[*Translating and the Computer 25*, November 2003 [London: Aslib, 2003]

The Philippine Style Chat - Instant Messenger Translation System

Dr. Imelda P. de Castro
Associate Professor
AB Translation Studies
De La Salle University - Manila, Philippines
e-mail: decastroi@dlsu.edu.ph

Introduction

The irreversible process of globalization and Internet revolution has changed the way people live their lives. The global spread of large-scale improvements of transportation and communication technologies has made the global world smaller than ever.

In relation to this, the Internet- especially the World Wide Web- has eliminated the barriers for communication and interaction with people from different countries. Unlike other communication and tools, the Web is available to any person in the world with an Internet connection. However, although the Web traces its roots from the United States, it reaches out to the four corners of the globe, wherein plenty do not know how to speak English. In fact, Internet research for IDC reports that within four years, only thirty percent of the Internet users will speak English as their first language. Because of the trend, translation becomes an important medium to communicate with various users on theWeb [LANN2001].

Language Translation, though an emerging sector, has shown progress in recent years. Part of its improvement lies in the development of productivity enhancing linguistic tools. Machine Translation is one of the major tools that analyzes and converts text from a source language to the target language. It does not require human intervention and preferable when integrated with organizational workflows for best results [LISA2001].

Future trends facing the industry include the need to use technology, process integration and develop a new tool. One development in that direction is the advent of Multilingual Chat Translation systems. These systems offer an opportunity to chat online with colleagues and friends that do not speak the same language. It is faster, reusable and cost effective than human translation [LISA2001].

Statement of the Problem

Businessmen encounter problems communicating with countries that speak in a different language.

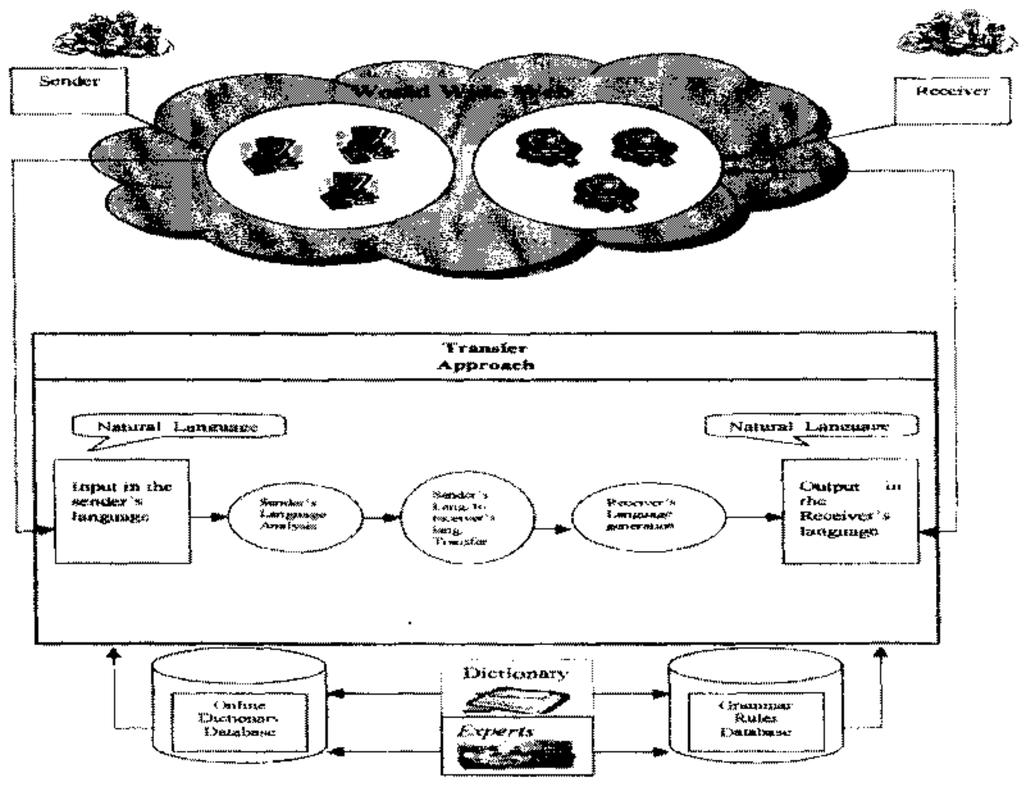
There are instances when miscommunication occurs. One instance is when one attempts to send a letter from one country to another that speaks in a different language. Because of the language barrier, the receiver might interpret the sender's message in the wrong manner. Thus, it might lead to confusion and at times arguments.

Another instance is when vendors cannot sell a product because their clients do not understand what they are selling. In order for the customer or client to invest in a certain product, the vendor must be able to fully communicate what the product is all about. In cases where language differences are an issue, the transaction may fail.

An alternative is to hire a human translator. However, translation is a highly skilled job that requires more than the mere knowledge of a number of languages. In some countries, the translators' salaries are comparable to highly trained professionals. Thus, it is costly to hire one. In addition, the delays in translation may prove costly. An average translator could only translate four to six pages of good quality translation per day and delays could erode the market lead-time of a new product [ARNO1994].

Also, when one conducts translations from one end of the globe to another, it is quite costly for it to be mediated by an interpreter over the phone. Also, there is an issue of confidentiality between the interpreter and the clients. The fact that the interpreter could acquire information over the course of translation is risky.

