International Journal for Research in Applied Science & Engineering Technology (IJRASET)
ISSN: 2321-9653; IC Value: 45.98; SJ Impact Factor: 6.887
Volume 6 Issue IV, April 2018- Available at www.ijraset.com

English-Ilokano Language Translator using Example-Based and Rule-Based Machine Translation

Rowell L. Casil[1]
[1]*Department of Graduate Programs Computer Science Technological Institute of the Philippines, Quezon City*

Abstract: ***Natural Language Processing is broadly defined as the automatic manipulation of natural language like text, by software. It provides both theory and implementations for a range of applications such as Machine Translation (MT). This study introduces a new way of implementing approaches for machine translation that utilized the strength of Example-based and Rule-based Machine Translation in translating English to Ilokano sentence. Since there are a lot of single words in Ilokano language that can be expressed in whole sentence in its equivalent English language, Example-based approach was used to translate those sentences. For the rest of the sentences, Rule-based approach was the idea for translating that involves analyzation, transfer and generation phases. The Stanford Log-linear Part-Of-Speech Tagger was used to analysed the input English sentence to get the part of speech (POS) for each word. Pattern grammar rules in English and Ilokano have been applied to check the grammar of the sentences. For the mixed translation, the combination of the two approaches was used to translate the sentence. The performance of the translator was being evaluated by comparing the reference output from the MT output. The accuracy of the translation results was 84% which means that the translations are acceptable and understandable.***
Keywords*: Natural Language Processing, Machine Translation, Example-based, Rule-based, Ilokano Language*

I. INTRODUCTION

Natural Language Processing (NLP) is an area of research and application that explores how computers can be used to understand and manipulate natural language text or speech to do useful things [1]. NLP research primarily focuses on gathering knowledge on how people use and understand their natural language so that appropriate tools and techniques can be developed that enable computer systems to mimic this level of understanding and allow them to manipulate languages to perform useful tasks. According to [2], there are two issues involving NLP: first is the language barrier problem and second is the extinction of languages in the Philippines.

Research in the field of NLP is not fully explored in the Philippines where different languages and dialects are used [3]. With the 7,641 islands of the Philippine archipelago, there are about one hundred and one languages that are spoken. Known Philippine languages which had already been processed in NLP were Tagalog, Cebuano, Ilonggo, Kapampangan and Ilokano.

Ilokano is the third most-spoken native language of the Philippines but there were only 3.9 % of the research in NLP in the country that involves Ilokano language. Other languages were English language 45.9% and Tagalog language 33.3 % [2]. NLP provides both theory and implementations for a range of applications such as Machine Translation.

Machine Translation (MT) can be defined as translation from one natural language (source language (SL)) to another language (target language (TL)) using computerized systems with or without human assistance. The SL and/or the TL medium might be text or speech, but most MT systems work with text [4].

Machine Translation is very challenging because of numerous issues due to language ambiguity like grammar, structure and even fluency of use. On the basic level, a machine translator simply converts sentences by substituting word to word from source language to target language. But only the word substitution would not be able to deliver desired results as it doesn't care about semantic and syntactic constraints of the target language. To improve the quality of translation, researchers developed and advanced linguistic paradigms of MT such as Rule-based Machine Translation [5], Example-based Machine Translation [6] and Statistical Machine Translation [7] among others.

These approaches have their own strengths and weaknesses. Statistical Machine Translation is more efficient in the use of human and data resources but specific errors are hard to predict and fix. On the other hand, Rule-based Machine Translation is an indirect approach that is expected to produce high quality translations. However, this approach has its weaknesses. Rule-based Machine Translation is effective only for a given language pair. Furthermore, some linguistic information still needs to be set manually.

Example-based Machine Translation addresses this problem. Instead of using rules, this uses a sample corpus as basis for translating a given text. Since rules are no longer needed, development and maintenance is no longer dependent on humans. Example-based Machine Translation is also good in translating idiomatic expressions. However, Example-based Machine Translation has its drawbacks. Aside from it dependency on the domain of the sample corpus, if the text to be translated was never seen in the corpus, the system will no longer be able to translate it.

Since there are a lot of single words in Ilokano language that can be expressed in whole sentence in its equivalent English language, Example-based Machine Translation is suitable for it. However, Example-based Machine Translation doesn't check for grammar which Rule-based Machine Translation do. Hence, this study will attempt to use Example-based and Rule-based Machine Translation in translating English to Ilokano language.

