Dynamic Weight-Based Approach for Thesis Title Recommendation

Frances Frangelico S. Friginal, Gerald T. Cayabyab, and Bartolome T. Tanguilig III
Technological Institute of the Philippines, Quezon City, Philippines
Email: {giofriginal26, gerald.cayabyab}@gmail.com, bttanguilig_3@yahoo.com

***Abstract*—The development of an application that can help the thesis adviser and student construct and recommend thesis titles through the use of Natural Language Processing (NLP) was the primary objective of this study. Resembling the concept of Information Retrieval under Query-Focused Summarization, words except stop words from the related literature and studies of the thesis and its title were used as a training corpus while the words except stop words of chapters one (1), two (2) and three (3) of the thesis or merely the test document acts as the query in the corpus, where it retrieved its weight from both training and test records. The result showed that out of forty-five (45) thesis titles, twenty-eight (28) different title formats were constructed by the thesis advisers with the support of the developed application. The advanced application obtained the accuracy score of 1.55 or "Accurate" and quality score of 4.06 or "Good" in fifteen (15) thesis titles.**

***Index Terms*—thesis title, electronic resources, natural language processing, information retrieval, query-focused summarization**

I. Introduction

Choosing a thesis title is interesting but most of the time it needs the professional guidance of experts to decide whether or not the chosen title is significant enough to work on as a study. There are considerations in writing a thesis title; grammar, style and creativity should be taken into account, furthermore, the student needs to consider the goals, issues, gathered data and findings in constructing a title as well. Selecting a topic is the first stage in conducting a thesis. Choosing the right thesis topic is vital. It takes much planning and data gathering. The student must not only appraise the possibility of the study, moreover, consider several things like interests, ideas, goals, capabilities, and budget. A topic is different from a title. A topic is the core of any study regardless of the field. It is the area of exploration where the student uses his or her knowledge and resources. Consequently, a title is the statement formulated from the topic where creativity and style are required to come up with a good one. The main objective in choosing a title is to capture the attention of the readers while providing them a picture of the study's main idea without having to read the entire article.

Jin (2003) acknowledged the use of automatic title generation through the use of Natural Language Processing to suggest titles for a particular document. However, the author stated that the challenging aspect in automatic title generation is that it requires both the understanding of the essential content of a document and the knowledge to create a headline that reflects the content in only a few words which helps people to quickly understand the relevant information contained in a document. The author also recognized the two broad approaches to title generation; the automatic text summarization based approaches and statistical approaches. Automatic text summarization approaches treat titles as summaries with extremely short length and use text summarization techniques directly for generating titles. Statistical approach emphasizes the idea of learning the title document correlation from training corpus and applying the scientific model to create titles for unseen documents. His previous work was focused on the examining and comparing the seven different statistical methods for word title selection. These include the nearest neighbor approach, K-nearest neighbor approach, decision tree approach, statistical translation approach, reverse information retrieval approach, Naïve Bayes approach with a limited vocabulary, and Naïve Bayes approach with a full vocabulary [1].

In accordance with Jin's theories, the proponent used the statistical approach towards automatic title generation combining a new approach which the proponent developed. The new approach is an application enabling the thesis adviser and student to construct and recommend thesis titles through the use of Natural Language Processing based on the concept of Information Retrieval under Query-Focused Summarization where electronic resources can be used for training documents. Training documents are the collection of valid studies and information that will add weight and importance to every word in the thesis document making it a candidate word for thesis titles. The valid studies will be derived from the related literature and studies of the thesis while test documents will be the thesis itself that will be used for recommending and constructing thesis titles with the help of training documents. Chapters one (1), two (2) and three (3) will be utilized as test documents. The proponent designed his procedures and styles strategically to process the candidate words from the pre-processing phase up to the selected candidate words organization. Furthermore, different techniques in Natural Language Processing were applied in four (4) major phases starting from the pre-processing phase, Term

Manuscript received January 29, 2016; revised July 26, 2016.

doi: 10.18178/ijsps.4.6.528-536

Frequency Inverse Document Frequency (TF.IDF) algorithm, computation of N-grams using Maximum Likelihood Estimates and the organization of the selected candidate words into the right sequence according to the chosen format.

II. Materials and Methods

In this part, the proponent will discuss the four (4) major phases used in the developed application: the pre-processing stage, the application of TF.IDF algorithm step, the computation of N-grams using Maximum Likelihood phase and the selected word organization phase. It will also show the system procedure, criteria and the system constraints of the developed application.

A. Pre-Processing Phase

The first stage of the procedure is the pre-processing steps. It is imperative to perform these steps before the main process can be executed. Every Natural Language Processing must be normalized first before the storing and processing of the inputted set of texts. This step is called Text Normalization. Spraot and Bendrick (2011) stated that Text Normalization is the process of transforming text into a single canonical form that it might not have had before. Normalizing text before storing or handling it allows for separation of concerns since the input is guaranteed to be consistent before operations perform on it. Additionally, they determined that there is no proper normalization procedure [2].

Jurafsky and Manning (2012) stated that there are three (3) stages of text normalization. These are the segmentation or tokenization of words in running text, normalizing words formats such as lemmatization and Part of Speech (POS) and segmenting sentences in running text. However, the proponent decided to change the order of text normalization starting from sentence segmentation followed by POS tagging, tokenization, and lemmatization [3].

In this study, with the help of Natural Language Toolkit (NLTK) libraries, text normalization will apply to document preprocessing procedure since the entire focus of this study is purely in text format.

The following section will show the method for preprocessing phase.