Theoretical Framework


Sender
World Wide Web
Receiver
Transfer Approach
Natural Language
Natural Language
Input in the sender's language
Sender's Language Analysis
Sender's Lang. to receiver's lang. Transfer
Receiver's Language generation
Output in the Receiver's language
Online Dictionary Database
Dictionary
Experts
Grammar Rules Database

. instant messenger translation system

The Philippine Style Chat is a proposed instant messenger translation system that works in a real- time environment. Given a scenario where two users do not understand each other due to different languages they used. The system provides a solution through the following steps. First, the users are subdivided into two parts: the sender and the receiver. The sender is connected to the Internet. Using the system, the sender sends a message in his/her own language. The input sentence goes into the system that uses the transfer approach to translation, and gets information from grammar rules and dictionaries. After going through the process of translation, the translated sentence is sent to the receiver in the language he/she understands.

The transfer process views translation as a three- phase process. First, analyze the input into a source- language syntactic structural representation. Second, transfer that representation into the corresponding target-language structure. Lastly, synthesize the output from that structure. Although this approach has the disadvantage of requiring another stage of processing, it holds an advantage of approaching the contrastive element of translation for it is at the transfer stage that the differences between the languages are revealed [SONN2000].

Grammar Formalism

Grammars are able to describe the syntax of languages. In the field of machine translation, grammar formalisms are treated as mathematical entities that can capture natural language [BORRA1999].

For this project, *context-free grammar* will be used. Context-free grammars are type 2 grammars that are widely used for syntactic description [TRUJ1999]. It is a formal system that depicts a language by describing how a legal text can be derived from symbols called an *axiom* or *sentence symbol* [ELI2001].

A context-free grammar consists of the following components:

> A set of *terminal symbols,* which are the characters of the alphabet that appear in the strings generated by the grammar.

> A set of *non-terminal symbols,* which are placeholders for patterns of terminal symbols that can be generated by the non-terminal symbols.

> A set of *productions,* which are rules for replacing (or rewriting) non-terminal symbols (on the left side of the production) in a string with other non-terminal or terminal symbols (on the right side of the production).

> A *start symbol,* which is a special non-terminal symbol that appears in the initial string generated by the grammar [NELS2001].”

To generate a string of terminal symbols from a Context-free grammar, the following steps are done:

- First, begin with a string that consists of the start symbol;

- Apply one of the productions with the start symbol on the left hand side, replacing the start symbol with the right hand of the production;

- Repeat the process of choosing non-terminal symbols in the string, and replacing them with the right hand side of some corresponding production, until all the non-terminal symbols are replaced by terminal symbols [NELS2001].

The close correspondence between the syntactic descriptions of the natural language and the context-free grammars have been made them a useful and popular tool in natural language processing. Also, since they are more convenient in defining syntax of programming language, efficient techniques for processing them have been developed. However, problems occur when situations like unbounded dependencies, which one cannot use context-free grammar to solve alone [TRUJ1999].

Project Objectives

General Objectives

The general objective is to develop a real-time instant messenger translation system.

Specific Objectives

The specific objectives are as follows:

> To collect relevant information from the sources and dictionaries that will help in translation.

> To use algorithms that will ensure good quality translation, which solves different types of ambiguity.

> To ensure that the translation will be done in an efficient manner, which takes into consideration the number of users using it.

> To incorporate features that will make translation easier and provide users with a "user-friendly" interface. *

> To constantly solicit users' feedback, so that the system will cater to their diverse needs.

Significance of the Study

Philippine Style Chat, being a translator system, dwells largely with the area of machine translation. In general, machine translation can be described as the process of translating from one human language to another. As a topic, machine translation is significant in various areas - socially, politically, commercially, scientifically, intellectually or philosophically [ARNO1994].

The *social* or *political* importance of machine translation is taken from the socio-political importance of translation in communities where more than one language is spoken. In this case, the only option other than hiring translator would be the adoption of a common "lingua franca". The problem with the creation of such language is that it involves the dominance of a certain language to the disadvantage of other languages becoming second-class or disappearing. Because of this, translation is important to effective communication- for ordinary human interaction and gathering of information one needs to play in society. One problem, however, is the shortage of translators due to the demand of translation. In the process, it seems that automation of translation is a social and political necessity for modern societies who do not wish to impose a common language on their members [ARNO1994].

The *commercial* importance of machine translation is a result of these factors. First, translation by itself, is commercially important. If a customer is made to choose between a product with an instruction manual written in English and one whose manual is written in Japanese, most English speakers will buy the former. Second, translation is expensive. Translation is a highly skilled job that requires more than knowing a number of languages. In some countries, translator's salary is comparable to highly trained professionals. Third, delays in translation are costly. Producing high quality translation of difficult material requires considerable time. Estimates would show that a professional translator might average around 4 to 6 pages of translation (200 words) per day. Due to translation delays of manuals of technical documents, it can easily lead to the delay of a product launch [ARNO1994].

Scientifically, machine translation is important, because it is a testing ground for various ideas and applications in Computer Science, Artificial Intelligence and Linguistics [ARNO1994].

Philosophically, machine translation is interesting, because it represents an attempt to automate an activity that requires the full range of human knowledge. One way of approach to machine translation is the extent to which one can automate translation is an indication of the extent to which one can automate "thinking" [ARNO1994].

Scope and Limitations of the Study

Scope of the Project

The developed system will create a real-time instant messenger translation device wherein users can communicate in real-time environment. Its scope is subdivided to two major parts: translation features and instant messenger related features.

Translation features

> Multilingual Aspect

- The system enables the user to choose and translate any of the following languages: English and Filipino. For future upgrades the system should be able to accommodate more languages.

