Morphological Analyzer

*Alcantara, Kim D.\*Calupas, Junika S.\*Echas, Jean Paula R.\*Ferreria, Jeremie Elaine P.\*Jucar, Nelly Jane E.\*Ramores, Marjorie Ann B.\*Saballo, Jeffrey III S.\*Yandoc, Kurt Lawrence\*\*Sagum, Ria S.*

*Polytechnic University of the Philippines, College of Computer and Information Sciences, Department of Computer Science*

***Abstract.- This paper presents a morphological analyzer that accepts Filipino verbs conjugated in different forms as inputs and analyzes them to produce the affixes used, the infinitive forms, and the tenses of the original input verbs. A prototype system was implemented and was fed with a file containing 2000+ Filipino verbs conjugated in various tenses using different types of affixes. The preliminary result showed that the accuracy rate was high in three expected outputs, i.e., tenses, infinitive forms, and affixes used.***

*Keywords: morphological analyzer, context-driven machine translation system, Filipino verbs.*

1. **Introduction**

Morphology is a branch of linguistics that deals with word formation, analysis and generation. It is concerned with how words are formed or created in a language from smaller units in a systematic way [1]. It has to find as well as to describe the mechanism behind this process. Each language has morphemes which are its smallest meaningful units. Realization of morphemes as part of a word is called morph. In formation of words from these

smaller units, certain processes such as inflection, derivation and compounding etc. are involved. Also involved is morph tactics that determines how morphs should be put together to form words.

Morphological analysis is the process in which a word is analyzed into its root word and associated morphemes. Morphological analyzers are programs used to provide morphological analysis of a language. They include recognition engine, lexicon/thesaurus, and algorithms to find out stem within an input word and identifying its suffixes. [2]. Filipino text is commonly pre-processed before it is indexed and one of the common pre-processing steps is stemming. Stemming is a process through which affixes are removed and words are converted to their stems [3]. Many different affixes can be used with a single word depending on the tense, gender, number, case, etc. and thus more complex rules for their stemming are needed.

Our Morphological Analyzer uses Natural Language Processing (NLP) to learn about the language morph syntactic structure and uses that knowledge to detect morphemes of Filipino verbs (Pandiwa). The inputted verb will undergo to several process to output there infinitive verb, affix/es and tense.

1. **Background of the Study**

The morphological analyzer presented in this paper is just one of the components of our Context-driven Filipino-English Machine Translation System. The analyzer will be used during the translation process. This morphological analyzer examines any Filipino verb in various possible inflections and produces the affix(es), the infinitive form, and the tense of the input verb. It should be pointed out

that we have no data as to how many infinitive forms are there in Tagalog. We don’t know of any study trying to count the infinitive forms of a certain language.

Natural languages are evolving. So it is difficult to categorically say that a particular language has that number of infinitive forms. The lexicon will have to be updated from time to time as the language

evolves. In this research, we don't bother so much on generating the root word because the entries in our lexicon will contain the infinitive or dictionary forms of the verbs and will be accessed using that form. We do, however, recognize the value of getting the root of a word, as being done in TagSA and TagMA, but extracting the root of the word is more useful in Information Retrieval System (IRS) than in a machine translation system. We also understand that the root of a word is used in some systems to get the right semantics of compound words in order to come up with a better translation. This is not necessary in our proposed Context-driven Filipino-English Machine Translation System, of which the morphological analyzer being presented here is just a subsystem. We choose the right English equivalent of a Filipino word by checking the neighboring words in a sentence and even in a paragraph. These neighboring words serve as context of the word being translated. So our lexicon must be a different one than the usual dictionary. The affixes are important because they help in determining the structure of the sentence, the meaning of the entire sentence, etc. The subject of the sentence as well as the object will be easily recognized once the affix has been determined already. For example, if the affix used in the main verb is “UM,” the subject of the sentence starts with the word “ang,” “si” for personal subject, or uses the nominative case of pronoun. The object starts with the word “ng,” “ni” for personal object, or uses the objective case of pronoun. This will be very helpful in the translation itself. As for the tense of the verb that will be output by the analyzer, it is also very useful for the translation process.

