NewsBytes: Tagalog Text Summarization Using Abstraction

Ervin G. Batang
College of Computer Management and Information Technology
Polytechnic University of the Philippines
Anonas St. Sta. Mesa, Manila, Philippines
teddy_vin11@yahoo.com

Don Erick J. Bonus
Computer Science / Information Technology Department
Technological Studies
Jose Rizal University
Mandaluyong City, Philippines
dej_bonus@yahoo.com

Mark Angelo T. Miano
College of Computer Management and Information Technology
Polytechnic University of the Philippines
Anonas St. Sta. Mesa, Manila, Philippines
MarkAngeloMiano@gmail.com

Ma. Regina L. Cruz
College of Computer Management and Information Technology
Polytechnic University of the Philippines
Anonas St. Sta. Mesa, Manila, Philippines
Reginacruz1986@yahoo.com

Ria A. Sagum
College of Computer Management and Information Technology
Polytechnic University of the Philippines
Anonas St. Sta. Mesa, Manila, Philippines
riasagum31@yahoo.com

Rubeleen Ann C. Yu
College of Computer Management and Information Technology
Polytechnic University of the Philippines
Anonas St. Sta. Mesa, Manila, Philippines
ruby_ann42806@yahoo.com

ABSTRACT
In this paper, we present an automatic Tagalog text summarizer that uses abstraction instead of the traditional extraction method of summarization. It employs Natural Language Processing and Generation to produce the summary. Summarization works by determining the subject of the sentence and then building phrases for that subject. A prototype was tested and evaluated based on the following matrices: sentence simplicity, and cohesiveness and understandability of the summary.

Categories and Subject Descriptors
H.3.3 [**Information Storage and Retrieval**]: Content Analysis

I.2.7 [**Computing Methodologies**]: Natural Language Processing

General Terms
Algorithms, Languages.

Keywords
Abstraction, Summarizer

1. INTRODUCTION
Extracting essential information from a source or a document to generate a shortened version for a specific user or users is the process of Summarization. According to Hann [3], summarization has become an integral part of everyday life. People keep abreast of world affairs by listening to news bites, base investment decisions on stock market updates and even go to movies largely on the basis of reviews they've seen.

There are basically two methods of summarization: extraction and abstraction. Summarization by extraction works by selecting original pieces from the source document and concatenating them to yield a shorter text. However, the result of this method is usually incoherent [8], making the text hard to comprehend. Abstraction, on the other hand, makes the summary more understandable for readers who wish to read the most important part of a document. It eliminated incoherent and sometimes 'choppy' summaries.

There are some summarizing tools available already [9], however, most of these tools use the extraction method of summarization.

Unlike the linear model in extraction methods, abstraction requires heavy machinery from natural language processing, including grammars and lexicons for parsing and generation. It works by analyzing the content of the source document and breaks it down into separate parts. It also searches for the main topic in the document and analyzes other sentences that might describe the topic. The main topic, along with its supporting sentences is further processed to generate the final summary. The purpose of this study was to develop an efficient algorithm for summarizing Tagalog news articles by using the abstraction method.

2. BACKGROUND OF THE STUDY
According to Pachantouris [7], an automatic text summarizer is a computer program that summarizes a text. Extracting essential information from a source document to generate a shortened version is referred to as the process of summarization. In his paper entitled "GreekSum: A Greek Text Summarizer", he presented an algorithm that consists a language independent summarization engine and several dictionaries. The NCSR "DEMOKRITOS" provided the Greek keyword dictionary. The algorithm was then compared with the Generic mode of SweSum based on which summary is better, which kept the most important information, coherence, and a 1 to 5 scale system. Although the results show GreekSum gave better summaries, it was also

mentioned that in most cases the contents of the summaries are identical.

Müürisep, et al. [5] stated in their study, 'ESTSUM – Estonian Newspaper Texts Summarizer', that as the amount of on-line information increases, more and more effort is dedicated to creating automatic summarization systems. Since the automatic text summarization is largely a language-specific task, suitable algorithms must be found for each natural language.

