**SumMe:**

**Filipino-English Summarizer using an abstractive semantic-based approach**

A Thesis

Presented to the Faculty of the

College of Computer and Information Sciences

Polytechnic University of the Philippines

In Partial Fulfillment

Of the Requirements for the Degree

Bachelor of Science in Computer Science

Aranzamendez, Paul Justin A.

Capalad, Justin Allen L.

Evangelista, Renz Gio A.

2015

**ABSTRACT**

Automatic Summarization is a research area that can be applied to a variety of domains. Summarization can be either extractive, which is basically a subset of the original document, or abstractive, which employs deeper understanding of the source document and utilizing natural language generation to truly create a summary “from one’s own words” and not just an extract of the source.   
One of the most common domain in automatic summarization are news articles, with researchers coming up with novel ideas of summarization for different purposes, employing different summarization types and kinds of summarization methods.

However, one particular thing to be noticed in this area is that most summarization systems have little to no tolerance to code-switching texts. Code-switching texts are texts that have a mixture of two languages in a sentence or paragraph that might confuse a reader that is not particularly knowledgeable on one of the languages in the code-switch. This scant tolerance in automatic summarization might result to a lower coherence and cohesion in the resultant summary.

In this paper we tackle Single document summarization using an abstraction method based on semantics. Aside from that, we also keep in mind the possibility of a code-switch in the contents of the text, specifically Filipino-English. We will then compare it with existing summarizers in the market to test for its effectiveness.

Table of Contents

[CHAPTER 1 1](#_Toc412070860)

[1.1 Introduction 1](#_Toc412070861)

[1.2 Background of the Study 3](#_Toc412070862)

[1.3 Statement of the Problem 3](#_Toc412070863)

[1.4 Research Hypothesis 4](#_Toc412070864)

[1.5 Conceptual Framework 4](#_Toc412070865)

[1.5.1 Conceptual Framework of the System 4](#_Toc412070866)

[1.5.2 Conceptual Framework of the Study 5](#_Toc412070867)

[1.6 Significance of the Problem 5](#_Toc412070868)

[1.7 Scope and Limitations 6](#_Toc412070869)

[1.7.1 Scope of the Study 6](#_Toc412070870)

[1.7.2 Limitations of the Study 6](#_Toc412070871)

[1.7.3 Scope of the System 6](#_Toc412070872)

[1.7.4 Limitations of the System 6](#_Toc412070873)

[1.8 Definition of Terms 6](#_Toc412070874)

[1.8.1 Conceptual Definition of Terms 6](#_Toc412070875)

[1.8.2 Operational Definition of Terms 7](#_Toc412070876)

[CHAPTER 2 9](#_Toc412070877)

[2.1 Review of Related Study 9](#_Toc412070878)

[CHAPTER 3 24](#_Toc412070879)

[3.1 Methodology 25](#_Toc412070880)

[3.2 System Architecture 26](#_Toc412070881)

[3.3 Research Paradigm 28](#_Toc412070882)

[3.4 Population of the Study 29](#_Toc412070883)

[3.5 Types of Sampling Technique 29](#_Toc412070884)

[3.6 Data Gathering Procedure 30](#_Toc412070885)

[3.6.1 Preliminaries 30](#_Toc412070886)

[3.6.2 Experimental Method 30](#_Toc412070887)

[3.7 Research Instruments 31](#_Toc412070888)

[3.8 Statistical Treatment 32](#_Toc412070889)

[Bibliography 36](#_Toc412070890)

[Appendix 38](#_Toc412070891)

[Prototype 38](#_Toc412070892)

List of Figures

[Figure 1 IPO Illustration of the System 4](#_Toc408802660)

[Figure 2 IPO Illustration of the Study 5](#_Toc408802661)

[Figure 3 System Architecture 25](#_Toc408802662)

[Figure 4 Iterative and Incremental Development 26](#_Toc408802663)

[Figure 5 Likert Scale 29](#_Toc408802664)

# CHAPTER 1

**THE PROBLEM AND ITS BACKGROUND**

This chapter shows the Introduction of the study, its background, the scope and limitations of the study. This chapter states problem that the study needs to answer and improve.

## 1.1 Introduction

A news article discusses current or recent news of either general interest (i.e. daily newspapers) or of a specific topic (i.e. political or trade news magazines, club newsletters, or technology news websites). A news article can include accounts of eyewitnesses to the happening event. It can contain photographs, accounts, statistics, graphs, recollections, interviews, polls, debates on the topic, etc. Headlines can be used to focus the reader’s attention on a particular (or main) part of the article. The writer can also give facts and detailed information following answers to general questions like who, what, when, where, why and how. (Article (Publishing), n.d.).

Summarization is the art of abstracting key content from one or more information sources (Hahn, 2001), and automatic summarization is the process of using a computer program to create a summary that retains the most important parts of the original document (Automatic Summarization, n.d.).

News nowadays are very detailed and narrates happenings and events vividly. News has been an important part in people's lives, most people starts the day reading or watching news. It helped people to know what's happening today and when someone ask about the news the person would just say the same thing a line of a song said, "*How do I know? I read it in the 'Daily News*" (Paxton).

People reads the Header of every news on Newspapers or on-line to get a thoughtful idea of what the news contains, problems are to wordy articles, and some are redundant that contains looping statements and information. Some writers tends to write long articles because of the Pay-By-The-Word system which the pay a writer gets depends on how many words that writer has written, but this system sometimes leads to redundant news. A blog by a writer once stated that "*Writers are terrible when it comes to talking about money. Many of them feel guilty or are afraid to negotiate for more money, which is insane, but true...Pay-by-the-word discourages reporting, does not reward more talented writers and adds an invisible siphon to editorial budgets.*" (Donatelli, 2013).

Information that news holds are important even if the news was way back 5 years ago. A song by The Rolling Stones said: "*Who wants yesterdays papers?, living a life of constant change, every day means the turn of a page*"(Stones, 1967).There are those researchers and people interested on some cases may read back news, but the idea is to provide summary of only news from before for people that are looking for specific news to read the whole only if that person found it interesting by reading the summary. Summary is shortening body of text and getting only the important details in it.

A summary on is very important just like some flashback news on the TV it holds small portion of time to be read or watched, but it has all important details there is on news. An episode of the TV series Breaking Bad titled "Face Off" shows the news the characters watched states brief but informative fraction of the whole news, where the Main characters are watching the news about an explosion that killed 3 people. The study shows that you can read news on-line without actually reading the whole part of it. Getting only the important details of the news helps lessen the time of scanning stacked news on-line.

