

# T2CMT: Tagalog-to-Cebuano Machine Translation

Jacqueline G. Fat  
Department of Mathematics & Computer Science  
College of Arts & Sciences  
University of San Carlos  
Talamban, Cebu City Philippines 6000  
(6332) 344-3801 local 328  
jgfat@usc.edu.ph

## ABSTRACT

T2CMT is a uni-directional machine translator for languages Tagalog and Cebuano, specifically it translates from Tagalog to Cebuano. The morphological analysis is based on TagSA (Tagalog Stemming Algorithm) and affix correspondence-based POS (part-of-speech) tagger. A new method is used in the POS-tagging process but does not handle ambiguity resolution and is only limited to a one-to-one mapping of words and parts-of-speech. The syntax analyzer accepts data passed by the POS tagger according to the formal grammar defined by the system. Transfer is implemented through affix and root transfers. The rules used in morphological synthesis are reverse of the rules used in morphological analysis. A bilingual dictionary from Tagalog to Cebuano was developed and is used by the different components of the system.

T2CMT has been evaluated, with the Book of Genesis as input, using GTM (General Text Matcher), which is based on Precision and Recall. Result of the evaluation gives a score of good performance 0.8027 or 80.27% precision and 0.7992 or 79.92% recall.

## General Terms

Algorithms, Design, Experimentation, Languages, Theory.

## Keywords

Machine translation, Parser, Morphology, POS Tagger.

## 1. INTRODUCTION

Machine Translation (MT) is a technology that automatically translates text from one human language into another. The source language (SL) and/or the target language (TL) medium might be text or speech, but most MT systems work with text.

The main distinction of MT systems is in terms of overall strategy: whether translation from SL to TL takes place in a single stage (direct translation), in two stages (via an 'interlingua'), or via the 'transfer' approach, where translation proceeds in three stages [2].

Machine translation, using the transfer approach, generally follows different phases: morphology, syntax, and semantics [1]. Morphology refers to the study of the structure of words or how words are formed. Syntax deals with how words can be combined

together to make larger phrases, such as, sentences. Semantics deals with real-world knowledge or the meaning of the sentence.

Research in the field of Natural Language Processing and Machine Translation is not fully developed in the Philippines where different languages and dialects are used. Within the 7,200 islands of the Philippine archipelago, there are about one hundred and one (101) languages that are spoken. This is according to the nationwide 1995 census conducted by the National Statistics Office of the Philippine Government. The languages that are spoken by at least one percent of the total household population include Tagalog, Cebuano, Ilocano, Hiligaynon, Bikol, Waray, Pampango or Kapangpangan, Boholano, Pangasinan or Panggalatok, Maranao, Maguindanao, and Tausug. Aside from these major languages, there are other Philippine dialects, which are variants of these major languages [15].

Roxas, et al. stated that Computational linguistics in the Philippines is currently focused on Tagalog using the LFG framework. Their study showed that not much has been done on the other Philippine languages with respect to the computational aspects of these languages towards a multi-lingual machine translation system. They recommended that further study be conducted on the design and eventual implementation of such an MT system involving Philippine languages [17].

## 2. EXISTING WORKS

There are notable works related to Machine Translation employing Philippine languages. Some works are MT systems [5, 7, 11, 14], others can be applied to MT systems [4, 6].

ISAWIKA! is a transfer-based English-to-Tagalog MT system that uses (Augmented Transition Network) ATN as the grammar formalism. It translates simple English sentences into equivalent Filipino sentences at the syntactic level [16]. Another transfer-based English-to-Filipino MT system was designed and implemented using the lexical functional grammar (LFG) as its formalism. It involves morphological and syntactical analyses, transfer and generation stages. The whole translation process involves only one sentence at a time [5]. Another work is a multilingual machine translation system designed for Tagalog, Cebuano and English. It exploits structural similarities of the Philippine languages Tagalog and Cebuano, and handles the free word order languages. It translates at the syntactic level only. It does not employ morphological analysis in the system [10].

CARLA (Computer Assisted Related Language Adaptation) is a system that allows the user to write linguistic rules to do automated morphological parsing and then transfers the text morpheme by morpheme to produce a rough draft of the input text in a related language. It works one sentence at a time. CARLA gives a very literal translation<sup>1</sup> from SL to TL [22]. As a result, CARLA works best between closely related languages with similar word order, grammatical and morphological structure, and cultural and idiomatic expressions.