The paper entitled "Use of Topic Segmentation for Automatic Summarization" [1] described a summarization system that uses generic text structure cues to detect structures in texts. It made use of a process called "layered topic segmentation" wherein key terms were associated with each topic or subtopic and outputs a tree-like table of content (TOC). The text structure trees reflect the most important terms at general and more specific levels of topicality and indicate topically coherent segments from which sentences were mined for inclusion into summaries.

The Online News Summarizer by Cruz et al. [2] is a direct application of an automatic text summarization system. It is an email service, which seeks to provide its users with daily, summarizes from various local and international news entities. Its primary component is an auto-text summarization, which uses support vector machines (SVMs) in determining which sentences are to be included in the summary, instead of using traditional approach to automatic text summarization such as lexical chains, machine learning based summarization, discourse trees, and sentence extraction techniques. The process starts by extracting news articles from the Internet. Since the news articles are in .html format and the summarizer engine takes text files as input, an html stripper was used to strip unwanted html tags in order to produce the text to be summarized. The summarizer engine will then determine the sentences to be included in the summary that is then emailed to the user.

According to Herzog [4], natural language generation is often characterized as a process that has to start from the communicative goals of the writer or speaker and needs to employ some sort of planning to progressively convert them into written or spoken words. With this in mind, the goal of the language producer is forced into linguistic nature, i.e. trying to produce particular words. The generation of the summary uses two techniques to generate a summary; strategically (deciding what to say) and tactically (deciding how to say it). Determining the large-scale structure of the text to be generated should also include content selection. Usually, this process involves a tree-like formation in which the leaves contain instructions that are then passed to a sentence generator, a task that can be further categorized into sentence planning.

3. NEWSBYTES

NewsBytes uses an abstraction algorithm that has three sub-processes; namely Parsing, Tagging and Scoring, and Generation. The document is parsed into words, tagged with their part of speech, and scored based on their importance. The study used the Phrase-Merging sentence generation method to produce the final summary.

3.1 The Parsing Process

Before NewsBytes can begin summarizing, it has to break-down the document into several parts – into paragraphs, then sentences, and finally, into words. It needs to be broken down before the tagging process could begin.

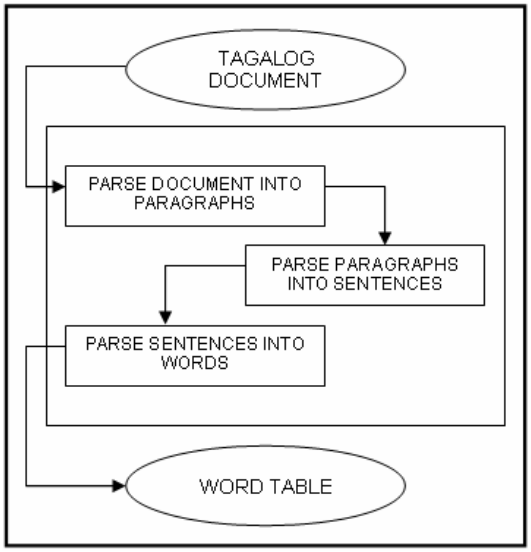

Figure 1. The Parsing Process

Figure 1 shows the parsing process. A Tagalog news article is broken down into paragraphs, and then into sentences, and then into words. All sentences and words are stored in a table, including the paragraph number for sentences, and the sentence number for words.

3.2 The Tagging and Scoring Process

The tagging process uses the information the parsing process produced. This time, NewsBytes uses its database for information about each word, as this process is as critical as generating a meaningful summary. The tagging process also includes the scoring process in which individual words are scored according to its importance. This determines which sentences should be included in the final summary.