> Approach to translation

- The system uses the transfer approach for machine translation. The transfer approach is a three-stage process; first the system analyzes the given sentence, translates the sentence then generates it to the target language. For a more detailed description of the transfer approach refer to the theoretical framework.

> Selection of Language

- The system allows the participants to determine the language they type in (source language) and view (target language).

> Mobile Messaging Shortcuts

- The system allows users to input different shorthand shortcuts: emotions, greetings and expressions. Ex. if the users types in "WRU"—"Where are you?" becomes the input for translation. This feature will be added to the instant messaging service as a way to lessen the number of keys the user has to press.

> Language Tools

- The system provides a selection of word processing tools like spelling and grammar checking as well as look-up dictionaries such as thesaurus. In addition, the tool simultaneously checks for spelling for error that may occur while typing. Also, the look-up dictionary will suggest words to replace the error. For these language tools, the services of Dr. De Castro of the Filipino department, a translation and grammar expert, have been enlisted as well as for the Filipino dictionaries.

Instant Messenger related features

> Status icons.

- The system provides icons to easily determine the language the user uses. As well as the language it is being translated to.

> User Profiles

- The system shows the contact information of each user in the participating chat rooms.

> Typing Notifications

- The system notifies that user when his chat friend typing.

> Emoticons

- The system provides a set of symbols that denote that user's expression or feeling.

> Organizations of Contacts

- The systems arrange your contact information into groups.

> Font Customization

- The system changes the style, color and size of your font.

> Environment communication tools

- The system allows the users to communicate information with the environment. It provides a section wherein users can click various options like Out for lunch, In a meeting etc- in a single click.

> Option for Interoperability with other messenger systems

- The systems will able to communicate with other messenger systems.

Limitations of the Project

> Number of Languages used.

- The proposed system will only translate two languages: English and Filipino only.

> Type of Language used

- The proposed system will only handle languages used in daily conservations and basic business languages examples of which are defined in selected books. These cover from basic greetings, to business related topics such as buying and selling..

> Translation of text only

- The system will only handle translation of written text. Voice translation will not be provided.

> Number of Sentence per send

- The user is only allowed to send one sentence at a time.

> Limit of Words in the dictionary database

- The dictionary database will be based on the words found in three books. The books are *Conversational Tagalog by Rufino, Practical Guide* to *Chinese-Filipino Conversation by Wangli and Co. and* Keeping Up with Your Chinese-Filipino by Young. It offers a vocabulary of approximately 4,000 words.

> List of grammar rules for each language

- The grammar rules provided will also be based on the rules provided in the two books. The books are *Lessons in Tagalog structures by Teresita Ramos and Analyzing the Grammar of English by Teschmer and Evans.* Each language will have around 110 grammatical rules each with their closest corresponding equivalents assigned by the research team.

> Mobile Messaging Shortcut limits

- This will be limited to about 100 MMS all of which will be common greetings and declarative statements that will be defined by the research group.

> Types of Sentences

- In order to have good translation, the types of sentences entered should be simple sentences. Compound sentences are discouraged. Also, the type of sentences entered covers declarative and interrogative sentences.

Project Justification

Justification for using machine translation

Background of Machine Translation

The art of Automatic translation or machine translation is the attempt to automate all, or parts of the process of translating from human language to another. At present, commercial presence in Machine Translation is getting a big lift even though the quality of machine translation output is not. Several factors have contributed to the implementation of this technology. Among them are translation quality, downsizing and integration and market factors [DALE2000]

In terms of translation quality, the technologies of machine translation traces back to the 1960's and 70's where each system requires millions of dollar funded largely by government agencies. At present, newer approaches using statistics and examples of existing translations (Example Based Machine Translation) has been introduced. However, it has not been proven to provide better results. That is why vendors currently developing approaches combining both linguistics and statistical methods, which may deliver higher quality translation in the future [DALE2000].

For the last decade, machine translation has dedicated their efforts to porting to applications from mainframes to UNIX microcomputers and then to Windows-based PC's then finally to web-based interfaces. This development makes it cheaper and easier get content out and into the machine translation system. In addition, machine translation is now a feature of popular applications, such as Lotus note and Microsoft Office. Access to various desktop applications increases translation volume, which in turn justifies the heavy investments to customize vocabularies and maintain the system [LION2001]

Another factor to be considered are market factors. During the last decade, product cycles have shortened not only in the hardware and software market, but also in the automobile and consumer electronics. This trend shrinks the commercial value of each product - related content. For example, shelf life of a printer is reduced to one year. If help and documentation delays last for eight weeks, the manufacturers will lose 15a% of its sales. Although Machine Translation may not offer a good translation, it may be more favorable than to incur loss due to translation delays. Another market factor is the volume of text, particularly in the area of customer and technical support, has grown into such extent that companies are forced to use faster translation. The third market factor is the tolerance for low-quality text. During the last decade, as the volume of information increased and the pace of exchange accelerated, user expectation about the quality of the text has decrease [LION2001].

Benefits of Machine Translation

Since Machine Translation holds great promise, there is a general misconception that machine translation was meant to replace human translation. Rather, machine translation only adds range of options to save time and cost and improve the services to customers and employees. Here are three possible goals that can be achieved with Machine Translation: saving cost, saving time and improve services.

Saving Cost

At present, there is no company that can reliably produced machine translation at the same quality as professional human translators. That is why cost savings cannot justify the use of Machine Translation when quality requirements are high. The reason behind it is increased fixed cost associated with maintenance and implementation. Cost savings can be achieved when quality is not the priority. This is possible when the Machine Translation contains a limited amount of customization. This is used in translating numerous documents or on real time translation systems such as the proposed project [LION2001].