II. OBJECTIVES OF THE STUDY

This study attempted to fuse approaches for machine translation utilizing the strength of Example-based and Rule-based.

A. The specific objectives of the study are as follows:

1) Translate English to Ilokano sentence using Example-based Machine Translation;
2) Translate English to Ilokano sentence using Rule-based Machine Translation involving the following phases;
3) Analyse the English sentence structure by performing processes segmentation, tagging, and syntax parsing;
4) Process the lexical transfer from the source text to the target text based from the output of the analysis phase; and
5) Rearrange the translated words using Ilokano grammar rules.
6) Translate English to Ilokano sentence using the combination of Example-based and Rule-based.

III. THEORETICAL FRAMEWORK

A. Review of Related Literature

There are a lot of notable studies involving Machine Translation with different approaches used such as Syntax Based Machine Translation System from English to Hindi Language[8], Grammar-Based and Example-Based Techniques in Machine Translation from English to Arabic[9], Thai to Isarn dialect Machine Translation using Rule-based and Example-based[10], ISAWIKA![11], Text Translation: Template Extraction for Bidirectional English – Filipino Example-Based Machine Translation[12], LFG-Based Machine Translation Engine for English and Filipino[13], T2CMT:Tagalog-to-Cebuano Machine Translation[14], and Bi-directional Ilocano-English Language Translator using Customized Moses Statistical Machine Translation System[15] among others. Existing MT systems in international and local setting use different architecture and translation paradigms.

B. Concept of the Study

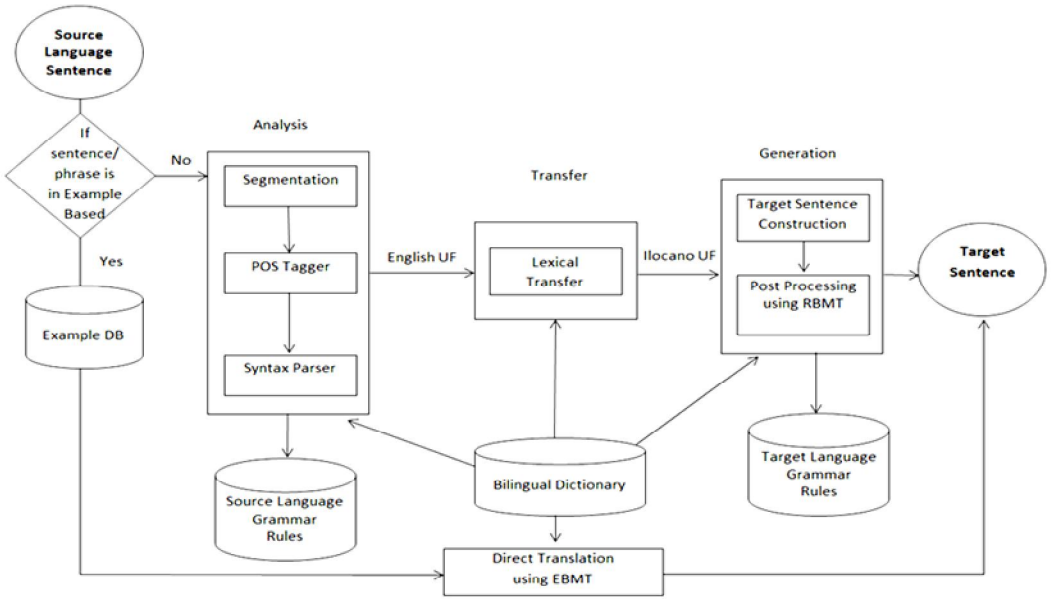

Fig. 1 Architectural Design

The architectural design of this research is shown in Fig. 1. The translator will examine the sentence if it is in the example database or not. If the sentence is in example database, direct translation to the target sentence will take place. Otherwise, it has to undergo three stages namely analysis, transfer and generation. Each stage uses the databases of bilingual dictionary and the set of rules. The analysis stage takes, as input (word, phrase, and sentence). The first step in translation is to split the sentence into words or phrases. If a part of the input is in the example database then we keep that part as it is and the remaining part is segmented into words. When the segmentation is done, the chunks or segments are actually translated and tagged independently. If the segment is example based, it is directly converted to Ilokano. Further, the remaining words are translated using English-Ilokano Dictionary. It then performs processes of POS-tagging and parsing. The output of this stage, the English underlying form (UF), will be passed to the next stage, which is transfer. Lexical transfer will be performed in this stage. The result of the transfer stage is the Ilokano UF, which will be fed up to the Generation stage. On finishing the tagging of all the segments, it rearranges them and constructs a valid sentence of it by applying proper grammar. The final outcome of the system is the Ilokano equivalent of the input sentences.