Research projects on morphological analysis and stemming present new approaches in its area. TAGMA (Tagalog Morphological Analyzer) is based on Optimality Theory (OT) and two-level morphology that handles both concatenative and non-concatenative phenomena for Tagalog verbs. Optimality Theory is a phonological approach that is proven effective in handling non-concatenative phenomena and has been applied for generation process but never been used in morphological analysis [9]. TagSA, a Tagalog Stemming Algorithm, was developed for all forms of Tagalog words. It can be used specifically for morphological analysis to derive root words. In addition, it can also be applied to information retrieval (IR) to conflate different word forms to a common canonical form. It uses the principle of iterative affix removal and is context sensitive [4].

Commercial translation softwares, which include Philippine languages, are ETTE 2000 2.4, Filipino Language Software, InterTran Web Site Translation Server, Wordtran, and the Universal Translator 2000. These translation softwares perform word-for-word translations.

### 3. T2CMT SYSTEM OVERVIEW

#### 3.1 Architectural Design

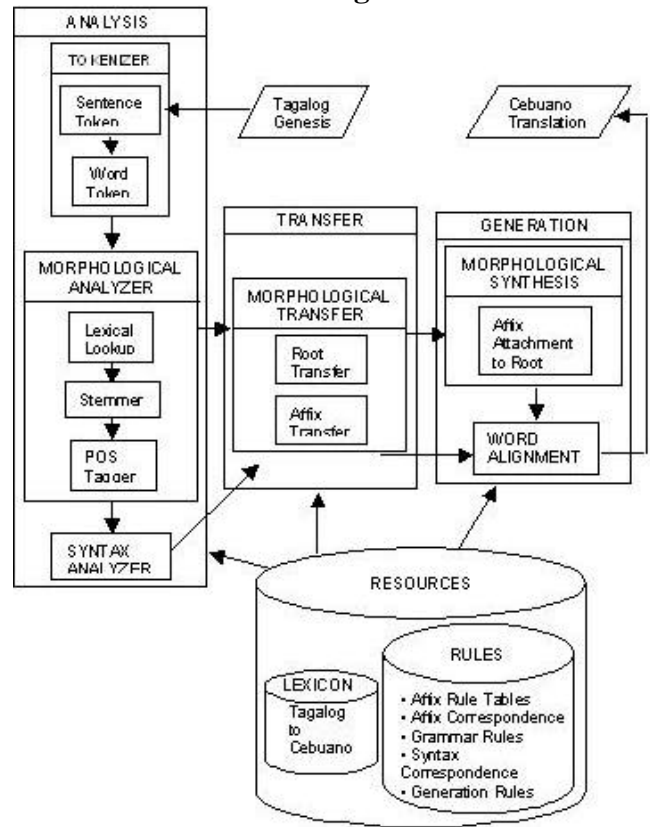


Figure 1. Architectural Design

The architectural design of T2CMT is shown in Figure 1. It has three stages: Analysis, Transfer and Generation. Each stage uses the resources of bilingual dictionary and the set of rules. The analysis stage takes, as input, sentences from the Book of Genesis. It then performs processes of tokenization or segmentation, lexical lookup and morphological analysis. The output of this stage will be passed to the next stage, which is transfer. Affix and root transfers will be performed in this stage. The result of the transfer stage will be fed to the Generation stage for morphological synthesis and word alignment. The final outcome of the system is the Cebuano equivalent of the input sentences.

#### 3.2 The Lexicon

Lexicons are the largest components of an MT system and the most expensive components to construct. The size and quality of the lexicon limits scope and coverage of a system, and the quality of translation that can be expected [14].

Tagalog-to-Cebuano dictionaries are currently not available whether in electronic or printed form. There are dictionaries, though, that contain the above languages (e.g. English-Tagalog-Cebuano-Bicolano dictionary). The Tagalog-to-Cebuano Machine Translator (T2CMT) needs a Tagalog-to-Cebuano dictionary containing root words only for it will handle both Tagalog

<sup>1</sup> A *literal* translation is one that follows very closely the word order and structure of the source text. In contrast, a *free or dynamic* translation changes structure and wording in significant ways to produce a text that sounds natural in the target language.

morphological analysis and Cebuano morphological synthesis. Since there is no such dictionary available as of present, a new one is built. These are the steps followed in the development of the said dictionary:

1. implementing TAGSA (Tagalog Stemming Algorithm) [4] in C
2. input the Book of Genesis Tagalog Version in the Tagalog stemmer and list all the root words generated by the stemmer in a text file. Generation of Tagalog root words using TAGSA is roughly produced due to its limitations.
3. manual look-up for the parts-of-speech and Cebuano equivalents of the generated Tagalog root words using available dictionaries ([6], [18], [7], [3], [21], [20] and the Tagalog and Cebuano [13] versions of the Book of Genesis)
4. using a C program to sort the list of dictionary entries in alphabetical order

### 3.3 Affix Correspondence Table

The affix correspondence table is used in the transfer of affixes and in part-of-speech (POS) tagging.