Saving time

In any Machine Translation arrangement stated above, the time needed to produce results is far less than human translation. As a result, it is easy to meet goal of saving time with Machine translation. There are two instances where companies can save time with Machine translation. Needing it now and Time-to-market [LION2001].

Needing it now

If there are ephemeral materials that are value for a brief time, but may be worthless after an hour or a day after, an approach using Machine Translation without post-editing is enough. Examples of this include real-time communication, news reports, financial data, and so forth. Our project, Philippine Style Chat falls under this category [LION2001]

Time-to-Market

In cases where time-to-market is crucial due to shortened life cycles, the timesavings provided by Machine Translation can add "shelf life", thereby increasing the total revenue per product release. Lion Bridge has several projects that assist companies to accelerate time-to-market. These projects include technical supports articles, technical drawings, and aircraft manuals [LION2001]

Improving Services to Customers. Partners and Employees

Machine translation systems will not threaten the translator's job. Instead, it can help them by taking over some of the boring, repetitive translation jobs [ARNO1996]. Since the volume or change rate of content is too high to allow human translation, there is not enough budget or the resources aren't available. Also, it allows human translation to concentrate on more interesting task, where specialist skills are needed. Machine Translation with customization and maintenance delivers the message, even if it is not perfect [LION2001].

Justification on the Language Used

Regarding the Language used, the project is geared towards the desire of the Philippine government to use Filipino as the medium of communication in the academe and government organizations. The Philippine constitution affirms it by stating: "Ang pambansang wika ng Pilipinas ay Filipino. Habang ito ay nabubuo, patuloy itong pauunlarin at pagyamanin batay sa mga umiiral na wika sa Pilipinas at ibang wika.

Alinsunod sa mga probisyon ng batas at kung mamarapatin ng Kongreso, gagawa ng hakbang ang gobyerno upang simulan at itaguyod ang paggamit ng Filipino bilang wika ng opisyal na komunikasyon at bilang isang wika ng pagtuturo sa sistema ng edukasyon."

" Para sa komunikasyon at pagtuturo, ang mga opisyal na wika ng Pilipinas ay Filipino, at hanggang walang ibang itinakda ang batas, Ingles."

Kontitusyong ng Pilipinas
Mga Bahagi ng Art.XIV,sek 6, 7

Filipino reference books, manuals and documents are in need at present. Sadly, the Philippine system has been filled with English literature. With the arrival of powerful formalisms, knowledge representations and computational architectures to address natural language processing, machine translation can provide a better way of converting current English reference to Filipino. Also, it may serve as learning tool for foreigners to speak and correspond in Filipino [BORR1999].

Another reason that made group decide to use English and Filipino as its base language is due to the presence of competent linguist here in the Philippines. Also, the groups found out that other existing translation system do not provide good translation for Filipino. However future projects may try to extend the language translated to other foreign languages as well [SHAR1992].

Justification on the Working Prototype Status

The research, Philippine Style Chat, is a working prototype. There are instances where the process of developing a prototype can be a project itself. This is the case present with our research. The reason behind it is that a complete Machine Translation system is quite complex to develop in a matter of one-year. Complexity is seen in the building of dictionary, grammar rules, the presence of different ambiguities and handling of different translation cases.

Statement of assumptions

In order to have a more efficient and better translation, the following assumptions should be met:

- The sentence construction should follow the subject-verb-object construct. Avoid run-on sentences and jargon. Sentences should be simple and direct.
- It is advisable to keep sentence length under 128 characters per send.
- The word count is encouraged to be within 25 words per sentence.

- Avoid abbreviations. Unless the abbreviations were included in the shortcut list, the abbreviations will not be translated.
- Special characters are skipped in the translation process
- If the translator identifies more then one meaning for the word and cannot deduce from the context which meaning is more correct, the translation may not make sense.
- If there are words that are not in the dictionary database, the sentence will undergo word for word translation. You may add any words in the dictionary for subsequent translation.
- The user of the system is also assumed to know the basics on hoe to operate a computer and the Windows operating system.

Operational Definitions

> **Chatting:** light, easy, informal talk or conversation.

> **Context Free Grammar:** are type 2 grammars that are widely used in syntactic description. It is a formal system that depicts a language by describing how legal text can be derived from symbols called axioms or sentence symbols.

> **Cross-platform:** ability of the software to run in different platform

> **Globalize:** to make global; especially to organize or establish worldwide.

> **Grammar Formalism:** mathematical entities that can capture natural languages.

> **Machine Translation:** attempts to automate all, or parts of the translation between two languages.

> **Mobile Messaging:** list of words or shortcuts used for text messaging and chatting.

> **Multilingual:** using or capable of using several languages.

> **Natural Language:** any type of languages naturally used by human, not an artificial or man-made language such as a programming language.

> **Natural Language Processing:** is a convenient description of all attempts to use computers to process natural language.

> **Noun Phrase:** can appear as subjects, objects, and in other contexts; sometimes proper names and pronouns are treated as phrases, given their distribution.

> **Parse:** to separate (a sentence) into its parts, explaining the grammatical form and function of each of the parts and their interrelation

> **Parsing:** the process of analyzing sentence to determine its syntactic structure according to a formal grammar. Not a goal in itself, but an intermediary step for the purpose of further processing, such as assignment of a meaning to the sentence.

> **Pragmatics:** branch of linguistics that deals with language in action.

> **Prepositional Phrases:** occurs in a variety of contexts and within almost all other phrase types.

