Automatic Target Word Disambiguation Using Syntactic Relationships

Ebony Domingo[1] and Rachel Edita Roxas[2]

[1] Computer Science Department, College of Science and Information Technology, Ateneo de Zamboanga, La Purisima St., Zamboanga, Philippines 7000
[2] Software Technology Department, Dela Salle University-Manila, 2401 Taft Avenue, Malate, Manila, Philippines 1000
{domingoeboc@yahoo.com, roxasr@dlsu.edu.ph}

Abstract. Multiple target translations are due to several meanings of source words, and various target word equivalents depending on the context of the source word. Thus, an automated approach is presented for resolving target-word selection, based on "word-to-sense" and "sense-to-word" source-translation relationships, using syntactic relationships (subject-verb, verb-object, adjective-noun). Translation selection proceeds from sense disambiguation of source words based on knowledge from a bilingual dictionary with sense profiles and word similarity measures from WordNet, and selection of a target word using statistics from a target corpus. Test results using English to Tagalog translations showed an overall 64% accuracy for selecting word translation with a standardized precision of at least 80% for generating expected translations using 200 sentences with ambiguous words (an average of 4 senses) in three categories: nouns, verbs, and adjectives, using 145,746 word pairs in syntactic relationships, extracted from target corpora (317,113 words).

Keywords: word sense disambiguation, machine translation.

1 Introduction

Target word disambiguation is a task in machine translation where a decision has to be made on which of a set of alternative target-language words is the most appropriate translation of a source-language word [1], a process familiarly known as *translation selection*. For instance, the correct translation of the word 'wash' in Tagalog could be *hilamos, hugas, laba,* etc. depending on the object noun of the source verb 'wash'.

Several methods have been developed for target-word disambiguation on different types of corpora using different nature of word translations. Techniques exploit monolingual corpora on either the target language (target language based) or the source language to resolve lexical ambiguities. Target language based approaches include the use of statistics on lexical relations [3], estimation of translation probability using a language model of the target language [5], [1]. Other methods exploit information from the source language for disambiguation such as distributional clustering [6].

A more recent and novel approach in translation selection is the hybrid method [2] based on the "word-to-sense and sense-to-word" relationship between source word and its translations, the method selects translation through two levels: sense disambiguation of a source word and selection of a target word. Other techniques worth mentioning in this field is the use of dependency triples on an unrelated monolingual corpus to select among translations of a given verb [4]. A more recent approach exploits content-aligned bilingual corpora for phrasal translations based on monolingual similarity and translation confidence of aligned phrases of two languages [7].

This research addresses resolving word translation ambiguity based on the idea of "word-to-sense and sense-to-word" relationship between source word and its translation using a bilingual dictionary and syntactic relations (subject-verb, verb-objects and adjective noun) on un-tagged, monolingual corpora in the target language with word sense disambiguation on source words.

2 Translation Selection

Majority of the methods for translation selection usually select a target word directly from a source word. Such direct mapping is referred to as 'word-to-word' relationship. Based on this, previous approaches could easily obtain statistical rules from corpora.

Although difficulty of knowledge acquisition is relieved, such methods are bound to select incorrect translations, even if the set of target words are reduced, since ambiguity of both source and target words are not taken into consideration. For instance the English word break can be translated to its various Tagalog senses as follows: *sira, durog, bali, basag, bakli, sakit, pinsala, suway, labad, kontra, laya, takas, bunyag, siwalat* and *hayag.*

The 'word-to-sense and sense-to-word' relationship mean that a word in a source language has multiple senses and each sense can be mapped into multiple target words [2]. Using such relationship, senses of the source words are disambiguated before selecting a translation. Since each sense covers a set of target words, information can be utilized needing less elaborate knowledge. For the word *break*, word-to-sense and sense-to-word relationships are as follows: (1) destroy: *sira, durog, basag, bali, bakli*; (2) hate: *pinasala, sakit*; (3) violate: *labag, suway*; (4) escape: *laya, takas*; (5) reveal: *bunyag, siwalat, hayag*.

Senses of source word are resolved first before selection of a target word. Knowledge for resolving word ambiguities can be extracted from various machine-readable dictionaries. As for this study, knowledge for word sense disambiguation was extracted from the English-Tagalog Dictionary [8], which contains sense definitions of an English word with a list of Tagalog translations grouped for each sense. The English word *break* has the following entry: "v. (1) to damage: *sumira, masira* (accidental), *sirain* (deliberate). He broke the machine. The machine broke down (stopped): *Nasira* (*Huminto*) *ang makina* … (6) to snap (off) as stick. Branch or stalk: *bumakli*, *bakliin*. He broke the stick into two: *Binakli niya ang patpat*. He broke (off) the stalk. *Binakli niya ang tangkay*. (7) to snap, break as string or wire: *malagot*, *lumagot*, *lagutin*. The wire broke. *Nalagot ang alambre*. (8) to break; at against; disobey*: sumuway, suwayin. Lumabag, labagin*. He broke the law. *Sinuway* (*Nilabag) niya ang* batas."