- Sentence segmentation
- POS tagging
  - ➢ The POS tagger uses an algorithm that corresponds to a particular part of speech, based on its definition, as well as its context, relationship with adjacent and related words in a phrase, sentence, or paragraph since a word can be more than one part of speech.
  - ➢ When it comes to hyphenated words, POS can also determine their type. For instance the word “centroid-based” or “compound-splitter” or “k-gram”, the POS tagger will label these words as nouns. Another instance is the word “one-sentence” or “break-even”, the POS tagger will mark these words as adjectives.
  - ➢ A possessive noun will be labelled as a noun and a possessive noun.
- Tokenization
  - ➢ Tokenize a set of text into words.
  - ➢ Remove stop words. Get the stop word’s list from SMART.
  - ➢ Remove arithmetic operators, relational operators, logical operators, increment and decrement operator and special characters such as currency symbols and other special symbols that cannot be part of a title.
  - ➢ Treat hyphenated compound words as one token
  - ➢ Consider closed compound words as one token unless the student chooses to make it two words.
  - ➢ Divide an open compound word and possessive nouns into two or three words.
  - ➢ Treat words that contain periods such as “m.ph” or “PhD.” as one token.
- Lemmatization
  - ➢ The hyphenated format will be treated as one word and in its original form, thus there’s no need to lemmatize. The same concept will apply to possessive nouns.
  - ➢ Convert all words into lowercase format except for abbreviations

B. TF*IDF Algorithm Phase

Nenkova and McdKeown (2011) stated that Query-Focused Summarization summarizes only the information in the input document that is relevant to a particular user query. For instance, in the context of information retrieval, given a query issued by the user and a set of relevant documents retrieved by the search engine, a summary of each document could make it easier for the user to determine which document is relevant. Snippets are example output of search engine. The author also stated that TF.IDF of words as traditionally used in information retrieval can be employed. This weighting exploits counts from an extensive background corpus, which is an extensive collection of documents, usually from the same genre as the document that needs to summarize. The background corpus serves as an indication of how often a word may be expected to appear in an arbitrary text. They also indicated that the TF.IDF weights of terms are useful indicators of importance, and they are easy and fast to compute. These properties explain why TF.IDF is one of the most commonly used features for extractive summarization [4].

Jurafsky and Manning (2012) stated that TF.IDF formula has two parts, Term Frequency (TF) and Inverse Document Frequency (IDF). Term Frequency measures the frequency of term usage in a document. Since every document is different in length, it is possible that a word would appear several times in longer documents. Thus, the term frequency is often divided by the document length as a way of normalization. In Inverse Document Frequency, the rare terms are more important than familiar words [5].

Also, Jurafsky and Manning acknowledged that raw term frequency is not the goal. They stated that a document where a term occurred ten (10) times was more relevant than a document where the term occurred once but it does not follow that it is ten (10) times more relevant. Relevance does not increase proportionally with the term frequency. Thus, they recommended getting the weight of term frequency formula by applying the logarithmic function plus the value of one (1) [6].

$$w_{t,d} = \begin{cases} 1 + log_{10} tf_{t,d}\ , & if\ tf_{t,d} > 0 \\ 0, & otherwise \end{cases} \quad (1)$$

Moreover, Jurafsky and Manning also discussed the concept of Inverse Document Frequency (IDF). IDF is another score used for weighting the matches' document to the query. The idea behind this concept is that rare terms are more descriptive than familiar terms such as common words. They stated that for familiar terms, it must have positive weights for words but lower weights than for rare times. The IDF weight is the document frequency of *t*, the number of documents that contain *t*. df_t is an inverse measure of the informativeness of *t* [7]. The formula for IDF is the following.

$$idf_t = log_{10}\left(\frac{N}{df_t}\right) \quad (2)$$

where *N* is the total number of documents in the entire document collection for search; df_t is the number of documents where term *t* occurs. Putting together the TF.IDF weight of a term is the product of its *tf* weight and its *idf* weight. It increases with the number of occurrences within a document and the rarity of the term in the collection.

$$w_{t,d} = \left(1 + log tf_{t,d}\right) \times log_{10}\left(\frac{N}{df_t}\right) \quad (3)$$

The following section will show the application of TF.IDF algorithm for candidate title word selection. Let:

- ***F** be the frequency of words*
- $\boldsymbol{w \in T}$ be the distinct words that used in the title from training documents except stop words
- $\boldsymbol{w \in TED}$ be the distinct words from test documents
- $\boldsymbol{w \in TRD}$ be the distinct words from training documents
- ***N*** be the documents in the training collection
- |***N***| be the number of training documents
- ***TD*** be the test document
- ***D1*** be the matrix for $\boldsymbol{w \in T}$ and N corresponding to $\boldsymbol{w \in TRD}$
- $\boldsymbol{D1_{[ij]}}$ be the number of occurrences between $\boldsymbol{w \in T}$ and ***N*** corresponding to $\boldsymbol{w \in TRD}$
- ***D2*** be the matrix for $\boldsymbol{w \in TED}$ and ***N*** corresponding to $\boldsymbol{w \in TRD}$
- $\boldsymbol{D2_{[ij]}}$ be the number of occurrences between $\boldsymbol{w \in TED}$ and ***N*** corresponding to $\boldsymbol{w \in TRD}$
- ***W*** be the weight value
- $\boldsymbol{w \in T}$ and $\boldsymbol{w \in TED}$ will be treat as queries q in a large set of ***N*** that corresponds to $\boldsymbol{w \in TRD}$.

1) Let ($\boldsymbol{w \in T}$) will be in ***D1*** corresponding to the *n* columns while *n* rows for (***N***).