> **Prototype:** all or parts of a system that looks like the system under consideration but does not have the complete functionally of a real system; but it does not perform all the functions of a complete system.

> **Production:** rules for replacing non-terminal symbols in a symbols in a string with other non-terminal or terminal symbols.

- **Semantics:** study of meaning in language.
- **Start symbol:** special non-terminal symbol that appears in the initial string generated by the grammar.
- **Syntax:** description of the word arrangement and of the relationships of meanings that these arrangements express.
- **Terminal symbols:** the characters of the alphabet that appear in the strings generated by the grammar.
- **Transfer Approach:** views translation as a three-phase process. First, analyze the input into a source language syntactic representation into the corresponding target language structure. Lastly, synthesize the output from that structure.
- **Verb Phrase:** includes predicates as well as phrases appearing in other context.

Methodology

The system that will be developed is patterned after Rapid Application Development. Rapid Application Development is a relatively new approach to systems development that emerged in the 1990's. Founded by James Martin, the methodology focuses on developing information systems that assure better and cheaper systems and more rapid development [HOFF1995]. Thereby, it addresses both weaknesses of the structured development methodologies: long development time and difficulty in understanding a system from paper-based description [DENN2000]

There are four necessary pillars for Rapid Application Development approach: **People, Tools, Methodology** and **Management. People** or **users** are much involved in the prototyping processes where end users and analyst work together to design and iteratively redesign interfaces, displays and reports for new systems the prototyping is conducted in a session that resembles Joint Application Development sessions. In the sessions, users perform detailed reviews of the system prototypes and specifications. Tools are also an important part of the process. CASE tools, which include code generators for creating bug-free code from designs of interfaces, help speed-up prototyping. A coherent methodology that spells out the proper tasks to be done in the proper manner and support from management are also important pillars for Rapid Application Development [HOFF1995]

The Rapid Application cycle contains the basic phases of any life cycle: planning, analysis, design and implementation, although he calls the phases Requirements Planning, User Design, Construction and cutover. It is treated as an adaptation of SDLC, but emphasizes on continuous iterative development rather than a strict step-by-step process. A way to view Rapid Application Development is that several SDLC phases occur simultaneously [HOFF1995].

Program Structure

Transfer Process

The transfer based translation process composed of three major steps. The first being the Analysis, followed by the transfer phase and finished off by the generation phase. The Analysis is centered on identifying the words to be translated, by deconstructing the sentence and by deconstructing the sentence and by identifying the words and their relationships to each other. There are three types of analysis: Lexical (which includes tokenisation and Morphological analysis), syntactic and semantic analysis.

After the analysis comes the transfer phase is step wherein the system transfers the sentence, and given the data from the analysis of the words from the analysis phase, analyzes it syntactically through the parser operations. That is breaking down the sentence more by exposing the word's grammatical associations to each other. To accomplish this, the system checks with its data dictionaries for the grammatical data it needs to identify the sentence patterns. And given the evaluation of the data the system can produce a grammatical equivalent of the sentence.

The generation phase of the transfer approach transverse the parse tree to generate the output sentence. Afterwards, it undergoes morphological generation by adding the necessary prefixes, affixes and suffixes to complete the sentence.

Analysis Phase

Lexical Analysis

There are three levels in the process of machine translation: the lexical level, the syntactic level, and the semantic level. The lexical level is the lowest of the three levels, dealing mostly with given alphabet of the source and target languages and a morphological analysis of the words. Low in the hierarchy it may be, it is nonetheless the most important, without the lexical level, translation to the next two levels would be next to impossible, as the lexical level is responsible for recognizing words that are not normally part of the dictionary DALE2000].

The first part of the lexical analysis is to ensure that the word to be analyzed is within the set parameters of the source language. This includes the identification of the source's alphabet source dictionary. In our project however, that aspect is taken care of by an ability of the messenger to select the source and target languages. After analyzing and determining the source language parameters, the next step is to analyze the word in the morphological level, which is the core of the lexical analysis DALE2000]

Tokenisation

Tokenisation is the process of breaking up the sequence of characters in a text by locating the word boundaries, the points where one word ends and another word begin. The words that are identified are frequently referred to as tokens. For example, the token "sentence" is under the token noun, which is different from the punctuation mark tokens [BORRA1999]. Also known as word segmentation, tokenization, tokenisation is well established and used for artificial languages such as programming languages. Further on, tokenisation is widely used in two types of languages: space-delimited and unsegmented languages [DALE2000].

In both space-delimited and unsegmented languages, specific challenges of tokenisation lie in both the writing system and the typographical structure of the words. There are three main categories into which word structures are classified. The morphology of a word can be classified as *isolating,* in which words do not divide into smaller units, *agglutinating,* in which words divide into smaller units (morphemes) with clear boundaries between the morphemes are not clear and the component morphemes can express more than one grammatical meaning [DALE2000]

Tokenisation in Space-delimited Language

Since the system Philippine Style Chat will be translating two languages, English and Filipino, which are classified as space-delimited language, let's take a look at issues affecting tokenisation in space-delimited language. In space-delimited language, most tokenisation ambiguities exist among uses of punctuation marks, such as periods, commas, quotations, apostrophes and hyphens, since the same punctuation mark can serve many functions in a single sentence, let alone a single text [DALE2000].

Example, the word "grammar" as noun becomes an adjective when added with an affix like "grammatical". Compounding processing exist when two new independent words are used to form a new word [AGNO1994]

Morphological analysis in machine translation considers the inflection and derivational processes and possibly compounding processes to analyze structurally each word and reduce them to their morphemes. Thus, the size of the dictionary will be greatly reduced since inflections and derivations of the word will not be included as part of the lexicon. Also, the tense, number, part of speech and role of a word in a sentence is determined through morphological analysis [BORRA1999].