1. **Related Work**

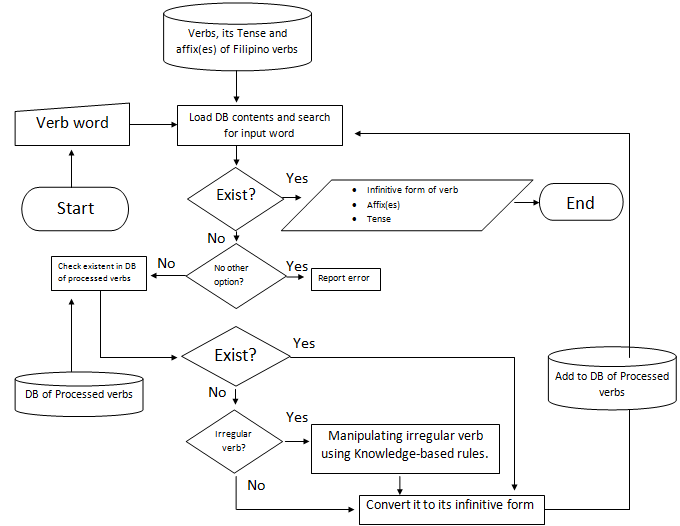
When someone attempts to create any machine translation system for natural languages, it is not possible to do away with morphological analyzers. One area of focus for morphological analyzers is the analysis of verbs. Verbs are usually conjugated according to tense, number, voice, and mode while nouns and adjectives are usually declined according to person, number, case, and degree. Some languages have simple forms of verb inflection or conjugation while other languages have complex forms of inflection. Japanese was originally thought to have simple verb inflection, thus it was not the central subject on Natural Language Processing (NLP). But in recent past, it had become an important subject (Hisamitsu and Nitta, 1994). Forother languages, verbs are also given much attention in research. These languages include Chinese (Kim, et. al, 2002), Japanese (Nakamura, 2007), and Korean (Hong, et. al., 2004; Jun, 2007) to name a few.

For languages that have complex ways of adding affixes and duplication of syllables or words, an excellent morphological analyzer is highly necessary. Filipino or Tagalog is one the most complex languages in the world. So several researches have been conducted in the area of morphological analysis for the Filipino language. One is the TagSA (Tagalog Stemmer Algorithm), which tackles on the extraction of the stem of any Tagalog word (Bonus, 2003). Another morphological analyzer, called TagMA (Tagalog Morphological Analyzer), is devoted to extracting the root word both for concatenative and non-concatenative formation (Fortes, 2002). A system called T2CMT (Tagalog-to-Cebuano Machine Translation) was developed (Fat, 2004) that used TagMA and TagSA for its morphological analyzer. Even before the TagMA and TagSA, there was already a morphological analyzer that was created and used in a prototype system that supported English-Filipino and Filipino-English machine translation system (Roxas, 1998).

TagMA (Fortes, 2002) produces three morphological structures (morpheme, CV, and

syllabication) to represent an input verb. To do this, it needs to scan the entire word in several stages. It starts by assigning the symbols “C” for consonants or “V” for vowels to the morpheme structure character by character to get the CV structure. Then it scans the CV structure character by character to assign some codes (0-2) to a “C” or “V” to get the syllabification structure. Then the input representation is fed to the GEN function in order to produce a candidate set. It then scans the input string syllable by syllable until the last syllable is encountered. Then the result is subjected to the EVAL function, where the output of the GEN function is checked against the two lexicons (root and affixes) and subjected to some constraints to be able to get the right root of the verb. Although TagMA was able to morphologically analyze 96% of the sample verbs accurately (Fortes, 2002), the process of analyzing an input verb is quite long and tedious. It only outputs the root, tense, and its affix for a particular input verb. It does not produce the infinitive or dictionary form of the original input verb. The infinitive form is also useful because that is what one usually uses to lookup a word in a dictionary. In fact, T2CMT (Fat, 2004) that used TagMA wrongly translated the Tagalog word “namatay” (to die) to “pinaagi” (by means of or through) in Cebuano, when in fact, the Cebuano translation should be “namatay” also. If a morphological analyzer produces the infinitive form, the translation would be correct. Our proposed system uses the morphological analyzer developed in (Roxas, 1989), which is being extended to cover more possible verb conjugations. If you pass “namatay” to our morphological