## 1.2 Background of the Study

Automatic Text Summarization is a process where a computer summarizes text and returns the text with shorter and with less redundancy. Automatic text summarization has two approaches, abstractive and extractive. Extractive summarization is identifying the most relevant passages in one or more documents, the passages which are sentences or phrases are extracted and pasted together to form shorter text compared to the original with little loss of information. Abstractive summarization compared to extractive is more difficult and challenging it parses the original text in a deep linguistic way and interprets the text in a semantical representation and then generates shorter text, abstract that has the same informational content (Hassel M. , Evaluation of Automatic Text Summarization, 2004). Two properties must be measured in evaluating summarized text, the Compression Ratio (how short the summary compared with the original) and Retention Ratio (how much information retained)(Dalianis & Hassel).

## 1.3 Statement of the Problem

The study aims to summarize news articles written in Filipino and English using abstraction methods under Semantic-based summarization.

**1.**  What is the rating of the system-generated summaries by SumMe in terms of:

a. Compression ratio

b. Retention Ratio

c. Understandability/Coherence

d. Cohesion

e. Q&A Task performance

## 1.4 Research Hypothesis

There is no significant difference between the evaluation of the SumMe and TextTeaser systems mentioned in terms of.

a. Compression ratio

b. Retention Ratio

c. Understandability/Coherence

d. Cohesion

e. Q&A Task performance

## 1.5 Conceptual Framework

### 1.5.1 Conceptual Framework of the System

The system will take as input a News article, either in a text file or a URL from a news website. The taken input will undergo three major processes which are the Preprocessing, the Semantic processing, and the Natural Language Generation. Preprocessing prepares the unstructured text into data that is readable by the Semantic processing block. Here, the data is processed to be a semantic representation which is then reduced and interpreted. The resulting interpretation is then fed into the Natural Language Generation block which then generates sentences and paragraphs for the given interpretation, which is the summary of the given input. Fig **1** shows the Conceptual Framework of the system.

\* Input Text file/news article either Filipino or English

\* URL of news article that will be summarized

\* Input contents undergo preprocessing, semantic representation, reduction and interpretation, which is then fed into the Natural Language Generator

\* Summarized input news article

**INPUT**

**PROCESS**

**OUTPUT**

Figure 1 IPO Illustration of the System

### 1.5.2 Conceptual Framework of the Study

The Study starts on gathering News articles taken from selected Online Resources, the gathered news articles will be used as input on the System. The output summarize news article of the system will be compared to the input news article as to measure the Retention Ratio, Coherence and Cohesion by giving expert(s) Experiment papers. The Compression Ratio will be measured by the developers as it only includes counting sentences numbers. On the experiment paper expert(s) are required to hypothetically formulate questions from the information given on the input article and by answering the questions using information from the output the Answer Recall Lenient (ARL), Answer Recall Strict (ARS) can be measured to get the Answer Recall Average (ARA). Fig 2 shows the Conceptual Framework of the study.

\* Recent news articles from selected websites (Filipino and English)

\* Output Summary of the System

\* Experimental Method

\* Expert(s) will compare the input and output of the system for Retention Ratio computation

\* Expert(s) are ask to make questions from the input article that must be answered by the output

\* List of SVO candidates extracted from the input and weighted, Subject, Verb and Objects are scored. The system will get the 10 highest scored

**INPUT**

**PROCESS**

**OUTPUT**

Figure 2 IPO Illustration of the Study

## 1.6 Significance of the Problem

This study might benefit:

Internet users, as the system will give them a gist of the news article first, giving minimal effort for readers to determine if a certain news article is of use to their intent. To this end, Students, researchers and secretaries might find this useful because it would make easier for them to find articles that match their interest, since the system’s output will show the general idea of the input article.

## 1.7 Scope and Limitations

### 1.7.1 Scope of the Study

The study will focus on the development and the assessment of the performance of the application with regards to of the summary of news articles. The study will show if there is a significant difference between the proposed system and other existing news summarization systems.

### 1.7.2 Limitations of the Study

The study will only span the assessment of the proposed application and a select number of existing news summarization systems in terms of evaluation metrics which are: Compression ratio, Retention Ratio, Understandability/Coherence, Cohesion and Q&A Task performance. The sample news articles will only be taken from rappler.com, inquirer.net, and the Philippine star.

### 1.7.3 Scope of the System

The system will summarize news articles from an input text file or URL. The system will parse both raw text and html encoded documents. The content will use preprocessing techniques and Semantic-based abstraction approach in summarizing the given document. The system will be a web app for easy access in the web.

### 1.7.4 Limitations of the System

The System will only accept input that is either in Filipino, English or both. The corpus that will be used for the Filipino language is comprised mostly of Tagalog words, with no regard for the other dialects.

## 1.8 Definition of Terms

### 1.8.1 Conceptual Definition of Terms

**Abstraction -** is the forming of ideas from specific given samples.

**Abstractive Summarization** - parses the original text in a deep linguistic way and interprets the text in a semantical representation and then generates shorter text

**Automatic Text Summarization -** is a process where computer summarizes text and lessen redundancy, but keeps the coherency

**Natural Language Generation** **-** deals with rule-bound text generation; it is used to provide coherent and concise output

**Natural Language Processing -** is a branch of Artificial Intelligence that deals with human language (natural language) to deal with mechanical problems either spoken or written.

### 1.8.2 Operational Definition of Terms

**Automatic summarization -** is the process where the application will process the article or news given by the user.

**Compression Ratio -** measures the shortness of summary of the output over the original length of the source document

**Data Scraping -** is a processed that involves URL input, Data scraping will get the information on selected html tags to be converted to text.

**Lexicon -** is a dictionary where semantically generated text will be stored to form structured texts

**News Article -** is a body of text that contains news entities, this will serve as the input of the system

**Parts of Speech Tagger -** will serve as an indicator or a trace for each word in a sentence. Each word in the sentence will have their tag according to the part of speech they are categorized.

**Question Game -** is a method that will be used to measure the accuracy of the expected output, it involves proponents that will read the output on three ways: the output without reading the original text, the output after reading the source document and reading the source document.

**Retention Ratio -** measures the output summarized text on how much information will be retained over the information on the source document.