The Affix Correspondence<sup>2</sup> Table is used in the Transfer module, specifically in the Affix Transfer sub-module. Each entry in the affix correspondence table is written as:

[tagalog\_affix]/[cebuano\_affix]/[attach\_to\_POS]/[result\_POS]

where

[tagalog\_affix] uses the dash (-) symbol: after a prefix (ika-), before and after an infix (-in-) and before a suffix (-han). PART\_RED signifies partial reduplication of the root word.

[cebuano\_affix] uses the digits 0 for prefix (0mo), 1 for infix (1in), and 2 for suffix (2on).

[attach\_to\_POS] is the part-of-speech (POS) of the root word to which the affix(es) will be attached. The POS "ANY" means that the affix(es) can be attached to any root word.

[result\_POS] is the part-of-speech of the word after the affixes are attached. The POS "ROOT" means that the resulting word will take the part-of-speech of the root word.

The parts-of-speech of the root words are listed as follows: ADJ (Adjective), ADV (Adverb), CONJ (Conjunction), DAT (Dative case), GEN (Genitive case), INTJ (Interjection), LIG (Ligature), NN (Noun), NOM (Nominative case), NUM (Number), PART (Particle), PN (Proper Noun), PREP (Preposition), PRON (Pronoun), and VRB (Verb).

An entry in the affix correspondence table looks like:

|  |                   |
|--|-------------------|
|  | mag-/0mag/NN/NN   |
|  | mag-/0mag/ANY/VRB |

<sup>2</sup> Some o

**Figure 4.8 Sample entry in Affix Correspondence Table**

The first entry "mag-/0mag/NN/NN" means that the Tagalog prefix *mag* (the hyphen after *mag* signifies that it is a prefix) has a corresponding Cebuano prefix *mag* (0 before *mag* signifies that it is a prefix). If the category of the root word is NN (the first NN in the entry), then the category of the resulting word (root word + affix) is also NN (the second NN in the entry).

The second entry "mag-/0mag/ANY/VRB" means that if the category of the root word is anything other than NN, then the category of resulting word is VRB.

### 3.4 Evaluation

The T2CMT system was initially tested with 16 sentences. The syntax of the test sentences is the same as the syntax of the test sentences of PinoyMMT [10]. Sentences following this syntax were selected from the Book of Genesis but sentences in the Book of Genesis follow complex sentence patterns. Hence, the words in PinoyMMT's test sentences were modified to suit the domain of this research, the Book of Genesis.

Two types of evaluation, human and automatic, were done on the initial test results of T2CMT. On the average, 78.7% of the human evaluators judged the translation output as "acceptable".

An initial automatic evaluation has also been done on the whole Book of Genesis generating the following scores:

precision = 43.09%  
 recall = 37.38%  
 f-measure = 40.01%

The following improvements has been done on the system:

1. pre-processing the input (Book of Genesis) such that only verbal sentences are retained;
2. adding functions in morphological analysis to handle irregular words and reduplication with assimilation  
 Ex. mangangahoy → manga- + kahoy  
 pumagitan → pa- + -um- + gitna  
 dalhan → dala + -han  
 takpan → takip + -an  
 lagyan → lagay + -an  
 sidlan → silid + -an
3. adding entries in the Affix Correspondence Table from the Book of Genesis.

A final evaluation has been on the whole Book of Genesis. The average precision, recall and f-measure scores are 80.27%, 79.92%, and 80.09% respectively. These scores fall beyond the range of good performance [19], which means that the system is able to

perform well in translating the Book of Genesis from Tagalog to Cebuano.

#### 4. CONCLUSION

The morphological rules for both Tagalog and Cebuano were studied, as well as its grammar rules.

Tagalog and Cebuano morphological rules hold both similarities and differences, in addition to its corresponding affixes. While analyzing these similarities and differences, Tagalog and Cebuano affix correspondences were found useful in determining the part-of-speech of word forms.

Differences in Tagalog and Cebuano grammar rules are found to be trivial, hence this research focuses on its similarities. Giganto (2003) also found that the rules of Tagalog and Cebuano are similar in structure.