2) Count $\boldsymbol{f}$ of $\boldsymbol{w \in T}$ where $\boldsymbol{w \in T}$ and $\boldsymbol{w \in TRD}$ in ***N*** overlapped as a value for each $\boldsymbol{D1_{[ij]}}$.

3) Compute the Term Frequency $\boldsymbol{tf_{w \in T,\ N}}$ using the formula (1).

$$\boldsymbol{W_{w\in T,N}} = \begin{cases} \boldsymbol{1 + log_{10} tf_{w\in T,N\ (w\in T)}}\ , & \boldsymbol{if\ tf_{w\in T,N}} > 0 \\ \boldsymbol{0}, & \boldsymbol{otherwise} \end{cases}$$

where $\boldsymbol{w \in T}$ is defined as the number of times $\boldsymbol{w \in T}$ occurs in ***N*** corresponding to $\boldsymbol{w \in TRD}$ for every value in $\boldsymbol{D1_{[ij]}}$. To get the weight of every $\boldsymbol{w \in T}$.

4) Compute for the total weight of $\boldsymbol{w \in T}$ using the formula:

$$\boldsymbol{Total\ Weight\ Score(w \in T)} = \sum_{(\mathbf{w\in T})\in N\cap \mathbf{w\in TD}} \left(\boldsymbol{1 + log tf_{t,d}}\right) \quad (4)$$

where all the **W** of $\boldsymbol{D1_{[ij]}}$ per row was computed by getting the summation of each value of $\boldsymbol{D1_{[ij]}}$.

5) Get the value of |**N**| and compute the Inverse Document Frequency by using the formula (2).

$$\boldsymbol{idf_t = log_{10}\left(\frac{|N|}{Nf_{w\in T}}\right)}$$

6) Calculate the dot value of $\boldsymbol{Total\ Weight\ Score(w \in T)}$ (3) and $\boldsymbol{idf_{w\in T}}$ (3) to compute the total ***W*** for each $\boldsymbol{w \in T}$**.**

7) Let $\boldsymbol{w \in TED}$ will be in ***D2*** corresponding to the *n* columns while *n* rows for (***N***).

8) Get the value of $\boldsymbol{f}$ of $\boldsymbol{w \in TED}$ where $\boldsymbol{w \in TED}$ and $\boldsymbol{w \in TRD}$ in ***N*** overlapped as a value for each $\boldsymbol{D2_{[ij]}}$.

9) Compute the Term Frequency $\boldsymbol{tf_{w \in TED,N}}$ using the formula (1).

$$\boldsymbol{W_{w\in TED,N}} = \begin{cases} \boldsymbol{1 + log_{10} tf_{w\in TED,N\ (log10(w\in T\ \ ))}} \\ \boldsymbol{if\ tf_{w\in TED,N}} > 0 \\ \boldsymbol{0}, \quad \boldsymbol{otherwise} \end{cases}$$

where $\boldsymbol{w \in TED}$ is defined as the number of times $\boldsymbol{w \in TED}$ occurs in ***N*** corresponding to $\boldsymbol{w \in TRD}$ for every value in $\boldsymbol{D1_{[ij]}}$ to get the weight of every $\boldsymbol{w \in TED}$.

10) Compute for the total weight of $\boldsymbol{w \in T}$ using the formula (4):

$$\boldsymbol{Total\ Weight\ Score\ (w \in TED)} = \sum_{(\boldsymbol{w\in TED})\in N\cap \boldsymbol{w\in TD}} \left(\boldsymbol{1 + log tf_{t,d}}\right)$$

where all the **W** of $\boldsymbol{D2_{[ij]}}$ per row was computed by getting the summation of each value of $\boldsymbol{D2_{[ij]}}$.

11) Get the value of |**N**| and compute the Inverse Document Frequency using the formula (2) $\boldsymbol{idf_{w\in TED} = log_{10}\left(\frac{|N|}{Nf_{w\in TED}}\right)}$.

12) Compute the dot value using formula (3) $\boldsymbol{Total\ Weight\ Score(w \in TED)}$ and $\boldsymbol{idf_{w\in TED}}$ to compute the total ***W*** for each $\boldsymbol{w \in TED}$.

13) Count $\boldsymbol{f}$ of $\boldsymbol{w \in TED}$ from ***TD***

14) Get the ***W*** of $\boldsymbol{w \in TED}$ by applying Term Frequency (1) $\boldsymbol{tf_{w \in TED,\ TD}}$ using:

$$\boldsymbol{W_{w\in TED,TD}} = \begin{cases} \boldsymbol{1 + log_{10} tf_{w\in TED,TD\ (log10(w\in T\ \ ))}} \\ \boldsymbol{if\ tf_{w\in TED,TD}} > 0 \\ \boldsymbol{0}, \quad \boldsymbol{otherwise} \end{cases}$$

15) Compute the final ***W*** by getting the summation of $\boldsymbol{Total\ Weight\ Score(w \in T)}$, $\boldsymbol{Total\ Weight\ Score\ (w \in TED)}$ and $\boldsymbol{W_{w\in TED,TD}}$

Resembling the concept of Information Retrieval under Query-Focused Summarization, words except stop words from related literature and studies of the thesis and its titles will be used as the training corpus while the words except stop words from chapters one (1), two (2) and three (3) of the thesis or simply the test document will act as the query in the training corpus; it will retrieve its weight from both training and test documents. See Fig. 1.

