Query-Focused Multi-Document Summarization for Earthquake News Written in Myanmar Language

Myat Myitzu Kyaw
University of Computer Studies, Mandalay
myitzuyii@gmail.com

Abstract

Due to the fast growth of online information on the Internet, there are a large number of journals, magazines, information portals and personal blog sites written in Myanmar language have now online editions. But there need more time to read and grasp it. Manual analysis of the conveyed documents becomes almost impossible. Therefore, the process of summarizing documents is becoming increasingly popular. It is a kind of text mining strategy which involves reducing a text document or multiple documents into a short set of words or paragraph which conveys the main meaning of the text. Although there are many language text summarizers but Myanmar language summarizer is still in research area. Summarizing the important information from the multiple documents written in Myanmar Language at short period of time is not mature. In this paper, when the user queries one or more options, the system extracts the relevant important sentences from the documents. After that, the system abstracts the important information in fewer words and then gives the word level summary to the user that is relevant to the user query options. The system focuses on Earthquake news written in Myanmar Language.

Keywords: Document Summarization, Myanmar Language Summarizer, Text Summarizer, Natural Language Processing, Multi-document Summarizer.

1. Introduction

With the explosive growth of information available on the World Wide Web, it has become more difficult to access relevant information from the Web at a short period of time. Text Summarization is the process of identifying the most salient information in a document or set of documents (for multi document summarization) and conveying it in less space. It became an active field of research in both Information Retrieval (IR) and Natural Language Processing (NLP) communities. Text Summarization process retrieves the most important content from the document. Its purpose is to identify a summary of a document without reading the entire document. Document summarization can be categorized along two different phases: abstract-based and extract-based. An extract-based consists of sentences extracted from the document while an abstract-based may brief words and phrases that do not appear in the original document. The summarization task can also be categorized as either generic or query-oriented. A query-oriented summary presents the information that is most relevant to the given queries, while a generic summary gives an overall sense of the document's content. For query-focused multi-document summarization, a summarizer incorporates user declared queries and generates summaries that not only reflect the important concepts in the input documents but also bias to the queries.

Multi-Document Summarization (MDS) is defined as the task of automatically producing a unique summary of a set of documents on the same topic. MDS is a more complex task than single document summarization as it aims to select sentences relevant to different query-related themes, inside a set of documents, rather than to only shorten a single source text. The purpose of a brief summary is to shorten the information search and to minimize the time by spotting the most relevant facts from the source documents. Generally, an effective summary should be relevant, concise and fluent. It means that the summary should cover the most important concepts in the original document set, contain less redundant information and should
be well-organized.

There are many language summarizers but lack of summarization systems for Myanmar language. Some modification should be added to meet with the specific features of Myanmar language for better performance. For some tasks, existing techniques need to be refined to suit Myanmar language. For certain features of the Myanmar language, new methodologies may have to be added. The system proposed new extraction idea for Myanmar language. Forward-backward algorithm concept is applied in extraction phase to understand the semantic. Longest matching approach is used to reduce redundant information in abstraction phase. Some abstraction methods will also be used to get the better result.

2. Related Work

Yanting LI and Kai CHENG proposed a novel algorithm, called Triangle Sum for key sentence extraction from single document based on graph

theory in Nov, 2011. The proposed algorithm works on single document without training data so that cost can be reduced. They described an efficient method which does not need to build any dictionary or training data or examples before the key sentences extracted from a document to extract triangles from connected graph. **[14]**

Vishal Gupta and Gurpreet Singh Lehal made a survey of Text Summarization Extractive Techniques in Aug, 2010. Summarization methods can be classified into extractive and abstractive summarization. An extractive summarization method based on statistical and linguistic features of sentences. An abstractive summarization method uses linguistic methods to examine and interpret the text. In this survey paper, Vishal Gupta and Gurpreet Singh Lehal presented many variations of the extractive approaches and their pros and corns **[11].**

Aijun Xu proposed query-focused document summarization using concept analysis. It uses Latent Semantic Analysis for finding out the concepts and identifying representative sentences of the concepts; then it uses the correlation measure between two matrices to measure similarity. The summary sentences are selected with the similarity one by one. Experimental result shows that the proposed summarization method is effective, and it can improve the performance of summarization. Also, there is a lot of room for improvement. For example, readability is an important aspect in the performance of summarization. **[1]**

Pinaki Bhaskar and Sivaji Bandyopadhyay presented a graph based approach for query dependent multi document summarization system in 2009. They tested their proposed algorithm with news articles from TAC 2008 data of Update Summarization track. The experimental results suggest that the proposed algorithm is effective and efficient. They tested with 48 document sets each of 10 news articles on a same topic. The proposed algorithm can be improved to handle more noisy WEB articles or work on other domain too. **[7]**

3. Theory Background

3.1. Text Summarization

Text Summarization is a kind of text mining strategy. It is the creation of a shortened version of a text by a computer program It summarizes brief information from a given document while preserving important concepts of the document. Text Summarization can be categorized as extractive and abstractive according to the way the summary is created.