Simply speaking, morphological analysis recognizing the words by identifying the prefix, infix, affixes and suffixes within them.

Examples:	Minamahal	Root: mahal Prefix: mina
	Iisipin	Root: isip Prefix: i- Suffix: -in

All the data on the pre-, in-, affix-, and suffixes, the rules surrounding their usage and corresponding target language equivalents and their grammatical roles in the sentence will come from a dictionary database that is derived form selected sources. Once the root word has been taken, the translation process goes to the next level of translation, the syntactic, with the data derived from the comparison search, the grammatical roles of each word and the rules from the tokenisation process.

There are several ways to implement the rules to input to produce the output tree. Also, there are different procedures or parsing algorithms by which an input string can be assigned a structure. In this text, we'll be discussing two types of parsing algorithms: bottom-up parsing and top-down parsing. [AGNO1994]

Bottom-up parsing is a method that starts out with the words in the sentence and built the tree "bottom-up". The first step is for each word in the sentence, look for a rule whose right hand side matches it. This states that every word should be labeled with a part with a part of speech (shown in the left hand side of the rule that matched it). The step is similar to looking up words in the dictionary. Then, starting from the left hand side of the sentence, find each rule whose right hand side matches will match one part of the speech. Keep doing this step, matching larger and larger bits of phrase structure until no more rules can be applied [AGNO1994].

On the other hand, a top down parser starts with S and attempts to guess what it is looking for in each sentence. It tries to guess if the sentence starts with the noun phrase and guess if the noun phrase consists of a noun and checks if the noun is located at the start of the sentence. Basically, it has a choice of possibilities. If the first choice is incorrect, it backs up and tries the next alternative. In the case of parsing a complicated grammar, it would eventually get the correct answer, perhaps only after many wrong guesses [AGNO1994].

Semantic Analysis

Semantics is concerned about the meaning the words have and how they combine to form the sentence meanings. Before anything else, it is important to distinguish *lexical semantics* and *structural semantics.* Lexical semantics deals with the meanings of words, while structural semantics has to do with the meaning of phrases, including sentences. In our project, we will be dealing with both types of semantics [AGNO1994]

Lexical Semantics

There are various ways to represent word meanings. However, for it to be effective, it has to be useful in the field of machine translation, in machine translation, the words are associated with the semantic features, which corresponds to their sense components [AGNO1994].

Examples:

Man =(+Human,+Masculine and +Adult)
Woman =(+Human,-Masculine and+Adult)
Boy =(+Human,+masculine and -Adult)
Woman=(+Human-masculine and-Adult)

Associating words with semantic features is useful because there is words that impose semantic constraints on what other kinds of words can occur with it [AGNO1994].

Example:

The verb *eat*-its agent (the one who eats) is animate-its patient (that which is eaten) is edible, concrete (not abstract like sincerity) and solid (not liquid like beer or coffee)

Encoding the constraint in our grammar rules by associating the features HUMAN and EDIBLE with appropriate nouns in the dictionary and describing that our entry for *cat* as something as cat=verb, AGENT=HUMAN, PATIENT=EDIBLE. The grammars will only accept objects *of cut* that have the feature EDIBLE. This type is coined as selectional restrictions and its filters out unwanted analysis (AGNO1994).

The first type of restrictions we'll be using in our project will be selectional in nature. selectional in nature, selectional restrictions is very useful device and found in most MT systems in a greater or lesser extent. Selectional restrictions in terms of semantic relations can help one of the thorniest problems for machine translation, namely translation of prepositions [AGNO1994].

Example:

The translation of the English preposition *on* can have two translations in Filipino, as in the following:

(1) English: at noontime
Filipino: sa tanghalian
(2) English: at home
Filipino: nasa bahay.

The choice of Filipino preposition depends on the type of the noun that follows it. Roughly, where a temporal noun follows the preposition, as in (1), it translates as sa. If however, the preposition is followed by a locative noun, it translates as nasa [AGNO1994]

Examples:

at, SR=TIME-sa
at, SR=PLACE-nasa

The semantic relations are assigned on the basis of the noun that follows the prepositions. This denotes that noun *noontime* must be marked in the dictionary with some temporal feature (ex semtype=TIME), while nouns like table are labeled with some locational feature (ex. Semtype=location) [AGNOr994].

Structural Semantics

Semantics is also concerned with linguistics systems such as tense and aspects and determination, all of which are important in translation. Consider the problem of how to translate the present tense in French to English where there are possibilities, stated in the following [AGNO1994].

(1) French: Elle vit a Londres
English (l)-She lives in London
English (2)-She has lived in London

Of course, one would try to formulate rules that would describe the conditions under which French present tense would be translated to English present. An approach would be dissect the English tense system. The English Tense system conveys two types of information. One is the time of event-both the present simple I SING and the present progressive I AM SINGING describes an event in the present. The other one is the nature of the event- e.g. the progressive stresses that the event is “in progress”. Therefore, let’s use the word TENSE to mean the time of the event and ASPECT to refer to the way it’s viewed. To refer to the way it’s viewed. Let’s use TIME REFERENCE to cover both and aspect [ARNO1994].