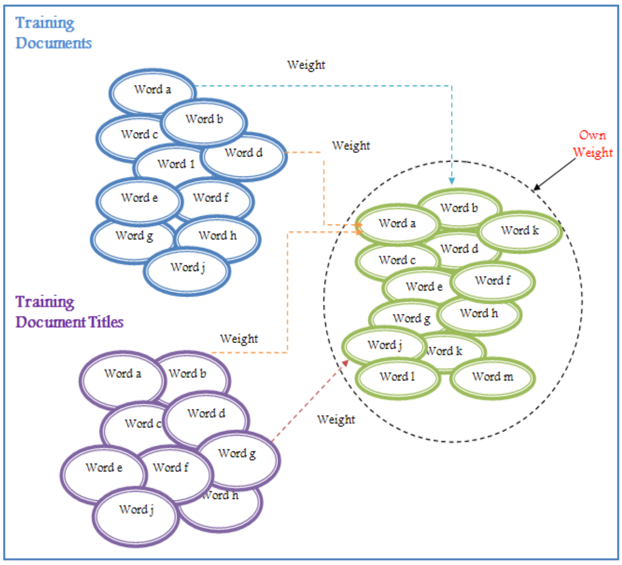

Figure 1. Developed application behaviour

Since the developed application is using the TF.IDF algorithm, to retrieve the weight of all the words used from training documents as well in its title words, it can give additional weight to the words used from the test document to get the candidate words that can be used as title words. Words from test documents have its weight since the developed application is applying the Term Frequency formula by default. The numbers of training documents can significantly affect the weight of each word used in the test document. Thus, it is important that those documents should be relevant to the test document. A word that doesn't exist in the test document has a slight probability to appear as candidate word since it does not have its weight in the test document.

C. N-grams Using Maximum Likelihood Phase

The main purpose of this phase is to normalize open compound and possessive word issues since it will undergo tokenization process. This phase is the process of getting the bigram and trigram by taking the probability that words may be considered as one term and be emphasized as one vocabulary.

Jurafsky and Martin acknowledged that using the Maximum Likelihood Estimate (MLE) is the best way to estimate bigram or N-gram probabilities. MLE estimate the parameters of an N-gram model by getting counts from a corpus and normalize the numbers so that they lie between zero and one [7].

To compute a particular bigram probability of a word *y* given a previous word *x*, it will compute the count of the bigram *C(xy)* and normalize by the sum of all the bigrams that share the same first word *x*.

$$P(w_n|w_{n-1}w_n) = \frac{C(w_{n-1}w_n)}{\sum_w C(w_{n-1})} \quad (5)$$

To simplify the equation: the sum of all bigram counts that start a given word must be equal to the unigram count for that word w_{n-1}.

$$P(w_n|w_{n-1}w_n) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})} \quad (6)$$

In this study, the proponent used the concept of N-gram model using MLE to assign the probability of sequencing of the selected title words. After the process of this phase, the thesis adviser or student can see both the output for unigram, bigram and trigram words. The proponent used bigram model alone. However, the proponent has its way to get the set of trigram words by getting the intersection of two (2) bigram words.

The following section will show the application of MLE for taking the bigram and trigram words.

Let:

- ***a*** be the last word in bigram x
- ***b*** be the first word in bigram y
- ***c*** be the first word in bigram x
- ***d*** be the last word in bigram y
- ***STW*** be the selected title words
- **S*∈N*** collection of sentences from the training documents
- **S*∈TD*** collection of sentences from the test document
- **TP** be the table for Maximum Likelihood Estimation value for each bigram
- $\mathbf{TP}_{[I,J]}$ be the probability of $\boldsymbol{STW}_{[nrow]}$ and $\boldsymbol{STW}_{[ncolumn]}$ occur together from **S*∈N*** and **S*∈TD***
- ***[TG]*** be the set of trigrams

1) Let ***STW*** will be in ***TP*** for *n* rows that correspond to $\boldsymbol{STW}_n$ and $\boldsymbol{STW}_{n-1}$ for *n* columns.

2) Get the value of $\mathbf{TP}_{[I,J]}$ using formula (6) $P(w_n|w_{n-1}w_n) = \frac{C(w_{n-1}w_n)}{C(w_{n-1})}$ Maximum Likelihood Estimation formula where the probability of the current word and the last word before the current word will occur based on **S*∈N*** and ***S∈TD.***

3) Get the highest probability value of $P(w_n|w_{n-1}w_n)$ in $\mathbf{TP}_{[I,J]}$ for each row.

4) Arrange the bigram words starting from the highest up to lowest.

5) if ***a*** = ***b*** then ignore ***b***, x+y = ***[tg]***

D. Selected Word Organization Phase

In this phase, the thesis adviser can organize the selected title words according to their preference. The proponent set simple rules on how the thesis adviser will organize those title words. Title formats are composed of words from the different part of speech, see Fig. 2. For instance, the thesis adviser wants to use the combination of verbs and nouns in the different arrangement, the advanced application permitted that to attain the goal of developed application's flexibility and thesis adviser or student's creativity. They have the right to decide the format of the title as they prefer it. The proponent set simple rules for title construction.