Extractive methods work by selecting a subset of existing words, phrases, or sentences in the original text to form the summary. Abstractive methods build an internal semantic representation. Abstraction use natural language generation techniques to create a summary closer to what a human might generate. Extraction techniques retrieve key clauses sentences while abstraction involves briefing the source document. Abstraction can condense a text more strongly than extraction. The abstractive summarization is similar to the way a person creates a summary. The programs for abstraction are harder to develop as they require the use of natural language generation technology, which itself is a growing field. There are two particular types of summarizations:

- Key phrase extraction - where the goal is to select individual words or phrases to "tag" a document.
- Document summarization - where the goal is to select whole sentences to create a short summary.

Summarization can be either *generic* summaries or *user-focused (query-based)* summaries. A generic summary presents an overall sense of the documents' contents. In those systems, the summary is about the whole document. In query-based text summarization systems, the summary is about the query asked. It presents the important information of the document that are related to the user's query. Summarization of multimedia documents, e.g. pictures or movies, is also possible. Some systems will generate a summary based on a single source document (*single document summarization*). Others can use multiple source documents (*multi-document summarization).*

3.2. Multi-Document Summarization

Multi-Document Summarization is an automatic procedure aimed at extraction of information from multiple texts written about the same topic. It extracts the most important information from a set of documents to generate a compressed summary. It is also creates information reports that are both concise and comprehensive. Multi-Document Summarization is much more complex than summarizing a single document, even a very large one.

There are three level of document summary:

- Word Level – It can give condense facts to the user.
- Sentence Level – It can give regenerated sentences to the user.
- Paragraph Level – It can give abstract summary closer to what a human might generate. It needs to paraphrase.

3.3. Forward-Backward Algorithm

The forward-backward algorithm has very important applications to both hidden Markov models (HMMs) and conditional random fields (CRFs). It is

a dynamic programming algorithm, and is closely related to the Viterbi algorithm for decoding with HMMs or CRFs. The proposed system use forward-backward algorithm concept to understand the semantic.

Inputs: Length m, set of possible states $\mathcal{S}$, function $\psi(s, s', j)$. Define * to be special initial state.

Initialization (forward terms): For all $s \in \mathcal{S}$,

$$\alpha(1, s) = \psi(*, s, 1)$$

Recursion (forward terms): For all $j \in \{2 \dots m\}, s \in \mathcal{S}$,

$$\alpha(j, s) = \sum_{s' \in \mathcal{S}} \alpha(j-1, s') \times \psi(s', s, j)$$

Initialization (backward terms): For all $s \in \mathcal{S}$,

$$\beta(m, s) = 1$$

Recursion (backward terms): For all $j \in \{1 \dots (m-1)\}, s \in \mathcal{S}$,

$$\beta(j, s) = \sum_{s' \in \mathcal{S}} \beta(j+1, s') \times \psi(s, s', j+1)$$

Calculations:

$$Z = \sum_{s \in \mathcal{S}} \alpha(m, s)$$

For all $j \in \{1 \dots m\}, a \in \mathcal{S}$,

$$\mu(j, a) = \alpha(j, a) \times \beta(j, a)$$

For all $j \in \{1 \dots (m-1)\}, a, b \in \mathcal{S}$,

$$\mu(j, a, b) = \alpha(j, a) \times \psi(a, b, j+1) \times \beta(j+1, b)$$

Figure 1. Forward-Backward Algorithm

4. Query-Focused Multi-Document Summarization

When the user chooses one or more query options i.e. , , the proposed system extracts important sentences which contain the query related information according to the query related features from multi-document of the same topic.

After that, the system abstracts the most relevant information from the extracted sentences. Finally, the system gives the world level summary to the user that is related to the query option. The proposed system develops both extraction and abstraction phases to get more accurate summary.