The structural difference between English and Ilokano Language is shown as a Parse tree Structure. Example in English sentence (Rowell cooked a delicious food.) translated into Ilokano sentence (Nagluto ni Rowell ti naimas a makan.). The sentence pattern of Ilokano is inverted from English sentence. Traditionally, the predicate (main verb and other clauses) comes first before the subject. Conjunctive Adverbs can also come first before the subject. In contrast to English, which is Subject, Verb, Object, or mostly Active in voice, the use of such is very rarely used in Ilokano. Fig. 2 and Fig 3 show the Parse tree structure of the English and Ilokano sentence respectively.

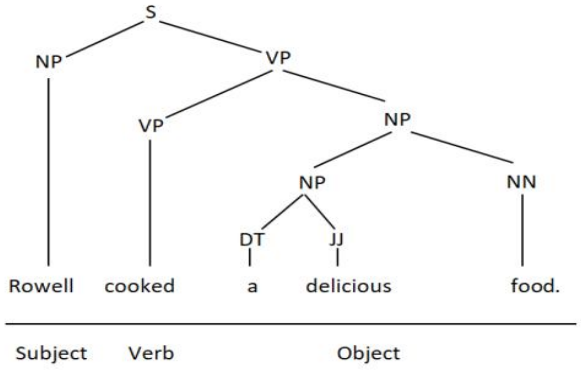

Fig. 2 Parse Tree of an English Sentence

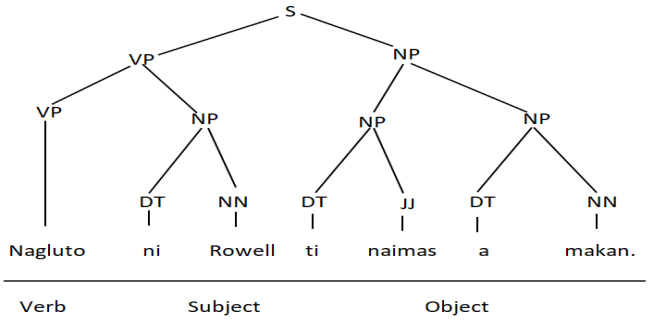

Fig. 3 Parse Tree of an Ilokano Sentence

C. Algorithm

The algorithm of the combination of Rule-based and Example-based is discussed in the following method steps.

STEP 1: Input the source text in English language.

STEP 2: When the sentence to be translated into the target language happens to be found in the example-based database go to step 4, if not skip step 4.

STEP 3: When the phrase of sentence to be into the target language happens to be found in the example-based database and the rest of the sentence is not in the example go to step 4 for the phrase and go to step 5 for the rest of the sentence, if not skip step 3.

STEP 4: Check the procedure according to which EBMT is based is the following:
4.1. The matching of input sentences against phrase (examples) from stored database.
4.2. The selection and extraction of equivalent target language or translated phrases.
4.3. Display the translated sentence.
STEP 5: Pass the source text to a Parser and get the output as (tagged POS).
STEP 6: Based from the output as tagged POS, arrange the English Pattern.
STEP 7: Retrieve the record of this pattern in order to know the subject, verb, and object.
STEP 8: Use the lexicon to translate the English word to Ilocano equivalent.
STEP 9: Apply the rules for verbs and their subjects, adjectives and the entities that they describe, and apply determiners if needed.
STEP 10: Rearrange them by applying proper grammar in Ilokano pattern sentence.

IV. OPERATIONAL FRAMEWORK

A. Materials

Stanford Log-linear Part-Of-Speech Tagger was used in segmentation and tag assignment of each word in the sentence .This software was implemented using PHP so WAMP was installed. The Example-based database was composed of English sentences equivalent to single word in Ilokano. The English-Ilokano lexicon was composed of most commonly used words in English and its equivalent Ilokano term. Aside from that, a lists of grammar rules for English and Ilokano sentence which serves as the foundation of RBMT approach, was used. Lastly, training sets of English sentences with different structures will also be used to evaluate the effectiveness of the algorithm.