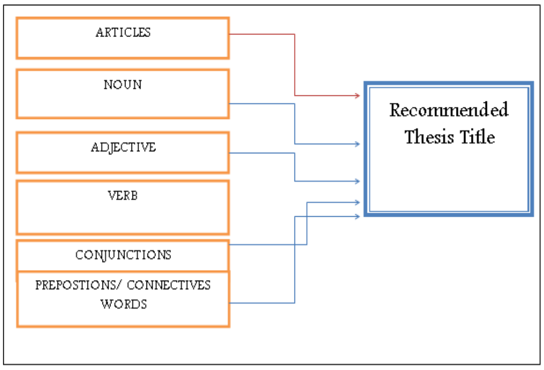

Figure 2. Thesis title components

These are the following:

1) Articles are optional in every first word on each part of the title.
2) Using preposition or connective words means that a new part of the title will be included.
3) Article "the" will be considered as the connective word if the first word in a title part is a verb that ends in "ing." Ex. "Identifying *the* Effective Elements"
4) Conjunctions are optional in every word
5) Conjunction "and" will be considered as a connective word if the following word has a different POS tag. Ex. "The Handsome (adjective) *and* Liana (noun)" which is different to "John (noun) and Liana (noun)."

E. System Procedure

This part explains how the developed application works from the input to the final output of the software.

Combining the four (4) major phases (Pre-processing Phase, TFIDF Algorithm Phase, MLE Phase and Selected Words Organization Phase), the proponent came up with the overall application processes. See Fig. 3.

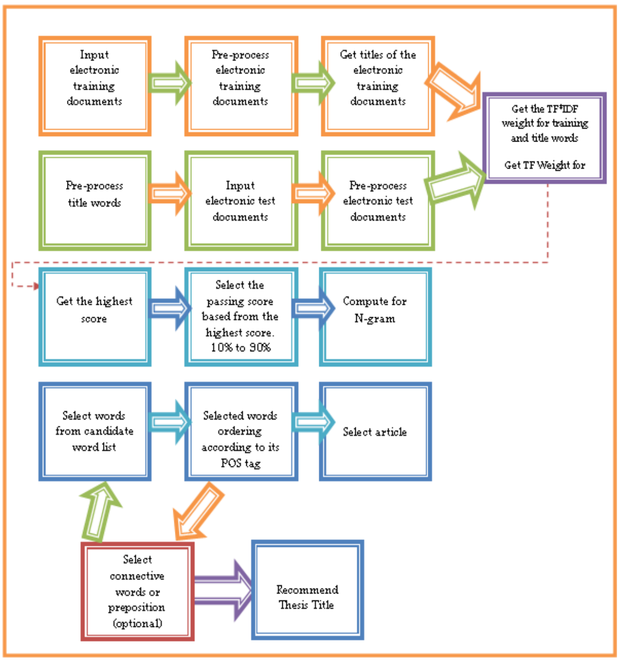

Figure 3. Overall developed application processes

F. Developed Application Features

The following section describes the characteristics of the developed application.

1) *Getting the passing score*

- Passing score determines the candidate words for the title. The thesis adviser or student can set a passing score based on the highest score computed from phase one.
  - For instance, the maximum score word is "system", with the TF*IDF score of 2.50. If the passing score is 90%, the passing score value will be 2.25, thus all the words with 2.25 and above will be included in the candidate word's list.
- The thesis adviser or student can select between 10% up to 90% as passing rate.

2) *Synonyms suggestion*

- The developed system is capable of suggesting synonyms of the word if its POS tag is verb or adjectives.
- This function can also acquire the different POS tag of a particular word under its synonyms.
- Adding Wildcards
- The developed system can include wild card words for flexibility purposes.
- The thesis adviser or student can add the necessary word that they want to include in their title as long as its POS tag is the noun, verb or adjective.
- This function can add color to the title especially if the word has the adjective or verb POS tag.

3) *Word ordering*

- After the thesis adviser or student selected words from candidate word's list, the developed system will automatically get the word's POS tag and put into its designated POS tag list, the same process will apply if the thesis adviser or student chooses to add wild card words.
- After all the selected words were collected and inputted in the designated POS tag list, the thesis adviser or student can choose the determiner.
- After the determiner was selected, they can choose the order of POS tag list. They can also set the conjunction of each POS tag list provided it is composed of two (2) or more elements in the list.
- The thesis adviser or student can include connective word/s or preposition/s which will form part of the new title.

4) *Adverb exclusion*

- The proponent conducted a survey on what POS tag is the most used in a title. The proponent collected Three thousand five (3005) technical thesis titles around the web in different fields such as in Information Technology, Computer Science, Medicine, Psychology, Engineering, Mathematics, Business Management and other Science branches. After the titles had been collected, the proponent applied the POS tagger in nlpdotnet.com/services/Tagger website. The website generated 12,413 tokens.

Table I shows that out of 12,413 tokens, the noun is the most commonly used POS in a title with 8,171 words or 66%, followed by the adjective with 930 words or 7% while verb with 161 words or 1%. Other POS tags such as determiners, conjunctions and prepositions had 3,128 words or 25% of usage. Although adverb POS has 23-word instances, it averaged to only 0%. It only depicts that in a title, adverb POS is not usually applied in technical thesis title. In this case, the proponent decided not to include adverb words in the candidate word's list.

TABLE I. POS TAG SUMMARY

POS TAG SUMMARY		
POS TAG	Number of Tokens	Percent Score
Noun	8171	66%
Verb	161	1%
Adjective	930	7%
Adverb	23	0%
Determiner / Conjunction / Preposition	3128	25%
Total Major Words	**9285**	**75%**
Total Words	**12413**	

G. *Criteria*

This section will show the constructed thesis titles and the various title formats that the thesis advisers created with the support of the developed application. It will also exhibit the accuracy and quality results of the developed application's output.

According to Jin, in general, there are two (2) major factors that will influence the quality of machine-generated titles. 1) Consistency, whether the machine-generated title can reflect the main content of the document since the function of a title is to provide a very brief summary of a document, and a good title should be able to indicate the main points of the document clearly and 2) Readability pertains to whether the machine-generated titles are readable to human subjects.