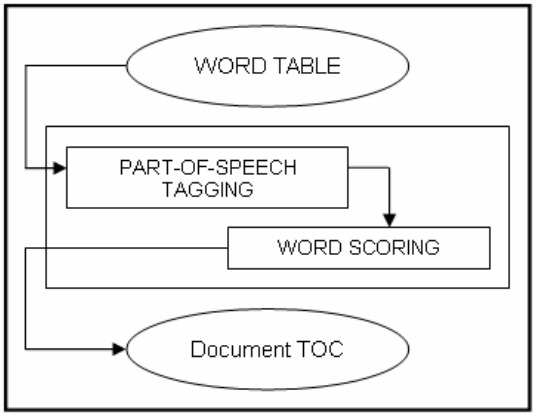

Figure 2. The Tagging and Scoring Process

Figure 2 shows how the tagging and scoring process work. It starts with each word being tagged with its corresponding part-of-speech taken from a database. The tags are used to determine the structure of the sentence. The sentence structure determines the subject of the sentence, and therefore, the topic. It then associates the remaining parts to the topic. Word scoring also occurs in this process. The word score is higher for important words, such as nouns, pronouns, certain verbs and adjectives, numbers, dates, times, places, and events. Lower word score is given for passive verbs, adverbs, common words, and articles. The scores of all words in a sentence are summed to produce the overall sentence score. The tagging process produces the Document Table of Contents (TOC) to be used by the generation process. The Document TOC contains the subjects or topics found in the document, along with its associated predicates, and the overall sentence score.

3.3 The Generation Process

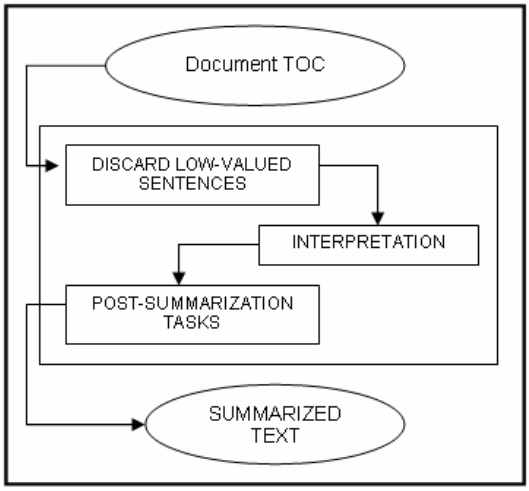

Figure 3. The Generation Process

Looking at Figure 3 above, the data from the Document TOC was used to generate the summary of the source document. In this process, the overall sentence scores will be sorted from highest to lowest, and discards the lower-half of the gamut. Sentences with high scores are considered important and are included in the final summary. The condensing of the text happens at the Interpretation stage, in which certain word groups (e.g. *mga lalaki at mga babae*) are condensed to general terms (e.g. *mga tao*). These interpretation definitions and instructions are stored in a database. When the interpretation is complete, the post-summarization tasks are carried out. These tasks furnishes the final text, and includes capitalizing the first letter of the sentence, indenting the paragraph, and inserting periods at the end of sentences – all to make the final, generated summary more readable.

4. EVALUATION

The abstraction summarizer was evaluated on three matrices: the simplicity of words used in the summary, its sentence cohesiveness, and the understandability of the summary.

In evaluating NewsBytes, the proponents compared its proptotype's generated summary to Microsoft Word's AutoSummarize [6] with a default 25% of the original. MS Word's AutoSummarize uses a language independent extraction method in generating summaries.

Table 1. Opinion Index used for the Evaluation

Range of Mean Values	Verbal Interpretation	
4.51 – 5.00	Highly satisfactory	HS
3.51 – 4.50	Very satisfactory	VS
2.51 – 3.50	Moderately satisfactory	MS
1.51 – 2.50	Fairly satisfactory	FS
1.00 – 1.50	Unsatisfactory	US

The proponents let fifteen randomly chosen CCMIT BSCS students and teachers use the NewsBytes and the MS Word's AutoSummarize, and compare the summaries produced by the application based on the matrices mentioned above and with the opinion index presented in Table 1 above. The respondents then fed two different news articles into both summarizers to measure the systems summaries' simplicity, cohesiveness and understandability.