3 System Workflow

A general overview of the system workflow is presented in the architectural design in Fig. 1. The components are: (1) the preprocessing of language resources for sense profiling (source to target lexicon, target lexicon and target corpora), (2) sense disambiguation and target word selection, and (3) translation preference. WordNet is also a resource used for word similarity measures since it organizes nouns and verbs into hierarchies of *is-a* relations.

The target corpora with 317,113 words are online Tagalog articles and the New Testament. 145,746 words in syntactic relationship (SR) were extracted from target corpora using a partial parser [9] and a bilingual lexicon [10]. The sense profiles consist of entries of source words in different senses along with translations in each sense, and content words extracted in definition and example sentences [8].

The process of translation selection proceeds from classifying senses of word in the input sentences through computation of word similarity based on WordNet hierarchy [11]. Sense probability (*sp*) represents how likely target words with the same sense co-occur with translations of other words in syntactic relationship with, in an input sentence, and is computed based on target word co-occurrence [2]. Then word probability which represents the probability of selecting a target word among all other target words in the same sense division is computed [2]. Finally, selection of a target word among all other translations of a source word is done by computing the translation preference for each translation. Thereby merging results from sense classifier, sense probability and word probability. Values from sense classifier and sense probability are added as a score of sense disambiguation. Score for word selection is computed by using a normalizing factor for word probability [2] to prevent discounting the score of a word for which its sense has many corresponding target words. Then selection is made on the target word with the highest computed translation preference factor.

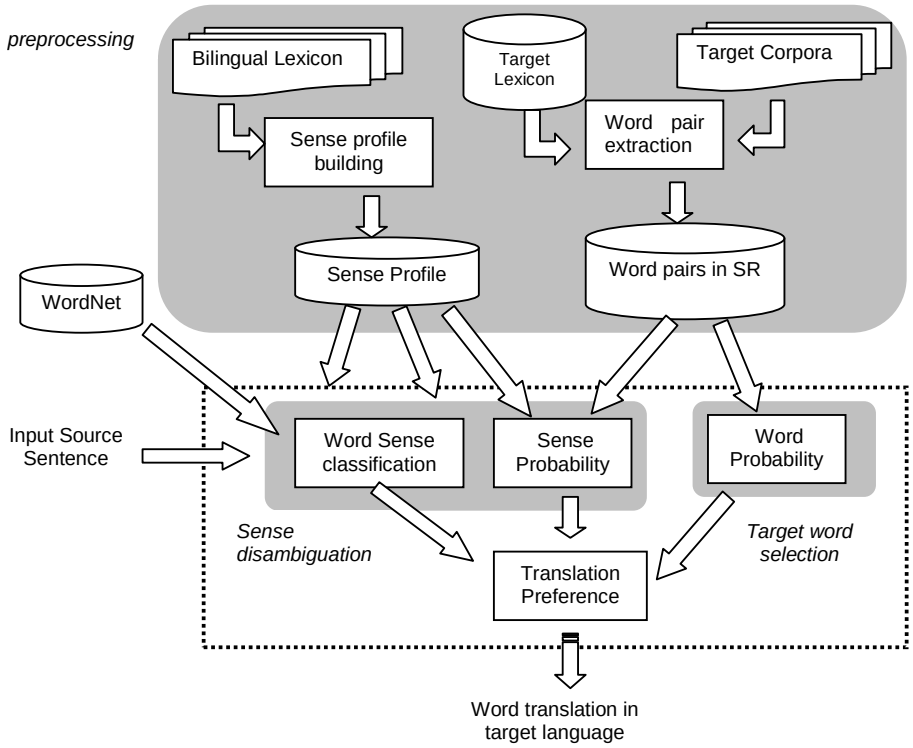

Fig. 1. The architectural design.

A set of 200 bilingual sentences extracted from various bilingual dictionaries and books is used for testing, with 244 extracted word pairs in syntactic relation using a memory-based shallow parser. From these word pairs, there were 217 nouns, 92 adjectives, and 148 verbs with average senses of 3, 5 and 6, respectively, for a total of 457 words. Translations of content words participating in syntactic relationship - subj-verb, verb-object, adjective-noun and subj-adjective – were obtained and evaluated on altering combinations of clues and measures.

For sense disambiguation, sense preference and sense probability were used. The sense classifier used words in sense definitions (*DEF*) and words in example sentences (*EX*) as clues for sense disambiguation. The accuracy of each module was evaluated by testing whether any target word of the sense that scores the highest is identified as a translation of its source word in a target sentence.