The author also stated that there are two evaluation metric types for evaluating machine-generated titles. These are the Automatic Evaluation Metrics and the Manual Evaluation Metrics. In Automatic Evaluation Metrics, the basic idea is that the score is based on the number of overlapped words between human-assigned titles and machine-generated titles. This concept is called F1 metric which is widely used for Information Retrieval measure to test accuracy. It composed of Precision and Recall score [8].

Jurafsky and Manning defined Precision as the percentage of selected elements that are correct and Recall as the percentage of correct items that are selected. They also discussed the use of the 2-by-2 Contingency Table (see Table II), wherein a particular piece of data being evaluated has essentially four states, on one axis we are choosing whether this piece of data correctly belongs to a class or not. These states represent True Positive, False Positive, False Negative and True Negative which will be used for the computation of Precision and Recall [9].

TABLE II. THE 2-BY-2 CONTINGENCY TABLE

Element of Class	**Not Element of Class**
True Positive	False Positive
False Negative	True Negative

Jin used the Precision for the machine-generated title on the human-assigned title is defined as the number of matched words between the machine-generated title and the human-assigned title divided by the length of the machine-generated title. On the otherhand, Recall for the machine-generated title on the human-assigned title is defined as the number of matched words between them divided by the length of the human assigned titles [1]. The formula of F1 metric can express as:

$$F1 = \frac{2*precision*recall}{precision} + recall \qquad (7)$$

The proponent used the F1 metric to obtain the accuracy result of the developed application's output on the original thesis title constructed by the thesis adviser alone. However, the overlapped words only applied to the major POS words such as noun, adjective, and verb. Helping words such as articles, conjunctions, prepositions and connective words were excluded in the formula. To get the corresponding representation of the percentage score (see Table III), the proponent set a range matching the level equivalent of accuracy based on the perfect score of three (3).

TABLE III. PERCENTAGE RANGE OF ACCURACY

Range	**Descriptive Equivalent**
> 2	Highly Accurate
≥ 1	Accurate
≥ .8	Moderate Accurate
≥ .7	Slightly Accurate
< .7	Not Accurate

For Manual Evaluation Metrics, the proponent used the metric of Jin where it tested the quality of the recommended thesis title based on the human assessment. Jin set up very simple standards for assessors since people can have different standards for proper titles and different people can have quite different opinions on the same titles. The author is considering the limitations of automatic evaluation metrics. The author believes that human judgment of human subjects is vital in evaluating the quality of machine-generated titles. Of course, because of the flexibility of titles, it could be quite difficult for an assessor to judge the quality of titles. People can have different standards for good titles as well as varied different opinions on the same titles. In addition, they set up very simple standards for assessors, i.e., a title is good as long as it reflected the primary content of a document and organized in a human readable way. The author also stated that by clearly defining the evaluation standard for the assessor, there is a way to avoid the fluctuation in the judgment due to diversity of tastes in titles [1]. The five (5) categories (see Table IV below) for human judgment are the following:

TABLE IV. THE MANUAL EVALUATION CRITERIA

Numerical Rating	Category	Implication
5	Very Good	The title reflects the main content of the document and the word sequence is readable.
4	Good	The title indicates the content of the document however some words are not smoothly connected with the others.
3	Satisfactory	The title contains significant phrases, however, some of the word orders are not correct.
2	Unsatisfactory	Many words within the title are not related to the document even though one or two of them are important content words. The whole word sequence is totally not readable.
1	Extremely Unsatisfactory	None of the words in the title makes sense to the document.

The proponent used the Manual Evaluation Metrics, a simple score scheme adapted from Jin, to obtain the quality of the developed application's output based on the thesis adviser's assessment. The score five (5) for the category of 'Very Good', four (4) for 'Good', three (3) for 'Satisfactory', two (2) for 'Unsatisfactory' and one (1) for 'Extremely Unsatisfactory'.

The average score of human judgment was used as the final evaluation metric thus, the higher the score of individual judgment the better the quality of machine-generated titles. The proponent applied the Arithmetic Mean, which is the most commonly used average or measure of the central tendency to obtain the overall scores of accuracy and quality results of the developed application.

H. Evaluation Procedure

The respondents of this research are composed of five (5) thesis advisers together with their three (3) thesis electronic documents from their advisee; thesis advisers with at least five (5) years of advising experience formed part of the main criteria. The proponent and the adviser met personally to use the developed application. The proponent instructed the adviser to prepare three (3) soft copies of theses without a title together with its supporting documents used in the related literature and studies. To avoid subjective evaluation, the adviser does not have any idea that the proponent will ask the original title of the thesis afterwards. After the adviser utilized the developed application, the proponent asked the adviser to write down on a piece of paper their constructed titles and to choose only one title that they preferred, since the developed application can create many titles in one thesis.

After using the developed application, the proponent presented the evaluation survey form which the adviser filled up. The adviser wrote down the original titles of each thesis and their personally preferred constructed title from the developed application for Automatic Evaluation Metric. They also wrote down the score of the constructed title for Manual Evaluation Metric. The titles that were constructed but not chosen by the adviser were included in the "Other Recommended Thesis" section of the survey form.

After the advisers completed the survey questionnaire, the proponent performed the necessary computations for Automatic Evaluation Metric for every recommended thesis titles from the developed application concerning the original title of the theses. The proponent also performed the necessary computations to get the overall accuracy score for Automatic Evaluation Metric, the overall quality score for Manual Evaluation Metric and the total number of different thesis formats constructed by the advisers including in the "Other Recommended Thesis" section.