We can think of a tense as expressing a relation between the time of the event and the time of speech. Thus, with the present (I SING), the time of the event (which we call E) overlaps with the time of speech (which we call S) Contrast it with the future (I SHALL SING) where the time of the event follows the time of speech. (E follows S) or the past where E precedes S. But problems occur when the simple past (I SANG) and the past-perfect (I HAD SUNG). Both cases show that the time of event is prior to the time. A solution to it is an additional point of time, which is called is called REFERENCE TIME(R) [AGN01994].

The following shows the notation

Sam has eaten	past-perfect	R	precedes	S,	E Precedes R
Sam ate	simple past	R	precedes	S,	E Coincides with R
Sam has eaten	present perfect	R	coincides with	S,	E

By analyzing the sentence, we now have the technique to represent the difference in tense and aspects of the examples above. In this way, if the source language is in present tense form, then we have a way of analyzing the sentence to ensure that the target language also falls under present tense form [AGNO1994].

Transfer

The transfer phase takes a representation of the source language as input and outputs the corresponding representation or structure of the target language. There are generally two types of transfer rules used and lexical transfer-based system. These are the structural rules and lexical transfer rules.

Lexical transfer transverse the structure, locates the first feature and the translation rules for the feature. If no translation rule exists, it is simply carried over to the target structure [BORR 1999].

Structural transfer rules involve source to target transfer of rules. The target structure is determined through the mapping database that assigns a Source Language based structure [BORR1999].

Generator

The generator phase function as the reverse process of taking an input string a producing representation for it. Therefore, it tries to generate a string from a constituent structure representation. In other words, one needs to do something with the words in order to get the correct form. The output would then be a list of tokens or words that form a sentence for the target language [BORR 1999]

Examples of the Translation Process

Example no. 1

Type of Translations: English to Filipino
Input Sentence: I love you
Output Sentences: Mahal kita

<u>Translation Process</u>

Analysis phase

First step: Tokenisation

The system breaks down the sentences “I love you”. To three separate words Love, You, and I through analyzing the white spaces in between the words in the sentences. It also recognizes that the tokenisation ends there through the period (.) located at the end of the sentence.

Diagram:

I love you. → word 1:I
word 2: LOVE
word 3: YOU

Step : Lexical Analysis
Morphological Analysis

The system cross checks each word with the data dictionary of the source language for possible pre-/in-/-affix/suffixes in the event that the dictionary does not recognize the word. Since the words I, Love and You don't have the pre-/in/affix/suffix, then morphological analysis is completed.

The dictionary entries also contain additional data about the words that are checked, for each word it is assigned a list of all possible equivalent meanings in the target language by cross checking the two databases as well as the grammatical data of the word, that being a noun, pronoun, subject, object or whatever possible case may be for that particular word. This data is then transferred for the transfer part of the translation process.

Sample data taken from dictionary database

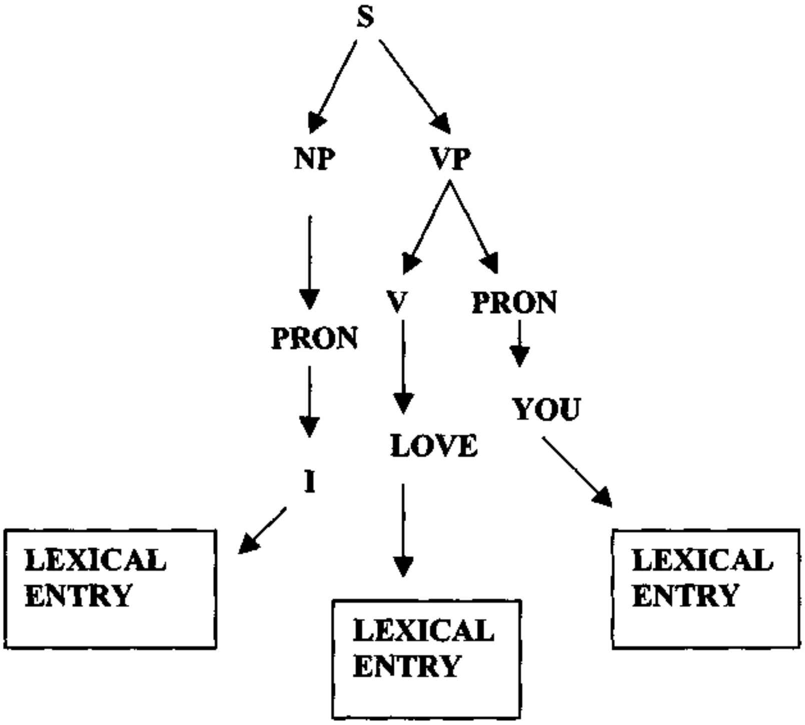
Word	Lexical Entries
I	Part of Speech: pronoun Root: I Number: Plural
Love	Part of Speech: Noun/verb Root: love Number: plural
You	Part of Speech: Pronoun Root: you Number: Plural

Third step: Parsing

After the system takes the data from the dictionary, it will generate the tree based on the grammar structure.

The parsing technique is a method of analyzing a sentence to determine its structure according to the grammar [BORRA1999]. In order to fully understand parsing, it has to be emphasized that the goal of an automatic parser is to take a formal grammar and apply the grammar to the sentence in order to: a) Check that it is indeed grammatical and b) Given that is grammatical, show how the words are combined into phases [AGNO1994].

Tree generated



Conclusion

Globalization has stamped its influence in the society today. In order to stay competitive with the times, development should be made in order to bring us closer to the global community. Communication and the Internet play a big role in fulfilling that need. However, the fact that people from different nationalities communicate using different languages poses a big problem. An online-real time translator can be a useful tool to help solve the dilemma.

