e-Wika: Philippine Connectivity through Language

Rachel Edita O. Roxas
Software Technology Department
College of Computer Studies
De La Salle University-Manila
2401 Taft Avenue, Manila Philippines
Tel: (63) (2) 536-0276/7
Fax: (63) (2) 536-0278
roxasr@dlsu.edu.ph

ABSTRACT
In this paper, we present what we have attempted towards connecting the Philippine islands through the digitalization of the Philippine languages and their respective applications, and what we intend to do in the future. We present the development of a multi-engine bi-directional English-Filipino Machine Translation (MT) system, and the building of various language resources and tools for this system.

Though linguistic information on Philippine languages are available, as of yet, the focus has been on theoretical linguistics and little has been done on the computational aspects. We report here our attempts in the manual construction of these language resources such as the grammar, lexicon, morphological information, and the corpora which were literally built from almost non-existent digital forms. Due to the inherent difficulties of manual construction, we also discuss our experiments on various technologies for automatic extraction of these resources to handle the intricacies of the Filipino language, designed with the intention of using them for the MT system.

We explore here what will be our possible future directions in terms of connecting the Philippine islands through the digitalization of our Languages.

Keywords
Corpora; language resources; language tools; lexicon; machine translation; morphology.

1. INTRODUCTION
The Philippines is an archipelago of more than 7,100 islands, with over 100 languages. The 1935 Constitution Article XIV, Section 3 states that "...Congress shall make necessary steps towards the development of a national language which will be based on one of the existing native languages..." due to the advocacy of then Philippine President Manuel L. Quezon for the development of a national language that will unite the whole country. Two years later, Tagalog was recommended as the basis of the national language, which was later officially called Pilipino. In the 1987 Constitution, Article XIV, Section 6, states that "the National language of the Philippines is Filipino. As it evolves, it shall be further developed and enriched on the basis of existing Philippine and other languages."

In 2000, I presented a paper entitled “Towards a Multi-lingual Machine Translation System for Philippine Languages” which was delivered as a Henry Sy Professorial Chair Lecture at DLSU, Manila, Philippines [18]. It was a vision towards the development of a multi-engine bi-directional machine translation (MT) system that is a computer software that performs machine (or automatic) translation of English texts to Philippine languages, and vice versa. For the past three years, my NLP research team is coming to the end of the first stage of providing a hybrid English-Filipino MT system which basically involves one Philippine language, Tagalog, as our Filipino language basis. But I believe that through the collaboration of many of us, that dream will come to fruition. I discuss here what we have already done towards the fulfillment of this dream, and what can we do as NLP researchers and citizens of this nation to connect Luzon, Visayas and Mindanao through the digitalization of our languages.

2. THE MACHINE TRANSLATION ENGINES
Natural language translation is a very complex task. Its automation presents more issues and difficulties in addressing, in particular, translation quality. During the past 50 years of research on machine translation involving various languages, different paradigms and methods have been suggested and employed to improve translation quality. The issue is: how effective is an approach in capturing the features of natural languages and the translation phenomena between these two languages so that translation quality can be considered at par with the translation of human experts? It has been shown in previous researches that this can not be achieved by a single MT paradigm. Hence, the multi-engine MT research attempts to integrate the various MT paradigms so as to draw a synergy out of the strengths of these paradigms. The current MT system outputs the MT engines’ respective generated translations and presents them to the user.

Figure 1 illustrates our multi-engine English-Filipino MT system which considers the rule-based and corpus-based approaches. Rule-based MT builds a database of rules for language representation and translation from linguists and other experts; while corpus-based MT automatically learns such information from sample text translations. Two corpus-based approaches are considered: example-based and template-based. An expansion slot allows for the integration of other MT engines into the system.
in the future. To implement these paradigms, language resources are built, such as the bilingual English-Filipino lexicon (or electronic dictionary), Filipino grammar, translation rules and annotated corpora. The corpus-based approaches require a part-of-speech tagged corpus. The MT paradigms will also use the morphological analyzer and generator to extract root words from texts. The next phase of the project is the language modeler that would choose the “best” translation, either as-is or integrating the outputs of the various engines.