B. Methods

The objective of the study is to translate an English sentence into Ilokano sentence using Rule-based and Example-based. Therefore, experimental method was applied to test the combine approaches and to distinguish the distinct relation from the data gathered and results variation.

C. Evaluation

Evaluation plays a major role in the field of Natural Language Processing. Evaluation is necessary for development of Machine Translators. To ascertain the performance of MT system, we employ evaluation process so that we may get precise report of MT development process. Evaluation depends on the subject matter, applied methodology or the application of its results. In general, evaluation can be understood as judgment on the value of a public intervention with reference to defined criteria of this judgment [16]. To evaluate the results of the translation, the reference output was compared to the MT Output.

V. RESULTS AND DISCUSSION

A. Example-Based Machine Translation

English Sentence: I will go Ilokano Translation: Mapanak
In this translation approach, it displayed the Ilokano equivalent of the sentence found in the Example-Based Database. The sentence must be exactly the same as the one in the database to be able to display the result.

B. Rule-Based Machine Translation

English Sentence: Rowell cooked a delicious food

Analysis
Tokens and tags:
token => Rowell
tag => NNP
token => cooked
tag => VBD
token => a
tag => DT
token => delicious

tag => JJ
token => food
tag => NN
Transfer
array (size=5)
'Rowell' => string 'NNP' (length=3)
'nagluto' => string 'VBD' (length=3)
'a' => string 'DT' (length=2)
'naimas' => string 'JJ' (length=2)
'makan' => string 'NN' (length=2)
Generation Ilokano Sentence: nagluto ni Rowell ti naimas a makan

In this translation approach, the sentence has to go three phases such as analysis, transfer, and generation. The analysis phase did the segmentation of sentence into words then gave the equivalent POS of the words. Retrieve a pattern based on the tagged-POS. For the Transfer phase, using the lexicon, it did find the best translation for each word. Finally, it rearranged the words based on the Ilokano pattern in the Generation phase.

C. Mixed Translation

English Sentence: I will go to Ilocos Norte
array (size=2)
'to' => string 'TO' (length=2)
'Ilocos Norte' => string 'NNP' (length=3)
Ilokano Sentence: mapanak idiay Ilocos Norte

The mixed translation approach is the combination of the example-based and rule-based translation. If the sentence composed of phrase found on the Example-based Database, it automatically displayed its Ilokano equivalent. The rest of the sentence will proceed to the Rule-based translation in which it will undergo again Analysis, Transfer, and Generation. Finally, they were combined to display the result.

In order to evaluate translation performance of the study, there are 150 sentences that were translated which are being compared to the reference sentences to the MT sentences. Table 1 shows the accuracy of the translation results. From the results, most of the problems occur from unknown word. Aside from that the lexicon size is quite small.

TABLE 1
ACCURACY OF ENGLISH – ILOKANO TRANSLATOR TABLE

Translation Approach	Sentences	Correct Sentences	Result (%)
Example-based Translation	35	31	88.57 %
Rule-based Translation	90	77	85.55 %
Mixed Translation	25	18	72.00 %
Total	150	126	84.00 %

VI. CONCLUSIONS

This English – Ilokano Language Translator presented a new approach of translating English sentence to its Ilokano equivalent that utilized the strength of Example-based Machine Translation, and Rule-based Machine Translation. It concludes the following:

A. The Example-based approach is very suitable for English – Ilokano Language Translation since there are a lot of single words in Ilocano language that can be expressed in whole sentence in its equivalent English language. It is a direct translation therefore the sentence must be exactly the same with the sentence in the Example-based Database to display the correct result.

B. The Rule-based is the main approach in this study composed of analyzation, transfer, and generation.

1) The analyzation of English sentence structure was done by Stanford Log-linear POS Tagger therefore the shortcomings of the said Tagger is also the shortcomings of the analyzation phase.
2) The lexical transfer aimed to find the best translation for each word. The words were limited only to the lexicon being created.
3) The rearrangement of Ilokano sentence was based on the given pattern grammar rules. The more the grammar rules, the better the generation.

C. The Mixed Translation was dependent only on the Example-based and Rule-based approach. The potentials of this approach were not fully developed in this study since the input sentence was only a simple sentence.

VII. RECOMMENDATIONS

In the future works, more parallel corpus and integrate Statistical Machine Translation in order to improve the accuracy of the translation. Compound and complex sentences to be accepted as input. More pattern rules on the two languages to cover more special cases that would improve the quality of the translated sentence. Further, work can be done to increase the performance of the machine translation.