**Semantic Graph –** the rules and standards for the correct structures of a sentence.

**Semantic Ontology -** serves as a database for the proper and coherent sentence.

**Sentence Splitter -** is a pre-process that will split every sentence on each paragraph of the given news and article

**Summary -** will be the output after the application is finished on processing and evaluating the given input

**URL -** stands for Uniform Resource Locator, is a protocol that identifies a specific address on the internet.

# CHAPTER 2

**REVIEW OF RELATED LITERATURES AND STUDIES**

This chapter covers the different literatures and studies that benefit and will be used as guidelines throughout the whole research. This chapter shows overview of the gathered literatures and studies from different researchers and authors that show relevance to the study and how it will affect the research and the study made by the researchers of this study.

## 

## 2.1 Review of Related Study

Automatic Text Summarization is split into two groups, text extraction and text abstraction. The text extraction identifies the most relevant passages in one or more documents. These passages, often sentences or phrases being extracted then combine together to form a non-redundant summary that is shorter than the original document (Hassel M. , Evaluation of Automatic Text Summarization, 2004).

Text abstraction is harder to do than text extraction in a way that it needs to parse the original text and interpreting the text semantically to produce a formal representation (Hassel M. , Evaluation of Automatic Text Summarization, 2004). The product of the text abstraction is much closer to what humans might produce when summarizing a document.

One of the important elements of summarization is to distinguish the part of speech of each word in a sentence. The significant duty in summarization is identifying the topic by finding the cue phrases.

To enhance the text summarizer and to acquire the topic of the sentence, Martin Hasselused Name Entity Recognition. The job of the Named Entity Recognition is to classify proper nouns the document. Clearly, the Name Entity carries clues to the topic of the document. But one of the problem they encountered in NER is a serious losses of sentences by prioritizing the elaborative sentences over introductory(Hassel M. ). Another problem is repetitive of the summary for the reason that a certain word is being repeated for numerous times in a document.

One of the existing summarization nowadays is the SweSum summarization engine (Dalianis & Hassel).SweSum is the first automatic text summarizer for Swedish news text created by Dalianis and Hassel. SweSum uses the most common text extractor paradigm. The method of this paradigm is to extract the most significant phrase or sentences from a text to create a new shorter non redundant text (Hassel M. , Evaluation of Automatic Text Summarization, 2004).

The domain of the SweSum is Swedish newspaper text. It utilizes several different topic identification schemes. One of the scheme is the bold tag used to emphasize contents of the text. Headings are given a higher weight. The most relevant information is always presented at the beginning of the newspaper text and are given higher scores. According to the others who used the engine that the performance is estimated to be as good as the state of the art techniques for English. Good Summaries at compression rates around 70% (retaining 30% of the words) can be obtained for original text of two to three pages in the news domain (Dalianis & Hassel, 2001). To identify the important topics of a text, Named Entity Recognition can enhance the identification of key text segments (Hassel M. ).

Additionally, the preprocessor for Swesum is SweNam that tags all found Named Entities with one of the four possible categories – name of persons, locations, companies, brands, products and time stamp. According to Dalianis &Astrom, the Named Entities found by SweNum are fairly reliable with a 92 percent precision. The recall is only 46 percent(Hassel M. ).

One of the problem of the using the Name Entity module tends to prioritize elaborative sentences over introductory that is responsible for some losses of background information. However, Name Entities clearly carry clues to distinguish the topic of a text. It is also helpful in knowing the main part of the different participants and their respective role in a text (Hassel M. ).

The SweSum works in three different passes, the first one is the tokenization and keyword extraction. The second is the ranking of the sentences and the last pass is the produced summary (Delianis & Hassel, How Short is Good? An evaluation of automatic summarization). SweSum’s domain consist mainly of Swedish HTML tagged newspaper text. SweSum ignores HTML tags that control format of the page but processes the HTML tags that control the format of text (Delianis & Hassel, How Short is Good? An evaluation of automatic summarization).Also the web page allows the user to specify the text to be summarized, and the degree of summarization. Besides that, the user is asked what language to be used for the text to apply what language-specific resources (Delianis & Hassel, How Short is Good? An evaluation of automatic summarization).

Furthermore, the function of the tokenization is to split the text into sentences. The SweSum then execute topic detection and identifies the important parts of the text by assigning scores to sentences according to the criteria. The summarizer uses dictionary with about 700.000 that contains the open class words and their stems (Delianis & Hassel, How Short is Good? An evaluation of automatic summarization).

Another automatic text summarizer is the GreekSum, it is based on well-known SweSum summarization engine. Basically it is a text summarizer for Greek news that uses a Greek key word dictionary provided by NCSR Demokritos, Athens (Pachantouris, 2004-2005).

Greek native speakers used SweSum’s built in function of Generic Summarization when evaluating the GreekSum [NewsByte].

A further method that is implemented by Martin Hassel is using a text pre-processor, called PRM (Pronoun Resolution Module) written in Perl and works in SweSum. (Hassel M. , Pronominal Resolution in Automatic Text Summarization, 2000) PRM uses lists of likely focuses (Sidner, 1984) and uses lexicon of nouns that contains information about each entry’s natural or grammatical gender (Hassel M. , Pronominal Resolution in Automatic Text Summarization, 2000). The procedure of this is that the nominal phrase is recognized and categorized then move it to a suitable list of category. They also found that using the PRM in the SweSum greatly improve the coherence for text rich with pronouns.

In the paper of Yihong Gong and Xin Liu, they propose two generic text summarization methods that create text summaries by ranking and extracting sentences from the original documents. The first method is IR methods that ranks sentences significance, while the other method latent semantic analysis technique that identifies important sentences semantically (Gong & Liu). They also compared the manual summarization created by human to the summaries of the generated by the two methods mentioned. Their observation on the experiment shows the larger the documents are, the more disparities between the summaries of humans and the summarization generated by the text summarization.

A system from SRA international, Inc. named The Knowledge Management (KM) uses extraction method for summarization. It features using morphological analysis, name tagging and co-reference resolution. To determine the optimal combination of the features in combination with statistical combination from the corpus to identify the best sentences to include in a summary, they used a machine learning technique (sra.com).

Another research made by Dalianis et al. is the ScandSum that aims to for the development on summarization tools particularly for the Scandinavian languages (Danish, Norwegian and Swedish). ScandSum is also uses extractive text paradigm and based on the SweSum. The research has been successfully ported Danish to Norwegian (Delianis & Hassel, How Short is Good? An evaluation of automatic summarization).

NorSum is the same as the SwesSum with the cooperation with the ScandSum network. They decided to collect a corpus containing manual summaries from the Norwegian newspaper, BergensTidende. The editorial works in the newspaper articles were shortened by removing the last few sentences or rewriting the whole text. Before inserting the news, it was slightly edited to fit the right format and were automatically divided into sentences that were each given a unique ID.