One of the major challenges is in the building of language resources such as the lexical and morphological information, lexicon, grammar, corpora and translation rules from almost non-existent digital forms. Linguistic information on Philippine languages are available, but as of yet, the focus has been on theoretical linguistics and little is done on the computational aspects of these languages. Computational issues involving Philippine languages, in particular Filipino, are considered in this study. Some of these issues include the complex verbal morphology of the language, free word order of sentences, and focus of the sentence.

The project considers various approaches in capturing natural languages and their intrinsic features and characteristics, and automating the processes involved in the representation and translation of natural languages. The approach will have to consider the effective representation of both the English and Filipino languages, and their translation. In all the components of the MT system, both the rule-based and the corpus-based approaches are considered. The details of these are discussed in the succeeding subsections.

2.1 The Rule-Based MT Engine

The rule-based MT builds a database of rules for language representation and translation rules from linguists and other experts on translation from English to Filipino and from Filipino to English. We need a mechanism to capture the languages and their corresponding translations.

We have considered lexical functional grammar (LFG). Given a sentence in the source language, the sentence is processed and a computerized representation in LFG of this sentence is constructed. An evaluation of how comprehensive and exhaustive the identified grammar is to be considered. Is the system able to capture all possible Filipino sentences? How are all possible sentences to be represented since Filipino exhibits some form of free word order in sentences? The next step is the translation step, that is, the conversion of the computerized representation of the input sentence into the intended target language. After the translation process, the computerized representation of the sentence in the target language will now be outputted into a sentence form, or called the generation process. Although it has been shown in various studies elsewhere and on various languages that LFG can be used for analysis of sentences, there is still a question of whether it can be used for the generation process. The generation involves the outputting of a sentence from a computer-based representation of the sentence. This is part of the work that the group intends to address.

The major advantage of the rule-based MT over other approaches is that it can produce high quality translation for sentence patterns that were accurately captured by the rules of the MT engine; but unfortunately, it cannot provide good translations to any sentence that go beyond what the rules have considered.

2.2 The Corpus-Based MT Engines

In contrast to the rule-based MT which requires building the rules by hand, the corpus-based MT system automatically learns how translation is done through examples found in a corpus of translated documents. The system can incrementally learn when new translated documents are added into the knowledge-base, thus, any changes to the language can also be accommodated through the updates on the example translations. This means it can handle translation of documents from various domains [1].

![Figure 1. Over-all Architecture of the Hybrid MT System.](image-url)
Another way to capture trends in the example translations is through templates, so instead of rules, templates are saved into the knowledge-base. Templates are patterns that are derived from the translation examples [1].

The principle of garbage-in-garbage-out applies here; if the example translations are faulty, the learned rules will also be faulty. That is why, although human linguists do not have to specify and come up with the translation rules, the linguist will have to first verify the translated documents and consequently, the learned rules, for accuracy.

It is not only the quality of the collection of translations that affects the overall performance of the system, but also the quantity. The collection of translations has to be comprehensive so that the translation system produced will be able to translate as much types of sentences as possible. The challenge here is coming up with a quantity of examples that is sufficient for accurate translation of documents.

With more data, a new problem arises when the knowledge-base grows so large that access to it and search for applicable rules during translation requires tremendous amount of access time and to an extreme, becomes difficult. Exponential growth of the knowledge-base may also happen due to the free word order nature of Filipino sentence construction, such that one English sentence can be translated to several Filipino sentences. When all these combinations are part of the translation examples, a translation rule will be learned and extracted by the system for each combination, thus, causing growth of the knowledge-base. Thus, algorithms that perform generalization of rules are considered to remove specificity of translation rules extracted and thus, reduce the size of the rule knowledge-base.

3. COMPUTATION LEVELS OF LANGUAGE REPRESENTATION

The Filipino alphabet consists of the Filipino ABAKADA, which consists of 20 letters, and another 8 letters (C F J Ñ Q V X Z) for the assimilation of borrowed words.

3.1 Morphological Processes

We consider rule-based and example-based approaches for both morphological analysis and generation. As with the MT systems, the rule-based approach requires the capturing of the morphological phenomena through the expertise of a human linguist, and transforming them into rules that can be represented in a computer, while the example-based approach automatically extracts morphological behavior from examples fed into the system. Since the Filipino language has very complex word morphology, the rule-based approach can be complemented by the example-based approach.