analyzer, it will give the infinitive form “mamatay” and tense is past or perpektibo. If you consult a dictionary, you will find that its English translation is “to die,” which is the intended meaning. Therefore, we believe that there is still a need to develop a morphological analyzer that uses the infinitive forms when checking a word in a dictionary.

**IV. System Architecture**

This morphological analyzer examines any Filipino verb in various possible inflections and produces the affix(es), the infinitive form, and the tense of the input verb.

**3.1 Extracting Affixes, Infinitive Forms, and Tenses**

In getting the infinitive form, the tense, and affix(es) of a Filipino verb, it first consults the lexicon if the word is there. If not, it checks whether or not all possible options were exhausted, in which case, it reports an error. If there are still options, it checks whether or not it is the first time to process the verb. If it is, it has to check whether or not the verb is irregular. If it is, it performs some manipulation and prepares for the formation of the possible infinitive form of the verb. If the verb is regular, the analyzer immediately tries to form the possible infinitive form.

The analyzer performs a certain number of passes until it finds the correct infinitive form of the input verb. If all options have been tried but still the generated infinitive is not found in the lexicon, it reports an error. For each pass, it tries to generate the infinitive form of the input verb. If the generated infinitive form does not exist in the lexicon, it is possible that the option used was not the correct one. So it tries another option. Eventually, it will generate the correct infinitive form. Then it reports the infinitive form, the tense, and the affix(es) used. Filipino verbs use one or more affixes. So it is possible that more than 1 affix is reported.For irregular verbs, they need to be modified a bit so that when they are passed to the section that generates the infinitive forms, it will generate the correct infinitive forms.

**V. Results and Discussion**

The prototype morphological analyzer was fed with 1,402 Filipino verbs conjugated in three different tenses plus some verbs used in everyday life. The verbs belong to the UM-, -IN, -AN, I-, MA-, and MAG- categories only. Table 2 shows the preliminary results. In determining the tense of the input verbs, it returned 98.79% accurately, which is very promising. The 2.21% error for determining the tense was primary caused by the irregular verb forms.

Table 2: The result of the initial testing.

|  |  |  |  |
| --- | --- | --- | --- |
| *Expected Output* | *Correct* | *Error* | *Undetermined* |
| Tense | 98.79% | 2.21% | 0% |
| Infinitive Form | 92.08% | 7.92% | 0% |
| Affix | 98.72% | 2.28% | 0% |
| Root Word | 84.45% | 15.55% | 0% |

In determining the infinitive forms of the verbs, it got a 92.08% accuracy rate, which is also good as initial results. The error rate was 7.92% and almost all of them resulted from irregular verb forms. As for determining the affixes of the verbs, the accuracy rate is only 98.72%. The error rate of 2.28% was caused partly by irregular verb formation but a large part of it was caused by the presence of “in” or “ni” that are always found any “I-“ verbs in their past and present tenses. And for determining the root words of the verbs, the accuracy rate is 84.45 %. The error rate of 15,55% was caused by failing to change the “o” to “u” of the penultimate syllable and for using of Tri-gram only.

Many of the errors will be greatly reduced by refining the algorithm. It should be noted that the testing was just a single-pass because the lexicon is not yet fully functional. We can expect a higher accuracy rate than what is presented here, if the multi-pass approach will be used because some errors above will be solved by the multi-pass approach.