The OGI/OHSU baseline multilingual multi-document summarization system is a summarization system developed by Fisher et al (2000).

They use sentence ranking based on simple features and sentence selection from a ranked list. SVMlight is used to learn preference ranking and use English text sentences when selecting from the ranked list.

According to Angheluta et al. (2002) study, they have developed a topic segmentation algorithm. It detects the structures in the text by using the generic text structure cues. They called the process “Layered topic segmentation that correlate key terms with each topic or subtopic and outputs a tree-like table of content (TOC). The trees use the most significant terms at general and more specific levels of topicality.

The result of the TOCs make both single-document abstracts and the multi-document abstracts and extracts.

Additionally, there are divisions in methods for evaluation automatic text summarization system, namely into intrinsic and extrinsic evaluation methods (Hassel M. ).

The intrinsic evaluation measures the system in of itself. It is done by comparing gold standard, made by a reference summarization system. The main focus of the intrinsic evaluation is on the coherence and how informative the summaries are (Hassel M. ).

While the extrinsic evaluation measures the efficiency and acceptability of the generated summaries like the relevance assessment or reading comprehension (Hassel M. ).

Extraction-based summarization sometimes suffers from coherence problem like anaphors or gaps in rhetorical structure of the summary.

Moreover, the precision and recall figures can be used to assess the performance in utility figures and content based methods. Precision and recall are often combined in F-Score for standard measures. One of the problem of this measure is that they are not efficient in distinguishing between many possible and the summaries differ in the content (Hassel M. ).

The KTH extract Corpus tool assists in the collection of extract-based summaries provided by human informants and semi-automatic evaluation of machine generated extracts in order to easily evaluate the SweSum summarizer. The corpus contains a number of original full texts and several man-made extracts for each text (Delianis & Hassel, Generation of Reference Summaries, 2005).

The compiled of the extracted corpus can analyzed automatically that the inclusion of the sentences in the several extracts for a given source text can easily be compared. The development of an automatic summarizer allows a quick adjustment and evaluation cycle (Delianis & Hassel, Generation of Reference Summaries, 2005).

In addition, the KTH extract tool collects the statistics on how many times a specific extract unit from a text has been included in a number of different summaries. The model summary can be composed using only the most frequently chosen sentences (Delianis & Hassel, Generation of Reference Summaries, 2005).

The group of text from two Swedish and one Danish group have been collected during iterations.

The Swedish extract corpus consists of a total of 301 Swedish text extracts submitted by 45 informants; average length of submitted extracts is currently 32.5 percent (31% and 34% for group 1 and group 2) (Delianis & Hassel, Generation of Reference Summaries, 2005).

While in Danish extract corpus consists of 135 Danish text extracts submitted by 15 informants; average length of submitted extracts is currently 32%.

Numerous equally good summaries for one specific source text effectively making the evaluation against one rigid reference summary unsatisfactory in the automatic text summarization. The evaluation methods have different compression rates that are optimal for different text types or genre, or even within a different text type or genre (Delianis & Hassel, Generation of Reference Summaries, 2005).

The research on summary generation technique mostly relies on the extraction of salient sentence from the original to generate the summary. Several methods for determining the importance of a sentence has been developed. Some algorithms calculate scores for each sentence considering the location of the sentence and word frequencies (Dalianis et al. 2003), while others use semantic information (Wordnet, for example), in order to find the hierarchy of concepts.

Natural Language Generation, the study of human language generation, is a multidisciplinary enterprise, requires expertise in areas of computer science and linguistics. The aim is to learn how computer program produce high-level, narural language from computer depiction of information (Batang, 2006).

Furthermore, according to Herzog (1999), natural language generation often is 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 force into linguistic in nature, 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 in turn to a sentence generator, task that can be further categorized into sentence planning (Batang, 2006).

There are different types of generation can be classified into four main categories:

The first one is the Canned text systems that uses the simplest approach for single-text sentence and multi-sentence text generation. They are trivial to create, but very inflexible (Batang, 2006).

The second one is the Template systems that are the next level of complexity. It relies on the application of pre-defined templates or schemas and is able to support flexible alterations. The template approach is used mainly for multi-sentence generation, particularly in applications whose text is fairly regular in sentence (Batang, 2006).

The third one is the Phrase-based systems that employ what can be seen as generalized templates. In such systems, a phrasal pattern is first selected to match the top level of the input, and each part of the pattern is recursively expanded into a more specific phrasal pattern that matches some sub portion of the input. At the sentence level, the phrases resemble phrase structure grammar rules and at the discourse level, they play the role of text plans. The last one is the Feature-based systems, which are as yet restricted to single-sentence generations, represent each possible minimal alternative of expression by a single feature. Accordingly, each sentence is specified by a unique set of features. In this framework, generation consists in the incremental collection of features appropriate for each portion of the input. Feature collection itself can be based either on unification or on the traversal of a feature selection network. The expressive power of the approach is very high since any distinction in language can be added to the system as a feature. Sophisticated feature-based generators, however, require very complex input and make it difficult to maintain feature interrelationships and control feature selection (Batang, 2006).

Many natural language generation systems follows a hybrid approach by combining components that utilize different techniques. This study used the summarization by abstraction with the Phrase-Merging summary generation (Batang, 2006).

The online News Summarizer by Cruz et al. (2003) 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 SVM machines. Instead of using traditional approach to automate text summarization such as lexical chains, machine learning based summarization, discourse trees, and sentence extraction techniques; the researchers used vector machines SVMs in determining which sentences are possibly included in the summary. The main component of the project is the summarizer engine. It first takes as input text file of the news article. It segments the text into words and sentences. It then scores these words based on a certain number of features. The scores are then input for the SVM module, which determines which sentences, are important, and should be included in the summary. The summarizer then produces the cohesive summary. The system will produce summarization results at three levels: keywords, salient sentences, rough summaries – this is the order of increasing complexity and increasing difficulty of their automatic generation. They used method in their summarization. In preprocessing method, it consists of lemmatization (finding root forms) and tagging (finding part of speech classes), and will use lexical database WordNet (Miller, 1990). The system uses syntactic knowledge and frequency analysis of the text to classify noun phrases into five categories of technicality. It also performs automatic syntactic category disambiguation based on text statistics, using the categories of the neighboring words (Batang, 2006)(Philippines, 2002)

In Determining Keywords method, the Online News Summarizer identified keywords using two methods. First, by considering the frequency of candidate noun groups in a large corpus, and second, by giving a large number of candidate rules to an inductive learning system. Examples of such rules that may help identify keywords are “select noun-phrases that occur frequently in the first paragraph” or “select phrases that occur in titles of sections.” Next, the system is trained on a body of summarized texts, and only the rules that perform well trained (Batang, 2006)(Philippines, 2002).