I. User Requirements

Purpose: To develop an application that can help the thesis adviser and student to construct and recommend thesis titles through the use of Natural Language Processing was the primary objective of this study. Specifically, based on the thesis adviser or student's preference, the developed application can support the construction and recommendation of thesis titles in different formats. In addition, the accuracy and quality results of the developed application's output were the final goal of the proponent.

Scope: This study focused on recommending and constructing thesis title according to the student or thesis adviser's preference, since title formats are composed of words classified under the different part of speech. The proponent wants to clarify that this developed application will only assist the thesis adviser and student to construct and recommend thesis titles but not generate one as a final thesis title. The thesis adviser or student has the liberty to choose or to add some words if they want to include it in the title.

Moreover, the developed application is only applicable to every English language and technical thesis. The student will submit their electronic documents such as chapters one (1), two (2) and three (3) and some of the related literature and related studies they used in the study. The developed application will only accept electronic resources that can convert into text forms such as E-journals, E-Manuscripts, E-books, E-Research Reports, E-Mail, E-thesis, and WebPages. Other multimedia files such as scanned documents, pictures, sounds, videos are excluded in the study. Lastly, aside from stop words, the developed application also omitted the words that are in adverb POS tag since it is not usually employed in the technical thesis titles.

Users: Thesis advisers and students can use the developed application.

III. RESULTS AND DISCUSSIONS

This section presents the results of the study from the respondents' responses to the following:

1) Construct and recommend thesis titles in different formats based on the thesis adviser or student's preference;

2) Obtain the accuracy result of the developed application's output on the original thesis title constructed by the thesis adviser alone;
3) Obtain the quality of the developed application's output based on the thesis adviser's assessment.

A. Constructed Thesis Titles in Different Format Results

POS Legend

- A - Adjective
- C - Connective Word
- N - Noun
- V - Verb

In this result, the proponent only showed the major POS words as well as the connective words. Connective words are a set of words that connects an independent part of a title to another separate part of the title. The proponent disregarded the minor POS words such the determiners and conjunctions since these small words are merely optional.

TABLE V. DIFFERENT POS FORMAT OF THE CONSTRUCTED THESIS TITLE

No.	POS	Count
1	A N C N	6
2	A N C N C N	4
3	N C N	3
4	N C A N C N	
5	N C A N	2
6	V N C A N C A N	
7	V N C N C N	
8	V C N C N	
9	A N C A N C A N	1
10	A N C A N C N	
11	V N C N	
12	A N C V C N C A N	
13	A N C V N C A N	
14	A V C N C A N	
15	N	
16	N A N C N	
17	N C A N C N V	
18	A C N V C N	
19	V C N C N C N	
20	N C N C A N	
21	N C N C A N C N	
22	N C N C N	
23	N C N C V N	
24	N C N C V N C N	
25	N C V N C N	
26	N CN	
27	V A N C N	
28	N A N C N C N	

These thesis titles were from the constructed titles of the thesis advisers with different formats that totalled to forty-five (45) thesis titles. Table V shows the twenty-eight (28) different title formats out of forty-five (45) thesis titles that were built by the thesis advisers. The highest total number of title format is the "A N C N" that acquired six (6) out of twenty-eight (28) or 21% of POS form count. The "A N C N C N" format acquired four (4) or 14%, while "N C N" and "N C A N C N" format got three (3) or 10%. The "N C A N", "V N C A N C A N", "V N C N C N" and "V C N C N" format got two (2) or 7% whereas others got only one (1) count in the POS format. The proponent clarifies that in this result, the primary goal is to show the different title formats that the developed application can bring into play and not the most preferred title format. Since the highest percentage is only twenty-one (21) percent, this is too small to justify the adviser's preferred thesis title format.

Table VI below shows that the thesis title number three (3), entitled "Locator of Gravestone using An Android-based Mobile Application within Manila Memorial Park" acquired the perfect score of three (3) or "Very Accurate". All the words overlap against the original thesis title entitled "Manila Memorial Park Gravestone Locator Application" as well as greater than in length which an advantage of the newly constructed title since it can have a more meaningful title. Oppositely, thesis title number six (6) entitled "The WOW Philippines Android-based Mobile Gameplay for Filipinos" acquired the lowest score of .6 or "Not Accurate". Only two (2) words overlapped against the original thesis title entitled "Operativeness of the World of Wonders: WOW Philippines Flash Game among People Currently Living in Metro Manila" and less than in length.

TABLE VI. THESIS TITLE F1 SCORE SAMPLES

No.	Original Title	Constructed Title with Support of the Developed Application	F1 Score	Descriptive Equivalent
3	Manila Memorial Park Gravestone Locator Application	Locator of Gravestone using An Android-based Mobile Application within Manila Memorial Park	3	Highly Accurate
6	Operativeness of the World of Wonders: WOW Philippines Flash Game among People Currently Living in Metro Manila	The WOW Philippines Android-based Mobile Gameplay for Filipinos	.6	Not Accurate

TABLE VII. OVERALL ACCURACY RESULTS

Total Number of Thesis Title	Average Accuracy Score	Descriptive Equivalent
15	1.55	Accurate

The proponent computed the overall average of the F1 score and the quality scores based on the fifteen (15) thesis titles selected by the advisers. Table VII below shows the overall F1 score that correspond to the overall accuracy result, and Table VIII shows the overall quality result. The quality scores for every thesis titles were purely based on the thesis adviser's judgment.