Table 2. Results of the Comparison between Microsoft Word's AutoSummarize and NewsBytes

Criteria	Microsoft Office Word's AutoSummarize			Sonicsoft NewsBytes' Abstraction			t-ratio	VI
	WM	VI	SD	WM	VI	SD		
Simplicity	3.11	MS	0.881	3.15	MS	0.681	0.238	NS
Cohesiveness	3.20	MS	0.887	3.43	MS	0.774	1.068	NS
Understandability	2.98	MS	0.892	3.37	MS	0.909	3.578	S

To arrived at the results presented in Table 2, a level of significance of 0.05, degree of freedom of 13, tabular value of 2.16 were used. Based on the results of the comparison, it could be gleaned from Table 7 that the respondents share their perception in simplicity and cohesiveness of the two systems, while they have different perception on the understandability. They believe that they have better comprehension using the NewsBytes' Abstraction approach over the Microsoft Word's AutoSummarize. This could be confirmed by the obtained t-values of 0.238 and 1.068, which fell below the tabulated value of 2.160, interpreted as not significant (NS), and a t-value of 3.578 for undestandability, which exceeded the tabulated value leading to the rejection of the null hypothesis which states that there is o significant difference between the extraction and abstraction methods of summarization.

5. CONCLUSIONS AND FUTURE WORKS

Most of the current automated text summarization systems use extraction method to produce summaries because it is easy to implement. Although using extraction is uncomplicated, there are three major difficulties:

- Finding out which are the most important sentences to use on the summary
- How to generate a coherent summary
- Remove all redundancies in the summary

In order to solve the difficulties in extraction method, the researchers preferred to use abstraction method.

NewsBytes' Abstraction for Tagalog news summarization does not employ statistical methods in producing a summary. Rather, it uses a database to determine which words are important, and to determine the sentence structure. It also contains rules that govern the correct interpretation of sentences. The precision of the summary relies on the database of NewsBytes.

NewsBytes' Abstraction is dictionary-based, therefore, the database should contain an extensive list of Tagalog words and interpretations. Thus, other methods may be explored towards the development of a non-dictionary-based summarizer.

The following are some recommendations to improve the algorithm used in the NewsBytes' Automatic Tagalog News Summarizer:

- Because Tagging is important to Subject Determination, improved sentence structure tagging can be used
- Machine Learning methods, where the algorithm learns new sentence structures from the source document, to further improve the summarization
- The scope of this study is only for news articles. The scope could be widened, not only for news articles, and even to other Philippine languages

6. REFERENCES

[1] Angheluta, R. et al. *The Use of Topic Segmentation for Automatic Summarization.* Katholieke Universiteit Leuven, 2002.

[2] Cruz, et al. *Online News Summarizer.* Thesis, University of the Philippines. Available: http://www.engg .upd.edu.ph/ ~naval/cs198/2002/ons/index.html. 2002.

[3] Hann, U. et al. *The Challenges of Automatic Summarization.* 2000. pp. 29-35.

[4] Hezrog, O. et al. *Language Generation.* [online]. Available: http:// www.lt-world.org/HLT_Survey/ltw-chapter4-all.pdf

[5] Müürisep, K. *ESTSUM – Estonian Newspaper Texts Summarizer*. Thesis, University of Tartu, 2005.

[6] Microsoft Corporation. *AutoSummarize a Word document*. Available: http://www.microsoft.com/education/

[7] Pachantouris, G., *GreekSum: A Greek Text Summarizer*. Master Thesis, 2004.

[8] Penn, G. *Text Summarization.* Available: http://www.cs.toronto.edu/~gpenn/csc401/summarization.ps

[9] Roca, S. C. *Automatic Text Summarization.* 2001. Available: http://www.uoc.edu/humfil/digithum/digithum3/catala/Art_Climent_uk/climent/climent.html