In Determining Salient Sentences method, Salient sentences will be determined by first identifying the activities and objects relevant to the subject matter communicated by the text. These will be essentially keywords, determined in the Determining Keywords phase. The sentences that convey the most information about the subject matter will then be selected. Selection will be based on a relevancy measure, obtained by counting referential links between the salient activities and objects (Batang, 2006)(Philippines, 2002).

In Assembling Summaries method, they applied sentence truncation. A sentence can be truncated in an orderly way by applying one of predefined discourse-level heuristics, triggered by the presence of specific lexical elements. For example, the sentence “the use of large, public domain linguistic resources,i.e.text corpora and on-line lexicons” can be compressed to “the use of large, public domain linguistic resources.” The relevant heuristic suggests dropping the part that follows “i.e.”. They used the approach to the extraction of rhetorical structure presented by Sumita et al. (1993), which groups sentences into tree of relations. For instance, one can distinguish the structure of examples, parallel argument (Firstly, Secondly, Then…), specialization (This is particularly true of…), and so on. The rhetorical structure acquired in this manner can then be compressed by removing entire sentences belonging to be a certain rhetorical category, such as specialization for example (Batang, 2006).

Tagalog-English code-switching (TECS) is widely accepted among bilingual speakers in the Philippines and communities around the world. Despite the prevalence of this phenomenon, there has been limited work focusing on this language pair in the codeswitching literature. The code-switching literature has been dominated by language pairs that are typologically similar. Further study on typologically dissimilar pairs such as Tagalog and English will be extremely valuable in understanding the mechanisms underlying code-switching (Labitigan, 2013).

Code-switching, one particular phenomenon of bilingual speech, refers to instances of alternating between two languages or varieties of the same language in the same conversation (Myers Scotton, 1983). The linguistic research on code-switching can be generally grouped according to two approaches: structural and sociolinguistic. The structural approach seeks to characterize how code-switching can be represented in the mind. The sociolinguistic approach views code-switching as a sociopragmaticphenomenon, focusing on the social motivations and functions of code-switching (Amuda, 1994). Although both of these main perspectives are invaluable in order to fully understand code-switching, this paper focuses on the grammatical structure of codeswitching(Labitigan, 2013).

Tagalog-English code-switching (TECS), or Taglish, is a variety of bilingual speech. Although it can be considered a prestige language variety, TECS has aincreasingly substantial presence in all socioeconomic classes in the Philippines, particularly in urban centers. TECS is also widely spoken by bilingual communities throughout the world (Labitigan, 2013).

It is well known among linguists that nouns are the most code-switched or borrowed forms. In the case of TECS, the nominal domain, which is the subject of this paper, provides many interesting phenomena that require explanation. For the remainder of the paper, I will often be using the term nominal phrase when referring to constituents in the nominal domain, that is, constituents headed by an N. Nominal phrase is a noncommittal term that helps our analysis in two ways. First, nominal phrases of English and of Tagalog seem to behave very differently. Nominal phrases in Tagalog remain achallenging topic of study, and there still remain many competing ideas about their structure and properties. Thus, a general term such as nominal phrase when referring to Tagalog relieves our analysis of some unnecessary complexity. Second, different types of English nominal phrases (i.e. Ns, modified NPs, conjoined NPs, DPs) seem to sometimes pattern together in TECS (for example, see section 3.2.3.1). Thus, an umbrella term helps to capture certain patterns in the data (Labitigan, 2013).

Unlike many other approaches to code-switching, the Matrix Language Frame (MLF) Model (Myers-Scotton, 1993 [1997]) is not simply a collection of descriptive constraints. Rather, it is a multi-layer model with interconnected parts that not only describes linguistic phenomena, but also provides an explanation for why these phenomena occur. At the core of the MLF Model are two key oppositions based on asymmetries in code-switching structures: the Matrix Language (ML) – Embedded Language (EL) opposition and the content-system morpheme opposition (Labitigan, 2013).

The first opposition stems from the fact that the languages involved in codeswitching do not participate equally. A higher level of participation does not refer to a greater number of morphemes or even the presence of certain morphemes, but rather the contribution of more abstract structure. The language that contributes more abstract structure can be referred to as the Matrix Language (ML), while the other language can be referred to as the Embedded Language (EL). The unit of analysis of the MLF Model is the CP (projection of complementizer). A CP is the highest projection of the clause. This unit of analysis for the MLF Modeldoes not only account for the data regarding distributions of the two participating languages, but it also provides an easily identifiable and consistent unit for comparisons across examples and languages. Referring to the CP also allows us to avoid the technical difficulty in defining and distinguishing among other terms such as sentence, clause, and utterance in our analysis. For each CP, there is a grammatical frame specified. This frame, called the ML,is abstract in nature; it does not itself include any actual morphemes, but rather, “it includes specifications about slots and how they are to be filled, based on directions from lemmas in the mental lexicon” (Myers-Scotton, 2002, p. 67). There is quantitative evidence suggesting that the ML cannot switch within a CP (Finlayson et al., 1998), making the CP the appropriate atomic structure for the study of code-switching. In monolingual speech, the ML frame of each CP is “vacuously transparent” (Jake et al., 2002, p. 72) since the frame is provided by the speaker’s only language. In bilingual speech, this frame may be provided by either one of the two participating languages or, in certain types of contact phenomena, by a combination of the two (Labitigan, 2013).

