 100 because of performance considerations, that is, the processing time needed for each word. Furthermore, it permits opportunities for fine-tuning the algorithm and performing re-executions.

From the source corpora, 50 words with the highest occurrence frequencies and 50 words with the lowest occurrence frequencies were identified. The effectiveness of the lexicon extraction algorithm was tested on these 100 words, by comparing the expected translations which were taken from an initial bilingual lexicon with the actual results of the algorithm. These words then underwent the *context extraction* process that outputs *feature vectors* for the contexts of the 100 terms. The context words, together with their part of speech tags, within these feature vectors are translated to the target language with the help of an initial bilingual lexicon to yield their target language counterparts and thus transforming the feature vector from the source language to a corresponding feature vector in the target language. The target feature vector is then compared to all the vectors found in the target corpora by computing their similarities. The generated sets of translations are then clustered first according to their syntax, then with their most similar senses. The translation corresponding to the top similar senses per cluster were gathered.

The F-measure was used in the research to determine the effectiveness of the approach to both high and low frequencies. In addition, a portion of the test cases were presented to point out some features offered by the extraction algorithm.

As a summary, the yielded translations captured most of the expected translations and placed them in high ranking positions. Furthermore, the Part of Speech tags and the discriminating algorithm complemented each other in grouping similar senses together. It was also able to select the correct translations for some low frequency words that have correct senses associated with the cluster scattered in the target corpus.

The F-measure was used to evaluate the overall performance of the system. The initial F-measure calculated was 7.32%. The generated value was influenced by several factors. The first reason was that it was strictly based on the output of the initial lexicon provided in the beginning of the study, thus, other correct translations for a particular source word were not considered. Secondly, alternate translations were also not considered in evaluation, in order to reduce the inconsistencies in testing. Thirdly, the sizes of some of the generated clusters were large. Thus, another form of testing was done on the results. The F-measure was modified to include the ranks generated by the algorithm. And the ranked F'-measure was calculated to be 10.65%, improving the results generated by the former evaluation. For the recall test, the system was able to yield 50.29% for high frequency words and 31.37% for low frequency words of the expected translations in all clusters. For the ranked precision test, the system was able to yield 56.12% for high frequency words and 21.98% for low frequency words of the expected translations within clusters. The precision and recall values were found to be within the range of values reported from previous researches, which range from 39 - 84.45% (Koehn et al., 2002 and Zhou et al., 2001). In contrast to the size of the corpora and lexicon used, the system used significantly smaller sized corpora (100K for source and 300K for target), versus the resources used in the related literature, wherein the smallest corpora resource used was 39M (Sadat et al., 2003) and the lexicon resource was 16,380 (Rapp, 1999).

It should also be noted that several factors were involved in generating the output. First, the algorithm performs well on high frequency words. Second, the sparseness of the data contributes to the drawback of the performance.

4. Conclusion

This research has presented an algorithm that automatically selects translations from comparable corpora involving a minority target language. Previous researches have only concentrated on acquiring translation from one language to another incorporating the use of parts of speech for defining the different usages of a word in a sentence.

An improvement introduced in this study involves the use of clustering algorithm to group together similar senses of a word. One of the contributions of this research is the combination of the word context extraction (Rapp, 1999) with the clustering technique (Pantel, 2003) and other clues like the part of speech tags in the source corpora. Other researches only concentrated on either context extraction or on clustering techniques only. This research first extracts the contexts of the source word, clusters them into their most similar sense, and then ranks the output through the assistance of the target corpora.

The system was tested on a set of 50 high frequency words and 50 low frequency words. The F-measure was used as the evaluation measure on the output translations of the initial lexicon. The initial F-measure calculated was 7.32%. And the ranked F'-Measure was calculated to be 10.65%, improving the results generated by the former evaluation. It was shown that the algorithm performs on a satisfactory level. This could be attributed to several factors such as the quality of corpora, preprocessing errors, and the inclusion of some function words in the target corpora.