Tagalog language exhibits complex morphological phenomena, which include concatenative behavior such as prefixation and suffixation, and non-concatenative behavior such as infixation and reduplication (either partial or full). Tagalog is also an agglutinative language, where for instance, prefixes may be successively combined. Research on computational morphology has been predominantly on concatenative morphology and on finite-state models of morphotactics [12]. Although attempts were made to handle non-concatenative phenomena, it has been on a limited capacity only [2; 3].

We experimented on a rule-based MA using Optimality Theory (OT) [9], since OT has been proven effective in generating non-concatenative phonology. Test results show a 96% accuracy when on 1,600 Tagalog verb forms (having 3 to 7 syllables) from 50 Tagalog roots which exhibit both concatenative and non-concatenative morphology. The 4% error is attributed to d-r alteration, an example of which is in the word lakaran, which is from the root word lakad and suffix -an, but d is changed to r. Unfortunately, time efficiency is affected by the exhaustive search performed since all candidates are generated, and erroneous ones are later eliminated through constraints and rules. To augment the rule-based MA, we also experimented on an example-based MA by extending the WordFrame model [20] from a seven-way split representation of morphological re-write rules from word pairs of a morphed word and its corresponding root. In the WordFrame model, the seven-way split re-write rules composed of the canonical prefix/beginning, point-of-prefixation, common prefix substrings, internal vowel change, common suffix substring, point-of-suffixation, and canonical suffix/ending. Infixation, partial and full reduplication are improperly modeled in the WordFrame model as point-of-prefixation as shown in the word (hin-)intay which should have been modeled as the word hintay with infix –in-. Words with an infix within a prefix are also modeled as point-of-prefixation as in the word (hin-)hlintay which should be represented as infix –in in partial reduplicated syllable hi-. In the revised WordFrame model, the non-concatenative Tagalog morphological behaviors such as infixation and reduplication are modeled separately and correctly. Unfortunately, it is still not capable of fully modeling Filipino morphology since some occurrences of reduplication are still represented as point-of-suffixation for various locations of the longest common substring. There are also some problems in handling the occurrence of several partial or whole-word reduplications within a word. Despite these problems, the training of the algorithm that learns these re-write rules from 40,276 Filipino word pairs derived 90% accuracy when applied to an MA. The complexity of creating a better model would be costly but it would ensure an increase in performance and reduced number of rules [5].

Work is still to be done on exploring techniques and methodologies for morphological generation (MG). Although it could be inferred that the approaches for MA can be extended to handle MG, the MG process can be ambiguous such that many various surface form of words can be generated from one underlying form.

3.2 The Filipino Grammar

The Filipino grammar was defined manually by formalizing observable conceptual rules and patterns of the language upon consultation with linguists, and specified using the lexical functional grammar (LFG) formalism. The main building block of LFG is the context free grammar which handles the syntax level of the language and the additional semantic actions that capture the semantic categories of words in the sentence such as subject and object.

One of the major challenges of the Filipino language is its free word order in sentence formation, wherein one sentence in
English can be translated to various sentences in Filipino. For instance, the English sentence *The man bought an umbrella from the store* can be translated into several different Filipino sentences while maintaining the semantics of the original English sentence:

*Ang lalaki ay bumili ng payong sa tindahan.*
*Bumili ang lalaki ng payong sa tindahan.*
*Bumili ng payong ang lalaki sa tindahan.*
*Binili ng lalaki ang payong sa tindahan.*
*Binilhan ng lalaki ng payong ang tindahan.*

Also, the construction of the sentence depends on the focus of the sentence. For the same sentence above, some of the possible translations in Filipino with varying foci would be as follows:

*Bumili ang lalaki ng payong sa tindahan.*
*Binili ng lalaki ang payong sa tindahan.*
*Binilhan ng lalaki ng payong ang tindahan.*

The foci in these Tagalog sentences are the *man*, the *umbrella* and the *store*, respectively. To add, within phrases, the Filipino language also exhibits the free word order phenomenon for adjectival and noun phrases.