Unlike many other models, the MLF Model is lexically based. That is, rather than relying solely on principles of monolingual phrase structure to develop accounts for codeswitching, the model underscores abstract procedures in and related to the mental lexicon. Some of these procedures involve phrase structure, but also include other abstract levels. The Abstract Level Model, largely stemming from psycholinguistic models for language production (Levelt, 1993), was developed by Myers-Scotton and Jake (2000) as a supporting model to the MLF Model, but can also stand alone as a description of the levels of abstract lexical structure. The Abstract Level Model designates three levels of abstract lexical structure: lexical-conceptual structure, predicate-argument structure, and morphological realization patterns. A lemma, or an entry in the mental lexicon that maps abstract structure to surface realizations, is represented at all three levels. Thus, the Abstract Level Model serves to trace the path of a linguistic utterance from its beginnings as abstract structure to its manifestation as surface structure.The origin of an utterance comes from an abstract bundle of languageindependent speaker intentions. These intentions activate an abstract entity known as the Conceptualizer, which refines the message and decides what information is to be communicated linguistically and para-linguistically. The Conceptualizer triggers semantic/pragmatic feature bundles, and the ones that are language-specific are then mapped onto lemmas in the mental lexicon. This mapping forms the first level of abstract lexical structure, lexical-conceptual structure. Once lemmas are active, their morphosyntactic properties (or instructions) can be accessed by the Formulator in order to generate hierarchical morphosyntacticstructures.This requires two levels of structure, which involve the language-specific encoding or structural assignment of relations between content morphemes. The first of these twolevels, predicate-argument structure, deals with how thematic structure maps onto grammatical relations, and then morphological realization patterns deal with how grammatical relations map onto surface structures (Labitigan, 2013).

One of the main aspects of the so-called “Web 2.0” is increased participation by website users, or a blurring of the distinction between the content provider and the content receiver. One form that this user interaction can take is the sharing of comments on products that users have purchased or services that they have used. Examples abound on websites such as amazon.com, flixster.com,and chapters.indigo.ca. The need for efficient and effective multi-document summarization of these user reviews and other kinds of evaluative text containing opinions and preferences is thus ever-growing (Cheung, 2008).

There are two main approaches to the task of summarization—extraction and abstraction (Hahn and Mani, 2000). Extraction involves concatenating extracts taken from the corpus into a summary, whereas abstraction involves generating novel sentences from information extracted from the corpus. It has been observed that in the context of multi-document summarization ofnews articles, extraction may be inappropriate because it may produce summaries which are overly verbose or biased towards some sources (Barzilay et al., 1999). However, there has been little work identifying specific factors which might affect the performance of each strategy in summarizing evaluative documents containing opinions and preferences, such as customer reviews or blogs. This chapter aims to address this gap by exploring one dimension along which the effectiveness of the two paradigms could vary; namely, the controversiality of the opinions contained in the corpus (Cheung, 2008).

# CHAPTER 3

**RESEARCH METHODOLOGY**

This chapter includes the steps on how the system will be done and how it will work, also the type of respondents , the type of sampling technique to be used and the instruments that will be used on the study and the data gathering procedure.

## 3.1 Methodology

The researchers will first research and gain a thorough understanding of the domain they are after, which is the online news article summarization. They will procure letters of request for them to be able to conduct surveys and ask for professional assistance from the faculty in the Collage of Arts and Letters in the Polytechnic University of the Philippines in the evaluation of the research’s output. The researchers will also collect a given number of news articles dated January 2014 to September 2014 from Rappler, Philippine daily inquirer and the Philippine star. This will serve as the Population for the study. Meanwhile , the 4th year students and Faculty from the College of arts and letters in PUP will serve as the population for the respondents.

The researchers will then proceed on programming the proposed application. This will be accomplished using Python platform. Other researchers of this study will then search for tools that may help in Natural Language Processing and Understanding, the overall look and feel of the web app, and finally the Generation of the summary itself, the Natural Language Generation. After programming is done, or at least a base implementation of a module is done, it is then tested for quality assurance. This repeats over and over until the researchers are satisfied.

After quality is deemed satisfactory by the researchers, integration, deployment and testing of the application is then done, using the Population of the study which contains a corpus of news articles from January 2014 to September 2014.

Data gathering will be done again, now focused on evaluating the summaries generated by the system. The evaluation consists of testing for the Compression ratio, Understandability, Cohesion, Retention Ratio and Q&A task performance.

Understandability and Cohesion will be evaluated using a likert scale and applying the statistical treatment weighted mean to the results. Compression Ratio is an intrinsic property that will be automatically evaluated by the system. Retention Ratio and Q&A task performance however will need manual evaluation by an expert.

The researchers then tabulate and analyze the data, from which the researchers will derive the conclusion and thus publish the results.

## 3.2 System Architecture

The system takes a News article URL or text file as its input. The input will then undergo the preprocessing part where the contents will be put through the sentence split, POS tagging, Language determination, among others. The preprocessed data is then put through a semantic processor which creates a semantic representation of the data with the aid of a Semantic ontology. The said representation is then reduced with the predefined heuristic rules. The resulting reduced representation is then interpreted and fed into the Natural Language generator which generates a body of sentences which is a shortened representative of the original document. **Fig 3** shows the architectural representation of the system.



Figure 3 System Architecture

## 3.3 Research Paradigm



Figure 4 Iterative and Incremental Development

The researchers will use iterative and incremental development on the study (shown on **Fig 4**). Iterative and incremental development is breaking down software development into a smaller chunks and uses phasing to satisfy software requirements.

The first thing that should be done is to plan what should be the elements needed for the development of the software. Determining the elements, what will be the tools needed for the system work and to finish it will be the next step. Analysis and design is visualizing the model throughout the development of the software cycle. In this phase, the implementation of the structure of news articles will be studied and applied to the coding. After implementing the structure of news articles, the researchers will test the system to know if it is working properly and efficiently. Then the researchers will evaluate if the results from the testing will be acceptable or not. If it is not, the cycle starts again by planning what should be needed to be done to fix the problems encountered until it will satisfy the objectives of the study. After all the software requirements will be met and the software development cycle will be finished, the software will now be ready for deployment.

## 3.4 Population of the Study

The Population of the study consists of all online news articles in news websites from Rappler, Inquirer.net, and the Philippine star. The sampling frame that will be used for this study consists of all online news articles from January 2014 to September 2014. Since these News Articles are considered recent for all were published this year.

The Population of the Respondents consists of personnel from the Polytechnic University of the Philippines. The sampling frame that will be used for the respondents consists of senior students currently enrolled in the Polytechnic University of the Philippines, and selected faculty member from the College of Arts and Letters.

## 3.5 Types of Sampling Technique

The sampling technique to be used for this study is the **Simple random sampling** technique. Simple random sampling chooses each sampling unit randomly and entirely by chance. Such that each individual has the same probability of being chosen at any stage during the sampling process and each subset of k individuals has the same probability of being chosen for the sample as any other subset of k individuals. (Yates, Moore, & Starnes). This way, the articles chosen for summarization is completely random.

## 3.6 Data Gathering Procedure

### 3.6.1 Preliminaries

The researchers identify what are the problems stated and what will be the possible solutions to solve it. It also involves knowing who will be benefited after finishing the project. After the researchers recognize the factors in solving the problem and what will be the benefits of the system to the user, the data gathering can be made now.