The algorithm is dependent on the quality of translation generated by the initial lexicon and it would generate better results if a high quality lexicon is provided for evaluation. An alternative approach is to develop an algorithm that does not completely rely on the use of an initial lexicon, as introduced by (Koehn et al., 2002). In this study, function words in the target corpora were manually removed and it would be better if a basis could be found for these set of function words. Schachter (1972) has listed a set of function words that can be used as the basis for the removal of function words. Data sparseness affected the quality of translations, thus, some smoothing techniques could also be implemented in order to complement the lexicon extraction algorithm.

Disambiguation techniques, the introduction of a Tagalog part of speech tagger and the removal of function words on the target corpora based on the available list would be good sources of improvement to the present system. In addition, an algorithm could also be developed to handle bi-directional translations, i.e. from Tagalog to English.

References

Ahrenberg, L., M. Merkel, D. Ridings, A. Sågvall Hein and J. Tiedemann. 1999. Automatic Processing of Parallel Corpora: A Swedish Perspective. Linkoping Electronic Articles in Computer and Information Science, ISSN 1401-9841, 4(2).

Anu, J. 2003. Porter's Stemming Algorithm (Visual Basic Version). [online]. Available: http://www.tartarus.org/~martin/PorterStemmer. (February 27, 2004).

Bonus, E. 2004. A Stemming Algorithm for Tagalog Words. *Proceedings of Philippine Computing Society Congress 2004*.

Fung, P. 1998. A Statistical View on Bilingual Lexicon Extraction: From Parallel Corpora to Non-Parallel Corpora. In Farwell, D., L. Gerber and E. Hovy, editors, *Third Conference of the Association for Machine Translation in the Americas*, 1-16.

Kaji, H and Y. Morimoto. 2005. Unsupervised Word Sense Disambiguation Using Bilingual Comparable Corpora. *IEICE Transactions on Information and Systems 2005*, E88-D(2): 289-301.

Koehn, P and K. Knight. 2002. Learning a Translation Lexicon from Monolingual Corpora. *In the Proceedings of ACL Workshop on Unsupervised Lexical Acquisition*.

Kikui, G. 1999. Resolving Translation Ambiguity Using Non-parallel Bilingual Corpora. *In Proceedings of ACL99 Workshop on Unsupervised Learning in Natural Language Processing*.

Mitton. 1987. Function word list from: Spelling Checkers, Spelling Correctors and the Misspellings of Poor Spellers. Function Word List. [online]. Available: http://www.cse.unsw.edu.au/~min/ILLDATA/Function_word.htm. (February 28, 2004).

Rapp, R. 1999. Automatic Identification of Word Translations from Unrelated English and German Corpora. *In Proc. of ACL-99*, 519-526.

Roxas, R. 1997. Machine Translation from English to Filipino: A Prototype. *International Symposium of Multi-lingual Information Technology (MLIT '97)*, Singapore.

Sadat, F., M. Yoshikawa and S. Uemura. 2003. Learning Bilingual Translations from Comparable Corpora to Cross-Language Information Retrieval: Hybrid Statistics-based and Linguistics-based Approach. *In the Proceedings of the Sixth International Workshop on Information Retrieval with Asian Languages*, 57-64.

Schutze, H. 1998. Automatic Word Sense Discrimination. *Computational Linguistics*, 24(1), 97-123.

Stevenson, M. and Y. Wilks. 2001. The Interaction of Knowledge Sources in Word Sense Disambiguation. *Computational Linguistics*, 27 (1).

Zavrel, J & Daelemans, W. 1999. Recent Advances in Memory-Based Part-of-Speech Tagging. *In VI Simposio Internacional de Comunicacion Social*, 590-597.

Zhou, M., Y. Ding, and C. Huang. 2001. Improving translation selection with a new translation model trained by independent monolingual corpora. *Journal of Computational linguistics and Chinese language processing*, pp 1-26, 6(1).