Because of this free-word order phenomenon in Filipino sentences, grammar representation and translation mapping from source to target language, that is English to Filipino, and vice versa, also pose problems of one-to-many and many-to-one relationships, respectively.

To add, morphological information play a crucial role in the accurate extraction of information from the surface form words, since it also determines the focus of the sentence and the over-all semantics. Other issues include the same grammar representation but various sentence interpretations which are based on the relationship of the action and the object.

### 3.3 Translation Rules

Transfer (or translation) rules are based on the lexical functional grammar specification of the two languages and are bi-directional. Rules can be specified by hand (rule-based), or can be derived automatically from a set of examples.

For our example-based MT system, transfer rules are learned using the general framework used by [6; 19], which has three steps: seed rule generation, compositionality, and generalization. In seed rule generation, seed rules that define the token sequence, expressed as a combination of POS tags and possibly constant words, token constraints and alignment scheme of a translation pair are generated. Compositionality infers rules of higher syntactic structure, that is, constituent labels are deduced by using the longest adjacent POS tags found in both the English and Filipino rule. The system groups together similar rules and generalizes it to encompass a wider range of unseen examples. In general, we have achieved 68% accuracy of translated sentences when all other information in morphology, lexicon and tagging are correct. The 32% error can be attributed to lack of information in the lexicon for possible disambiguation by the semantic analyzer [1].

Another way to capture trends in the example translations is through templates and chunks, so instead of rules, templates and chunks are saved into the knowledge-base. Sentence templates are used for translation, and when no appropriate templates are available, phrase translation chunks are used for the construction of the translations. We adopted the machine learning techniques to implement the similarity template learning algorithm performed by [6], and introduced template refinement and derivation of templates from previously-learned chunks. Test results show that strict chunk alignment with splitting (SCAS) in training, instead of loose chunk alignment, wherein correspondences are not required, improves quality of learned input. Test results show an accuracy of up to 96% in the testing of a translation corpus. Our results show possible extensions on a more stringent match disambiguation procedure, an algorithm that accepts chunks with empty contents, a feasible chunk refinement method, more comprehensive linguistic resources, or an approach to Cicekli’s difference template learning algorithm [11]. This is used during the template extraction phase, to ensure that the system does not learn templates whose fixed elements contain only common or noise words, e.g., *The X ⇔ Ang X*. This also prevents the system from learning long chunks for X, which would be very difficult to use during translation.

### 4. LANGUAGE RESOURCES

In this particular NL application, we require a lexicon and corpora. We describe here briefly these resources.

#### 4.1 The Lexicon

The lexicon (or dictionary) is a collection of source words with the corresponding translation in the target language, and their features (such as part-of-speech tag, sample sentences, and semantic information). The base lexicon for the Multi-engine MT project is the dictionary of the Commission on the Filipino Language (CFL) of the Philippine government, which contains approximately 10,000 English source words with a total of about 25,000 Filipino meanings. Each English entry in the lexicon contains the POS tag of the word and the corresponding Filipino meaning, with co-occurring word information, and other English features or attributes (e.g., other acceptable spellings or synset id based from WordNet) for proper translation. Semantic features address the problem of ambiguity since many words have many-to-many relationships and determining the proper translation is vital. A study is currently being explored in building a Tagalog WordNet so that the base lexicon can be extended to handle semantic features as in the WordNet.

But since languages are in the process of evolution, it is imperative that the project provides some way to be able to determine and capture new words and probably new meanings of words in the languages considered in this study. New terms can be added into the base lexicon through automatic lexicon extraction from documents on English and Filipino. New lexicon terms are automatically learned from sample documents. Two approaches have been experimented on using parallel documents (translations of each other) and non-parallel documents (non-translations of each other). In using parallel English-Filipino corpora in the automatic lexicon extraction [13], a probabilistic approach is used to determine candidate translations of English and Filipino words through the principle of co-occurring words in parallel English-Filipino documents, since it is assumed that semantically related words usually co-occur (i.e., words that appear together) in documents in a language and would also co-occur in documents in a different language. Only 57% accuracy is derived from training using about 40,000 words, since there are
other sources of errors such as tags which were not verified by linguists.