First, the researchers pinpoint the circumstances to solve the problem by identifying what are the needs of the system to work properly and the needs of the user to meet the expected output. second, finding a suitable and reliable source news articles. third, identifying the benefits of the system to the user as well as to the community. By finding the benefits, the researchers gathered information on how will it affect to the community and to the user. The researchers also will identify who will take the surveys in order to know significance of the study. Fourth, the researchers identify what tools are needed in order to solve the problem stated.

### 3.6.2 Experimental Method

After identifying the problems to be solved, who will be the user and what benefits can be obtained, data gathering can be made. The researchers will find possible solutions to the problems stated on the related topics. The researchers will also identify the benefits of the system to the user and to the community by identifying the factors affecting each other. By giving surveys to a group of people, the researchers will classify what are their speculations and what can the researchers do for improvement. In order to solve the problem stated, the researchers will select tools that are needed to the system. The requirement also includes what software and what platform is to be used by the system.

For the type of questionnaire, Likert Scale survey will be used to allow the participants to respond more effectively to the questions. Furthermore, the researchers will be able to surmise the data more efficiently.

**LIKERT SCALE**

|  |  |
| --- | --- |
| **SCALE** | **INTERPRETATION** |
| 5 | Strongly Agree |
| 4 | Agree |
| 3 | Neutral |
| 2 | Disagree |
| 1 | Strongly Disagree |

**Figure 5 Likert Scale**

## 3.7 Research Instruments

The researchers will provide questionnaires for population to be used to gather information from Individuals that read news on line or from newspapers, each questionnaires indicates questions to answer and weigh in several questions about the research. Gantt Chart was also used as a clerical instrument, the Gantt chart states schedule of activities by the researchers. Activities and deadlines like when things need to be done and what should be done today are all placed on several cells showing activity progress per month. Undone activities are marked red and once done it'll be marked green and yellow if it is currently in progress and lastly black if the activity is dropped down or taken off. Gantt chart helped the researchers to be on date on things. Mechanical Instrument used is Python for the programming part. Django will be used for Web designing part where Web functions will be connected using Python. The researchers are also Python literates that make it easier for the researchers to manipulate the edges in Python programming. The database of the research will be accessed using MySQL.

## 3.8 Statistical Treatment

To assess the generated summaries of the system, the following formulas and treatments were used:

A Likert scale is used for the evaluation of Understandability and Cohesion

**1.** The **Compression ratio** is the property of a summary that shows how much shorter the summary is than the original, denoted by:

Here, the compression ratio is further subdivided into two parts: length of compression ratio *by the number of sentences* and *by the number of words.*

**2. Retention ratio** is the property of a summary that shows how much information is retained in the summary from the original, denoted by:

To assess the system’s ability to convey key facts of the source article, the Q&A Task performance evaluation is going to be used, in which experts are asked to read through the source document and marking central passages. Questions that corresponds to certain factual statements in the central passages are then made.(Hassel M. , Evaluation of Automatic Text Summarization, 2004) After which the questions are then answered using the system generated summary, judged as *Correct, Partially Correct,* and  *missing.*(Mani)

**3. Answer Recall Lenient(ARL)-**  an accuracy metric for the Q&A task, denoted by:

**4. Answer Recall Strict(ARS)-** an accuracy metric for the Q&A task, denoted by:

Where

* **n1** is the number of *Correct* answers
* **n2** the number of *Partially Correct* answers
* **n3** being the number of questions.

**5. Answer Recall Average** is accuracy metric for the Q&A task, which is the average of the **ARL** and **ARS.**(Mani)

**6.** A **Mean** is the measure of the central tendency of a probability distribution.



Where

* is the summation of frequency
* N is the number of respondents

**7.** A **Standard Deviation** measures the amount of variation from the average, denoted by



Where

* **C:\Documents and Settings\lenovo\My Documents\Downloads\x-bar.png**is the mean
* x is the sample
* n is the respondents

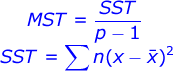
**9. Anova** is a statistical test which analyzes variance. It is helpful in making comparison of two or more means which enables a researcher to draw various results and predictions about two or more sets of data.

Anova Formula

Where

* F is Anova Coefficient
* MST is Mean sum of squares due to treatment
* MSE is Mean sum of squares due to error.

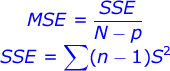
Formula for MST is given below:



Where

* SSE = Sum of squares due to error
* S = Standard deviation of the samples
* N = Total number of observations

Formula for MSE is given below:



Where

* SSE = Sum of squares due to error
* S = Standard deviation of the samples
* N = Total number of observations.

For the evaluation of the summaries, the **Sample mean and standard deviation** for each of the abovementioned formulas (*Compression Ratio, Retention Ratio, Answer Recall Strict, Answer Recall Lenient and Answer Recall Average)*. An Anova test is then conducted to determine if the system developed is significantly different from textTeaser and other existing summarizers in terms of the evaluation criterion given in the first problem statement.

# Bibliography

*Article (Publishing)*. (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Article\_(publishing)

*Automatic Summarization*. (n.d.). Retrieved from Wikipedia: http://en.wikipedia.org/wiki/Automatic\_summarization

Batang, E. (2006). *NewsByes: A Tagalog News Text Summarization by Abstraction.* Philippines.

Cheung, J. (2008). *Comparing Abstractive and Extractive Summarization of Evaluative Text: Controversiality and Content Selection.*

Dalianis, & Hassel. (2000, 2004).

Dalianis, & Hassel. (2001).

Dalianis, H., & Hassel, M. (n.d.). *Generation of Reference summary.*

Davis, G. (n.d.). Sampling. In *Davis/Gallardo's Straight Talk About Communication Research Methods 1st edtion* (p. 164). Kendall Hunt Publishing Corporation.

Delianis, & Hassel. (2005). *Generation of Reference Summaries.*

Delianis, & Hassel. (n.d.). *How Short is Good? An evaluation of automatic summarization.*

Donatelli, J. (2013). *Pay-By-The-Word Journalism is Ridiculous.*

Gong, Y., & Liu, X. (n.d.). *Generic Text Summarization Using Relevance and Latent Semantic Analysis.* C & C Research Laboratories USA.