4 Results and Discussions

Accuracies of the different combination of clues for the sense classification are computed. Using clues both found in the definition and example sentences (DEF-EX) produced better results than using clues found in sense definitions only (DEF). Accuracy for verbs and adjectives increased at using both clues from definition and example sentences (DEF-EX). However, accuracy for nouns is higher when only clues from example sentences (EX) are used. Overall accuracy is 61.27%.

Translation preference results are presented in Fig. 2. Alterations on combinations of preference, sense probability (*sp*) and word probability (*wp*), were compared against three baselines: random selection, first translation of the first sense (1^{st} sense) and most frequent translation (*mft*). The sense classification (*sc*) shows the accuracy of selecting the first translation of the sense that scores the highest. Combinations considered are sense probability and word probability (*sp* x *wp*), sense classification and word probability (*sc* x *wp*), and all measures ((*sc*+ *sp*) x *wp*). Based from the results of the tests, the overall accuracy is 64% with a standardized precision of above 80%.

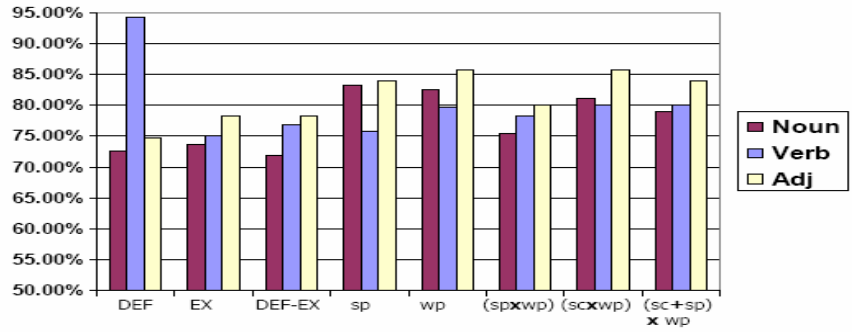

Fig. 2. Standardized Precision for Accuracy of Each Measure of Translation Selection

5 Conclusions

This research presented a method that automatically resolves target-word ambiguity which resolves senses of source words through word similarities and target word selection through word co-occurrence. Evaluation on 200 sentences with highly ambiguous words with altering combination of the measures for resolving target-word ambiguity has shown a 64% accuracy of the expected translations with a standardized precision of above 80%. The method is highly dependent on clues found in the sense profile for disambiguation of source words.

Despite a satisfactory result, the algorithm could not properly disambiguate some source words because of inadequate clues in sense definition as well as example sentences for a certain sense of a source word. Some smoothing techniques can further improve results since 0-values produced by sense probability and word probability affected the quality of translation. The system can further be improved with the integration of morphological analysis.

Acknowledgments. This project is partially funded by the Philippine Council for Advanced Science & Technology for Research & Development, Department of Science and Technology, Philippine Government.

References

1. Prescher, D., Reizler, S., and Rooth, M.: Using Probabilistic Class-based Lexicon for Lexical Ambiguity Resolution. Proceeding of the 18th International Conference on Computational Linguistics (2000)
2. Lee, H. A., Yoon, J., and Kim, G. C.: Translation Selection by Combining Multiple Measures for Sense Disambiguation and Word Selection. Inter Jour of Computer Processing of Oriental Languages, 16(3) (2003)
3. Dagan, I. and Itai, A.: Word Sense Disambiguation Using a Second Monolingual Corpus. ACL (1994)
4. Zhou, M., Ding, Y., and Huang, C.: Improving Translation Selection with a New Translation Model Trained by Independent Monolingual Corpora. Journal of Computational Linguistics and Chinese Language Processing, Vol. 16 No. 1 (2001) 1-26
5. Koehn P. and Knight, K.: Knowledge Sources for Word-level Translation Models. Proceedings of the Empirical Methods in Natural Language Processing Conference (2001)
6. Kikui, G.: Resolving Translation Ambiguity using Non-parallel Bilingual Corpora (1999)
7. Aramaki, E., Kurohashi, S., Kashioka, H., and Tanaka, H.: Word Selection for EBMT based on Monolingual Similarity and Translation Confidence. HLT-NAACL Workshop 2003 (2003)
8. English, L. J.: English to Tagalog Dictionary. Congregation of the Most Holy Redeemer (2003)
9. Abney. S.: Tagging and partial parsing (1996)
10. Tiu, P.E.: Lexicon Extraction from Comparable Corpora. MSCS Thesis. College of Computer Studies, De La Salle University (2003)
11. Pedersen, T., Patwadhan, S. and Michlizzi, J.: Wordnet::Similarity – Measuring the Relatedness of Concepts. In Proceedings of 5th Annual Meeting of the North American Chapter of the ACL. (2004).