In using the non-parallel comparable (that is, non-parallel but within the same domain) corpora, we derive 50% accuracy of extraction of translation terms of the source word to its equivalent target word using a corpora within the same domain with 381,553 English and 92,610 Tagalog terms, with 4,817 and 3,421 distinct root words, respectively [19]. An improvement introduced in this study involves the use of clustering algorithm [15] to group together similar senses of a word using word context extraction [17] and other clues like the part of speech tags in the source corpora. It 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. Errors are attributed to quality of corpora, preprocessing errors, and the inclusion of some function words in the target corpora.

We have derived further improvements to the lexicon extraction algorithm through several aspects. The most important of these improvements is the use of a seed lexicon in the preprocessing phase. We claim that introducing the seed lexicon’s properties to the whole process early produces better similarity measurement results, since most of the ‘noise’ are temporarily ‘silenced’. In this way, we ensure that only the necessary features are made available during the extraction process. Another major difference of the new scheme is its candidate word translation evaluation, wherein a multi-pass procedure is performed. We perform iterative similarity measurements on each word pair as the bilingual dictionary is updated. This is to further solidify the hold of top performing candidate translations on their initial positions. The argument is that, if a candidate translation does not change its similarity measurement over multiple iterations relative to newly introduced candidates, then it must be the most correct translation. Since the search is exhaustive in nature, the accuracy is higher, but then computation time is more.

In the representation of assimilated words in the lexicon, there may be various entries of words in the lexicon when following various spelling conventions [7]. For instance, the word *congregation* can be spelled as *konggregasyon* or *kongregasyon*.

### 4.2 The Corpora

Our work requires both parallel English-Filipino corpora and mono-lingual corpora. The example-based MT engines require parallel English-Filipino corpora, while components of the system such as the part of speech Filipino tagger only requires Filipino corpora. We have built bilingual parallel English-Filipino corpora which consist of 207,000 words, and currently half of the Filipino corpora creation are explored. AutoCor is our method for the automatic acquisition and classification of corpora of documents in closely-related languages [8]. It is an extension and enhancement of CorpusBuilder, a system that automatically builds specific minority language corpora from a closed corpus [10]. We address a problem with the Tagalog corpus generated by CorpusBuilder which contains documents in other closely-related languages to Tagalog and not Tagalog documents alone. We used the query generation method odds ratio which was reported to produce best results in CorpusBuilder, and introduced the concept of common word pruning to the language models of closely-related languages, which was found to improve the precision of the system. The method was tested on 3 most closely-related languages in the Philippines, namely: Bicolano, Cebuano and Tagalog. Each of the target languages was tested for query lengths 1 to 5, with 100 generated queries per query length, both with and without common word pruning. Results show that common word pruning improved the precision of the system by 53% for Bicolano, 18% for Cebuano, and 20% for Tagalog for query lengths 4, 1 and 2, respectively.

We developed the Filipino tagset with 9 general POS tags, 60 specific POS tags and 5 other tags (for punctuation and currency symbols), based on the Penn Treebank tagset which contains 36 POS tags and 12 other tags [14]. We identified tags for nouns, pronouns, determiners, adjectives, conjunctions, verbs, adverbs, cardinal number, and punctuation marks. One of the main differences in the tagset of English is that the tag for the word *ay* which is literally translated to *is* in English. In the sentence *Ako ay kumakain* (*I am eating*), the Tagalog *ay* is not considered as a verb but as a lexical marker since it only signifies that the sentence is in the form *subject + ay + predicate* and is not in the natural order. The natural rendition of this sentence is *Kumakain ako* which is of the form *predicate + subject*. Also, singular and plural personal pronouns have distinct tags, while gender in Tagalog pronouns such as *siya* (*he or she*) is not captured. Interjections are also tagged in Tagalog. The English *if* which is translated to *kung* in Tagalog is tagged as a conditional adverb rather than a conjunction.