Hahn. (2001). *The Challenges of Automatic Summarization.*

Hassel, M. (2000). *Pronominal Resolution in Automatic Text Summarization.*

Hassel, M. (2004). *Evaluation of Automatic Text Summarization.*

Hassel, M. (n.d.). *Exploitation of Named Entities in Automatic Text Summarization for Swedish.*

Khan, S. (2014). *A Review on Abstractive Summarization Methods.*

Labitigan, R. (2013). *Tagalog-English code-switching: issues in the nominal domain.* Philippines.

Mani. (n.d.). *The TIPSTER SUMMAC Text Summarization Evaluation.*

Pachantouris, G. (2004-2005). *GreekSum.*

Paxton, T. (n.d.). *Daily News.*

Philippines, U. o. (2002). *Online News Summarizer http://www.engg.upd.edu.ph/~naval/cs198/cs198/2002/ons/index.html.*

Sidner, C. (1984). *Focusing in the Comprehension of Definite Anaphora. Berwich& Brady eds., Computational Models of Discourse.* London.

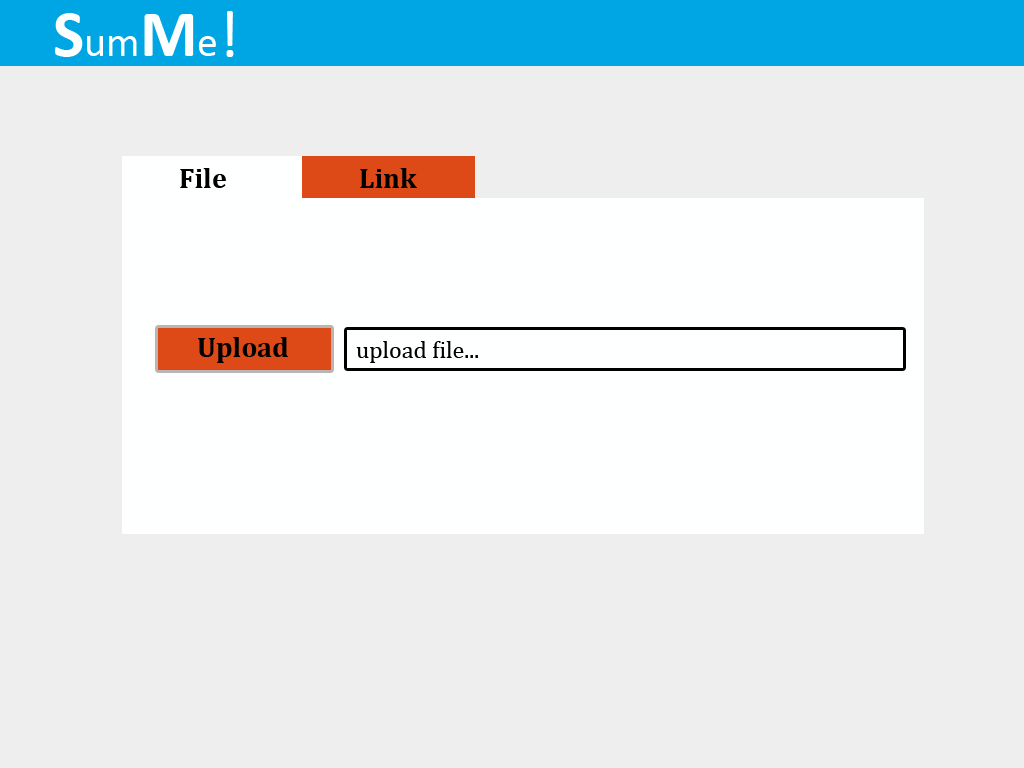
*sra.com.* (n.d.).

Stones, T. R. (1967). *Yesterday's Papers.*

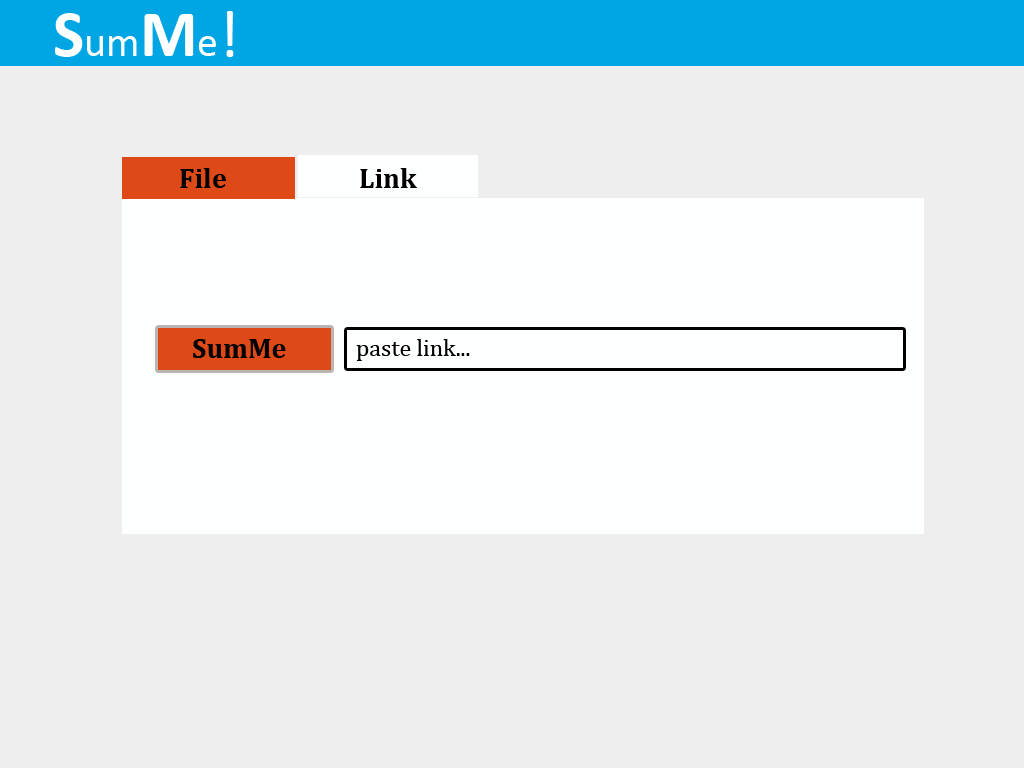
Yates, D., Moore, D. S., & Starnes, D. S. (n.d.). *The Practice of Statistics, 3rd Ed. Freeman.*

# Appendix

## Prototype



The sample prototype presented above shows one of the possible input type the system is going to accept. In this case, the input type is a text file (.txt).



The illustration above shows another way on how to input file to the system, this shows input using URL links from the internet to be processed.