TABLE VIII. OVERALL QUALITY RESULTS

Total Number of Thesis Title	Average Quality Score	Descriptive Equivalent
15	4.06	Good

IV. CONCLUSION

In the experiment conducted by the proponent, there were twenty-eight (28) different title formats out of forty-five (45) thesis titles that were built by the thesis advisers. The highest total number of title format is the "A N C N" that acquired twenty-one (21) percent of POS form count. The "A N C N C N" format acquired fourteen (14) percent, while "N C N" and "N C A N C N" format obtained ten (10) percent. The "N C A N", "V N C A N C A N", "V N C N C N" and "V C N C N" format got seven (7) percent whereas others got only one (1) count in the POS format. Based on the analysis and interpretation, the developed application can make different title formats based on the various POS and connective words positioning.

Thesis title number three (3) entitled "Locator of Gravestone using An Android-based Mobile Application within Manila Memorial Park" acquired the perfect score of three (3) or "Very Accurate" since all the words overlapped against the original thesis title entitled "Manila Memorial Park Gravestone Locator Application" as well as greater than in length which an advantage of the newly constructed title since it can have a more meaningful title. Oppositely, thesis title number six (6) entitled "The WOW Philippines Android-based Mobile Gameplay for Filipinos" acquired the lowest score of .6 or "Not Accurate". Only two (2) words overlapped against the original thesis title entitled "Operativeness of the World of Wonders: WOW Philippines Flash Game among People Currently Living in Metro Manila" and less than in length.

The proponent computed the overall average of the F1 score and the quality scores based on the fifteen (15) thesis titles selected by the advisers. The experiment attests that by getting the overall accuracy and quality results of thesis titles with scores of 1.55 or "Accurate" and 4.06 or "Good" from the experiment, it proves that the developed application can support the thesis adviser and student to construct and recommend thesis titles.

ACKNOWLEDGMENT

The authors would like to thank their family and friends for showing support, to other mentors, Technological Institute of the Philippines and lastly to Almighty God.

REFERENCES

[1] R, Jin, *Statistical Approaches toward Automatic Title Generation, ACM Digital Library*, 2003, pp. 26-34.
[2] R. Sproat and S. Bendrick. (2011). Text normalization. [Online]. Available: http://www.cslu.ogi.edu/~sproatr/Courses/TextNorm
[3] D. Jurafsky and C. Manning. (April 22, 2012). 2 - 4 - Word Normalization and Stemming. [Online]. Available: https://www.youtube.com/watch?v=2s7f8mBwnko
[4] A. Nenkova, and K. McKeown, "Automatic summarization," *Foundations and Trends in Information*, vol. 5, no. 2-3, pp. 103-233, 2011.
[5] D. Jurafsky and C. Manning. (April 22, 2012). 19 - 3 – TF*IDF Weighting. [Online]. Available: https://www.youtube.com/watch?v=43WvJU4LaUg
[6] D. Jurafsky and C. Manning. (April 22, 2012). 19 - 3 - Term Frequency Weighting. [Online]. Available: https://www.youtube.com/watch?v=43WvJU4LaUg
[7] D. Jurafsky and C. Manning. (April 22, 2012,). 19 - 3 – Inverse Document Frequency Weighting. [Online]. Available: https://www.youtube.com/watch?v=a50Hv_N-yHA
[8] D. Jurafsky and C. Manning. (April 22, 2012). 4-2 - Estimating N-gram Probabilities. [Online]. Available: https://www.youtube.com/watch?v=o-CvoOkVrnY
[9] D. Jurafsky and C. Manning. (April 22, 2012). 6 - 7 – Precision and Recall. [Online]. Available: https://www.youtube.com/watch?v=2akd6uwtowc

Frances Frangelico S. Friginal was born in Quezon City, Philippines on April 26, 1988. With a Bachelor's Degree in Computer Science at the University of the East, Manila and is currently completing his thesis for his Masters of Information Technology degree at the Technological Institute of the Philippines, Quezon City. An experienced faculty teaching Information Technology and Computer Science, System Analysis and Design, NET Framework Programming Languages and Mathematics. He is also engaged in research in Natural Language Processing Application.

Gerald T. Cayabyab was born in Dagupan City, Philippines on June 7, 1980. He is currently taking up his Doctoral degree in Information Technology at the Technological Institute of the Philippines-Quezon City and the SAP Coordinator and professor in the College of Information Technology Education. He earned his Bachelor's degree of Science in Information Technology at Saint Louis University, Philippines in 2001. He finished his Master's Degree in Information Technology from STI College, Manila, Philippines in 2008. He presently is teaching major courses in the field of Information Technology such as Multimedia System Development, System Design and Development, Information Resource Management, Quality Assurance and Mobile Development for both undergraduate and graduate students.

Dr. Bartolome T. Tanguilig III, a Computer Engineering graduate of Pamantasanng Lungsodng Maynila, Philippines in 1991. A Masteral Degree holder in Computer Science from De La Salle University, Manila, Philippines in 1999, and a Doctor of Philosophy in Technology Management from Technological University of the Philippines, Manila in 2003. His work includes the following: Assistant Vice President for Academic Affairs and concurrent Dean of the College of Information Technology Education and Graduate Programs of the Technological Institute of the Philippines, Quezon City, a member of the Commission on Higher Education (CHED) Technical Panel for IT Education (TPITE), chairperson of the CHED Technical Committee for IT (TCIT), founder of Junior Philippine ITE Researchers (JUPITER), board member of the Philippine Society of IT Educators (PSITE), member of the PCS Information and Computing Accreditation Board (PICAB), member of the Computing Society of the Philippines (CSP) and a program evaluator and an accreditor of the Philippine Association of Colleges and Universities Commission on Accreditation (PACUCOA).