The machine translation system, which was designed, tested, and evaluated, showed good performance with a score of 0.8027 or 80.27% precision and 0.7992 or 79.92% recall.

#### 5. RECOMMENDATIONS

The following are for future works: integration of ambiguity resolution in dictionary lookup and affix correspondence lookup, handling of multi-words and word derivatives, employing thought-for-thought translation to capture multi-words translation, adding semantic analysis, and extend the scope of the domain and the dictionary.

#### 6. REFERENCES

- [1] Arnold, D. (1997). What is LFG [online]. Available: <http://www.essex.ac.uk/linguistics/LFG/WhatIsLFG.html>. (May 3, 2001).
- [2] Arnold, D., et al. (1995). Machine Translation: An Introductory Guide [online]. Available: <http://www.essex.ac.uk/linguistics/clmt/MTbook/HTML>. (May 3, 2001).
- [3] Bautista, J., Enriquez, M., & Jamolangue, F. (2001). Pocket Dictionary English-Tagalog-Visayan-Ilonggo-Cebuano Vocabulary. Manila: Marren Publishing House Inc.
- [4] Bonus, D.E. (2003). A Stemming Algorithm for Tagalog Words. Manila: De La Salle University. MS Thesis.
- [5] Borra, A. (1999). A Transfer-based Analysis Engine for an English to Filipino Machine Translation Software. Manila: University of the Philippines Los Banos. MS Thesis.
- [6] Cabonce, R., S.J. (1983). An English-Cebuano Visayan Dictionary. Manila: National Bookstore.
- [7] Carlsen, J. E. (2002). English-Tagalog Lexikon First Edition [online]. Available: <http://swefil.com/pdf/engtagv1.pdf>. (May 20, 2004).
- [8] Cubar, N. (1974). Complex Sentences in Tagalog, Cebuano, and Hiligaynon. Manila: University of the Philippines.
- [9] Fortes, F.C. (2002). A Constraint-based Morphological Analyzer for Concatenative and Non-concatenative Morphology of Tagalog Verbs. Manila: De La Salle University. MS Thesis.
- [10] Giganto, R. (2003). Exploiting Structural Similarities of Philippine Languages For A Multilingual Machine Translation System. Manila: De La Salle University. MS Thesis.
- [11] Grand Rapids, MI: Christian Classics Ethereal Library (2002). The Holy Bible: Cebuano Translation [online]. Available: <http://www.ccel.org/ccel/bible/c1.toc.html> (Feb. 3, 2004).
- [12] Green, R., Turian, J., Melamed, I., Shen, L., Argyle, A. (2004). General Text Matcher (GTM) [online]. Available: <http://nlp.cs.nyu.edu/GTM/>. (October 13, 2004).
- [13] PBS: Philippine Bible Society (1981). Maayong Balita Alang Kanimong. Manila: United Bible Societies.
- [14] Reinhard, S. (2003). Machine Translation: Role of the Lexicons in MT [online]. Available: <http://www.cogsci.uni-osnabrueck.de/~reinhard/MT/MT04.pdf>. (May 5, 2004).
- [15] Roxas, R. and Borra, A. (2002). Policies for Machine Translation Research & Development in the Philippines. Survey on Research and Development of Machine Translation in Asian Countries, Thailand, May 13-14, 2002.
- [16] Roxas, R., Devilleres, E., Giganto, R. (2000). Language Formalisms for Multi-lingual Machine Translation of Philippine Dialects. De La Salle University, Manila, 2000.
- [17] Roxas, R., Devilleres, E., Giganto, R. (2001). Computational Representation of Philippine Dialects: Towards a Multi-Lingual MT. 38th Annual Conference of the Association for Computational Linguistics, Hongkong, October 1-8, 2000.
- [18] Sagalongos, F. (1968). Diksyunaryong Filipino-Ingles. Manila: National Bookstore.
- [19] SDSU: San Diego State University (2000). Machine Understanding and Data Extraction [online]. Available: <http://www-rohan.sdsu.edu/~ling354/understanding.html>. (October 27, 2004).
- [20] Tagalog Dictionary (2004). Available: <http://www.tagalog-dictionary.com/cgi-bin/>.
- [21] Tungol, M. (1987). Modern English-Pilipino-Cebuano Dictionary. Manila: Merriam & Webster Bookstore, Inc.
- [22] White, S. and Stone, R. (2004). Introduction to CARLA STUDIO for Philippine Languages. Document version 0.9.