**VI. Conclusion and Recommendation**

A morphological analyzer that analyzes the Filipino verbs and produces their affixes, infinitive forms, and tenses has been presented. The prototype was tested with 1,050 verbs (regular and irregular) conjugated in three different tenses and the initial results showed high accuracy rate for determining the tense, the infinitive form, and the affix used. The higher accuracy rate is expected to even improve once the lexicon is in place and the multi-pass approach is already implemented.

Future work includes the implementation of the lexicon that employs an innovative way of getting the appropriate meaning to support context-driven machine translation system. We will also add other verb forms that were not yet included in the current prototype and perform further tests when all different forms have been taken into consideration. We will also increase the data set in the next testing.

**VII. References**

<http://www.mathcs.duq.edu/~packer/Courses/Psy598/Ling-Morphology.pdf>

<http://thiscourse.com/ucsd/lign101/su11/docs/59089132/>

<http://sanskrit.jnu.ac.in/rstudents/mphil/muktanand/Chapter%201.pdf>

<http://allthingslinguistic.com/post/40962717566/protolinguist-resources-teaching-yourself-morphology>

<http://www.blackwellpublishing.com/content/BPL_Images/Content_store/Sample_chapter/0631203184%5CAronoff_sample%20chapter_What%20is%20morphology.pdf>

http://www.acroterion.ca/Morphological\_Analysis.html

Akshar Bharati, Vineet Chaitanya and Rajeev Sangal. "Natural Language Processing: A Paninian

Perspective", Prentice-Hall of India, New Delhi, 1995.

Authors

**Kim D. Alcantara** was born on October 3, 1993. He is 4th year student of Polytechnic University of the Philippines and taken up Bachelor of Science in Computer Science. Has a basic knowledge in computer programming in different languages such as C++, Java and C#. He is also proficient in using some Microsoft Office tools. He has a good communication skill and a good documentary.

**Junika S. Calupas** was born on August 12, 1994. She is recently taking up her College degree in Computer Science from Polytechnic University of the Philippines, Manila. She knows how to code in Turbo C programming, C#, and Cobol. She also has knowledge in database MS Access as well as MS SQL and has a background in using Microsoft Office tools such as MS Word, Publisher, Excel, etc.

****

**Jean Paula R. Echas** was born in Munioz, Quezon City on January 12, 1995. She is a 4th year student of Polytechnic University of the Philippines, Manila and taken up Bachelor of Science in Computer Science. She is interested in Web Designing, Photo Editing and has knowledge in Database MS Access as well as MS SQL

**Marjorie Ann B. Ramores** was born in Caloocan City on February 04, 1994. She is a 4th year student of Polytechnic University of the Philippines and currently taking up. Bachelor of Science in Computer Science. Has a basic knowledge in computer programming in different languages such as C++, Java Cobol and C#.

**Jeremie Elaine P. Ferreria** was born on July 27,1993 at San Juan, Manila. She is a graduating student in Polytechnic University of the Philippines taking up Bachelor of Science. She is interested in photo editing, web designing and has a basic knowledge in Cobol, C-Sharp and MS Office tools.

**Kurt Yandoc** was born on March 8, 1994.He is 4th year student of Polytechnic University of the Philippines and taken up Bachelor of Science in Computer Science. Has a basic knowledge in computer programming in different languages such as C++, Java and C#. He is also proficient in using some Microsoft Office tools.

**Nelly Jane Jucar** was born on November 16, 1993. She is a 4th year student of Polytechnic University of the Philippines and currently taking up Bachelor of Science in Computer Science. Has a basic knowledge in computer programming in different languages such as C++, Java Cobol and C#.

**Jeffrey Saballo III **was born on December 15, 1992. He is 4th year student of Polytechnic University of the Philippines and taken up Bachelor of Science in Computer Science. Has a basic knowledge in computer programming in different languages such as C++, Java and C#. He is also proficient in using some Microsoft Office tools. He has a good communication skill and a good documentary.