VIII. ACKNOWLEDGMENT

In conducting a study, much is required. This includes time, money, efforts and determination. A lot of challenges encountered but it is nice to know that there were people who have, in one way or another helped them in the completion of this study. Therefore, I wish to acknowledge with profound gratitude to all who made this humble piece of work come true. I would like to thank Mr. Felizardo Reyes Jr., Dr. Maria Christina Aragon, Mr. Great Allan Ong, Jan Viktor Adora, and Dr. Bert Gamiao for being instrumental in this study.

REFERENCES

[1] Chowdhury, G., “Natural language processing. Annual review of information science and technology”, 37(1), 51-89, 2003.
[2] Raga, Jr. R., “Reflections on the Awareness and Progress of Natural Language Processing Research in the Philippines.”, Philippine Computing Journal Dedicated Issue on Natural Language Processing, pages 1-9, 2016.
[3] Roxas, R. and Borra, A., “Policies for Machine Translation Research & Development in thePhilippines.”, Survey on Research and Development of Machine Translation in Asian Countries, Thailand, May 13-14, 2002.
[4] Javier, C.J et al., “FILSPAN: A Filipino –Spanish Language Translator”, College of Computer Management and Information Technology, 2008.
[5] J. Centelles and M. R. Costa-Jussa, “Chinese-to-spanish rule-based machine translation system”, Universitat Polit`ecnica de Catalunya (UPC), Barcelona and Instituto Polit´ecnico Nacional (IPN), Mexico, 2014.
[6] H. Somers, “Review article: Example-based machine translation”, Machine Translation, vol. 14, no. 2, pp. 113–157, 1999.
[7] P. F. Brown, J. Cocke, S. A. D. Pietra, V. J. D. Pietra, F. Jelinek, J. D. Lafferty, R. L. Mercer, and P. S. Roossin, “A statistical approach to machine translation”, Computational linguistics, vol. 16, no. 2, pp. 79– 85, 1990.
[8] Shashi Pal Singh, Ajai Kumar, Pragya Sahu,Preeti Verma, “Syntax Based Machine Translation using Blended Methodology”, 2016 2nd International Conference on Next Generation Computing Technologies (NGCT-2016) Dehradun, India 14-16 October 2016.
[9] Alawneh, M, Sembok, T, and Mohd, M., “Grammar-Based and Example-Based Techniques in Machine Translation from English to Arabic”, 5th International Conference on Information and Communication Technology for the Muslim World, 2013.
[10] Unlee, P and Seresangtakul, P., “Thai to Isarn dialect Machine Translation using Rule-based and Example-based”, 13th International Joint Conference on Computer Science and Software Engineering (JCSSE), 2016.
[11] Roxas, Rachel and Borra, Allan, “IsaWika! Machine Translation from English to Filipino: A Prototype [Final Report]”. ICS, UPLB, 1997.
[12] Go, Kathleen L., Morga, Manimin R., Nunez, Vince Andrew D., Veto, Francis Germiline S., and Ong, Ethel C, “TExt Translation: Template Extraction for a Bidirectional English-Filipino Example-Based Machine Translation”, 3rd National Natural Language Processing Symposium - Building Language, De La Salle University, Taft Avenue, Metro Manila.
[13] Borra, Allan B., Chan, Erwin Andrew O., Lim, Chris Ian R., Tan, Richard Bryan S., and Tong, Marlon N., “LFG-Based Machine Translation Engine for English and Filipino” College of Computer Studies. De La Salle University. Manila, Philippines, 2007.
[14] Fat, Jacqueline, “T2CMT: Tagalog-to-Cebuano Machine Translation”, Department of Mathematics and Computer Science, College of Arts and Sciences, University of San Carlos, Cebu City, Philippines.
[15] Joshua R. Bautista, Claude D. Bayla, Kwinnie F. Fianza, Dominique B. Mamis, Job L. Tangangco, Jerwin C. Yango, Dalos D. Miguel, “Bi-directional Ilocano-English Language Translator Using Customized Moses Statistical Machine Translation System (SMTS)”, School of Computing and Information Sciences, Saint Louis University, Baguio City, 2015.
[16] Gupta V, Joshi N, and Mathur I., “Subjective and Objective Evaluation of English to Urdu Machine Translation”, Apaji Institute, Banasthali University, Rajasthan, India, 2013