While our linguists are manually tagging our corpora, we attempted to build automatic POS taggers as well. The example-based tagger requires a large corpus where words are associated with the corresponding part of speech tags, while the rule-based tagger based on Brill’s tagger [4] also learns tagging rules from examples. Initial tagging can be done through our template-based Tagalog tagger [16] since accuracy of this tagger is only 83% for general tags and 77% for specific tags, and verification of tags are done manually so that tags can be confidently considered correct.

### 4.3 Summary

Manual and automatic creation of language resources are explored for lexical and morphological knowledge bases, lexicons, grammar, corpora and translation rules. The accuracy of the language translation is largely dependent on the comprehensiveness and correctness of the language resources for Filipino and for English-Filipino translation, thus, great pains are exerted to build as much as we can and as accurately as we can. But, we are a long way off from our target. For example, in an automatic lexicon extraction study conducted by [17] from non-parallel texts, the corpus used had as much as 163 million words, which were taken from German and English newspapers articles. To date, the multi-engine MT project has built the corpus for Filipino of 207,000 words, but only half of the tags so far are verified by expert linguists.

Interdependence of language resources is also a major problem in this study. For instance, the part-of-speech tagger for Filipino which uses a stemmer requires a lexicon. Due to minimal resources, excluding the lexicon in the part-of-speech tagger
architecture caused many errors in the final analysis of the tags generated by the system.

Efforts are made in the building of language resources for the implementation of a multi-engine English-Filipino machine translation system, from almost non-existent digital forms. We consider here the computational aspects of these languages since there are already sufficient materials that address the theoretical aspects. We address both manual and automatic constructions of these language resources, problems associated with these and the solutions provided.

5. NL APPLICATIONS FOR CONNECTIVITY

Our English-Filipino machine translation system has given us a hands-on experience on an NLP application and the necessary linguistic resources for this particular application. I contend and propose that we come up with NL applications that will involve as many Philippine languages as possible and as many NL researchers all over the country, as well as being socially-relevant. We also explore ways in which we can use and re-use existing language resources and tools on Philippine languages that we have already built for the successful implementation of these NL applications.

One such application is the e-Laban sa Kahirapan which is a project being proposed for the initial work to be done on an ICT initiative towards poverty alleviation. It involved ICT improvements to the current community-based monitoring system (CBMS) which stems from the growing demand for a regular source of up-to-date information that is delegated at the community level. The Philippines has been involved in the development of a CBMS network in Asia and Africa, with IDRC-Canada funding. However, much work has to be done to improve the current Philippine CBMS. The project aims to point out that the “next wave” of ICT for Development (ICT4D) initiatives in governance is on leveraging knowledge capital and ICT. It is a strong belief of the proponents that these elements can empower community stakeholders (LGUs, POs, NGOs, etc.) for the purpose of socio-economic development.

The project will directly benefit the communities, especially those who are currently involved in poverty alleviation initiatives. Other stakeholders that would directly benefit from the project are the Local Government Units (LGUs), which can use the portal as a leverage tool to better allocate government resources. The proposed portal will act as a “one-stop shop” or “single-window entry” for monitoring & evaluating community development initiatives. Aside from the community and local executives, other interested stakeholders are non-government organizations (NGOs), financial institutions and donor agencies whose task is the advancement of sustainable economic development in rural communities.

We envision a system that includes an NL generation component that will output NL in various Philippine languages depending on the choice of the user. It could include a question-answering system and a report generation system. Natural language resources would include a lexicon and grammar knowledge-base, and NL tools would include an MA and an MG, a NL parser and sentence generator.

To be able to successfully undertake this project and other projects of this nature with the purpose of connecting the Philippine islands through language, we need the involvement of researchers from various parts of the country who are willing to work on automation of their respective local languages, and various applications that we intend to develop for these languages.

6. ACKNOWLEDGMENTS

This project is funded by the Philippine Council for Advanced Science and Technology Research and Development, Department of Science and Technology, Philippine Government. The author would also like to thank the other project leaders of the Hybrid MT project, namely: Allan Borra, Charibeth Cheng, Nathalie Rose Lim, Ethel Ong, Solomon See and Michelle Wendy Tan. Thanks is also extended to all our undergraduate and graduate students who have helped us, in one way or another, towards the development of the English-Filipino Machine Translation system and the needed language resources described in this paper.

7. REFERENCES