The instant messenger translation system the group will be developing will translate the Tagalog-based Filipino language to the English language and vice versa. The mode of communicating will be using are natural languages and the message sent will only accommodate up to two sentences only.

Several concepts are directly related to the development of an online, real -time chat translator. This study tackles the general concept of communication to specific concepts like Natural Language Processing and Machine Translation. It also provided an overview of the different online real-time machine translation systems available in the market today.

Bibliography

I. Internet Resources

[AIKA2001] Aikawa, T ., Melero, L. Schwartz and A. Wu. Multitlingual Sentence Generation. In *Proceedings of the 8th European Workshop on Natural Language Generation, Toulouse.* http.//research.Microsoft.com/nlippubs.asp.2001.

[AMIC2001] Amichat.www.amikai.com/EN/Products/asp/amichat.jsp.2001.

[AMTR1999] Amtrup,Jan. Architecture of the Shiraz Machine Translation System. Computing Research Library. 1999.

[BUSH 1996] Bush, Michael. Paper presented at the 1996 Symposium of the Computer Aided Language Instruction Consortium. (CALICO) Albuquerque,NM. May 29,1996. *http://industry.java.sun/javenews/ stories/story2/0.1072,391.36.00.html.2001*

[COMP2001] Computer User : High Tech Dictionary. htpp://www.computeruser.com/resources/dictionary/popup definition.php.?lookup=53.200.[ELI200n
ELI Translation Construction Made Easy. www.cs.colorado.edu/~eliuser/elionlined43/syntax 1.html.2oo1

[GVU1998] GVU WWW User Surveys. http.//www.gvu.gatech.edu/user surveys/survey-1998-10/graphs/use/q01/htm. 1998

[HARD2001] Hardy, Michael, Universal Translations Abound. www.multicity.com/about/press/inthenews/ptj062501.html.June 25, 2001

[HOW 1999] How Compilers Are Constructed. http.//ece-www.Colorado.edu/~ecen4553/Content/Overview.html.1999.

[PARA2002] Paralink.IM Translator. http://www.paralink.com/ims/index.html.

[INFO2001] Infoworld. IBM introduces Instant Online Translation. www.itworld.com/Tech/2420/IW 10108.nibontrans/2001

[LANG2002] LanguageForce. Universal Translator 2000. Languageteacher.com 3930 Swenson Street #310, Las Vegas. NV 89199 http.//www.languageteacher.com/textsoft/ut2000.html.20Q2.

[LION2001] Lionbrigde Technologies Inc. when to use Machine Translation. A Lionbridge White Paper.2001.

[MCCO2001] McCornell,Brian. Building a Multilingual Chat Client.http.//picotwebloggr.com/stories/storyReader$56.html.2001

[MYSQ2002] MySQL. The most Popular Open Source Database. http://www.mysql.com/companv/index.html.

II. Book and Articles

[ALE1972] Alejandro, Rufino. Conversational Tagalog. National Bookstore Inc. Pines cor Union Sts, Mandaluyong City. 1972.

[ALMA 1996] Amario et al. patnubay sa pagsalin. Pambansang Komisyon sa Kultura at mga Sining, Metro Manila. 1996.

[ARNO1994] Arnold, Balkan, Meijer et al. Machine Translation: An Introductory Guide. Blackwells/NCC. U.S.A. 1994.

[BERL1960] Berlo, David. The Process of Communication. Holt, Reinhart and Winston,Inc. U.SA.

[BORR1998] Borra, Allan. A transfer Based Analysis for a English to Filipino Machine Translation Software. Master Science in Computer Science Thesis. December 1998.

[DALE2000] Dale, Robert. Symbolic Approaches to the Natural Language Processing. Macquarrie University. Sydney, Australia.2000.

[HOFF1995] Hoffer, George and Valecichi. Modern Systems Analysis and Design. The Benjamin Cummings Publishing Company. Inc. 1995.

[LANG 1999] LanguageForce. Professional Universal Translator 2000. Tutorial. Languageforce, Inc. 1999

[NOBI2001 Nobilitt James, The Online Language Learning: So What's New>FIPSE Summit Conference. April 28,2001.

[PRIN1996] Princeton Review, Grammar Smart. A guide to perfect Usage. Princeton Review Publishing. U.S.A. 1996.

[PREN1998] Prentice Hall. Webster New World Dictionary. 3rd College Edition. Prentice Hall Publishing. 1998.

[RAMO1971 Ramos, Teresita. Tagalog Structures. Honolulu: University Of Hawaii Press. 1971.

[SAMS 1999] SAMS> Teach Yourself Java in 21 days. Sams Publishing 1999.

[SCHO2002] Schonberger, Victor and Hurley Deborah. "Globalization of Communications" Governance in a Globalizing World. January 2000.

[SHARI1992] Shari, Lawrence and Pflegger. Software Engineering. The Production of Quality software. St. Martins Publications. 1992.

[SOMM2000] Somers, Harold. Machine Translation. UMIST. Manchester England.2000

[TESC1995] Teschmer and Evans. Analyzing the Grammar of English. 1995.

[TRUJ1999] Trujillo, Arturo. Translation Engines: Techniques for Machine Translation. Springer-Verlag, London Limited 1999.

[WANG1998] Wang, Weidong And Xiumei. Practical Guide to Chinese-Filipino-English Conversation. Inkwell Publishing co., Inc. 1998.

[YOUN1991] Young, Johnny. Keeping Up With Your Chinese-Filipino (Business Edition). Oregem Int'l Publishing Co. Inc. Diliman, Quezon City. 1991.