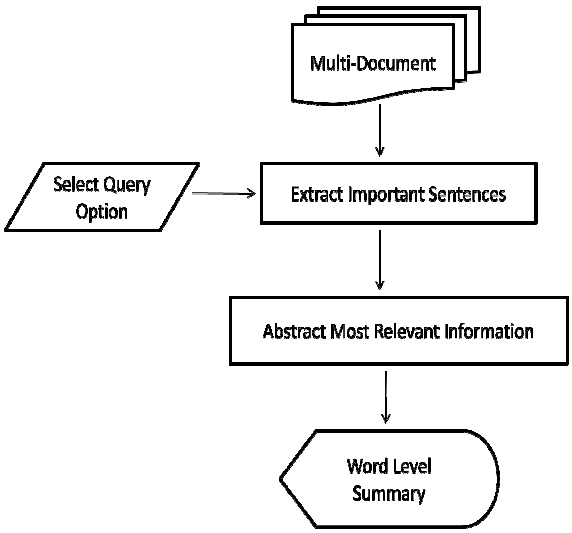

Figure 2. System Overview

မြန်မာနိုင်ငံ ရှမ်းပြည်နယ် လွယ်မွေ့မြို့အနီး အင်အားပြင်းလျင်တစ်ခုလုပ်ခတ်

၂၀၁၁ မတ် ၂၄ ရက်နေ့ မြန်မာစံတော်ချိန်ညနာရီ ၂၉မိနစ်၃၀စက္ကန့်ခန့်ကရှမ်းပြည်နယ် ကျိုင်းတုံမြို့မှ အရှေ့တောင်ဘက် ၃၅ မိုင်ခန့်အကွာ လွယ်မွေ့မြို့အနီးတွင် အင်အားရစ်ချိတာစကေး ၇ ဒသမ ၀အဆင့်ရှိ မြေလျင်တစ်ခုလုပ်ခတ်သွားခဲ့ပါသည်။လျင်ကြောင့် အနည်းဆုံး လူ ၅၁ ဦးသေဆုံးခဲ့ပြီး ဒဏ်ရာရသူ ၂၅၉ဦး ရှိသည်ဟုဒေသခံများနှင့် ချယ်နယ်FM က ပြောသည်။အဆိုပါလျင်ကြောင့် တာချီလိတ်မြို့၊ တာလေမြို့၊ နားယောင်းကျေးရွာ၊ မိုင်းလင်းကျေးရွာ၊ မိုင်းကိုးကျေးရွာအုပ်စု ကျကုန်ရွာစသည့်ဒေသများတွင်ပျက်စီးဆုံးရှုံးမှုများ ရှိကြောင်းသိရှိရသည်။လျင်ကြောင့် တာချီလိတ်မြို့၊ တာလေမြို့၊ နားယောင်းကျေးရွာ၊ မိုင်းလင်းကျေးရွာ၊ မိုင်းကိုးကျေးရွာအုပ်စုကျကုန်ရွာတို့တွင် စုစုပေါင်းအိမ်ခြေ ၂၂၄ လုံးပျက်စီးခဲ့ပြီး ခြောက်ခန်းတွဲ လူနေလိုင်းခန်းတစ်ခု၊ သာသနိကအဆောက်အအုံ ၁၁ ခုဟုနဆိုင်ရာအဆောက်အအုံကိုးလုံးပျက်စီးဆုံးရှုံးခဲ့ရသည်။လျင်ဖါအရှိန်ကို မြန်မာနိုင်ငံရှမ်းပြည်နယ်၊ ရန်ကုန်တိုင်းဒေသကြီး၊ မန္တလေးတိုင်း ဒေသကြီးနှင့် ထိုင်းနိုင်ငံ၊ တရုတ်နိုင်ငံနယ်စပ်ဒေသများ အထိပါခံစားသိရှိခဲ့ရသည်။

Figure 3. Sample Earthquake News

4.1. Typical Earthquake News Features

According to the sample earthquake news, there are eight typical earthquake news features. They are -

- -၂၀၁၁ ၂၄
- -
- - ၇ ဒသမ ၀ ( )
- -

•

•

•

•

According to the earthquake news features, there are three main categories.

- Time

- Place

- Condition

4.2. Query Relevant Features

- Time

- Place

- Condition

4.3. Preprocessing Phase

First, download Earthquake news from Myanmar News sites such as Weekly eleven, Seven Days, The Voice, etc. Then, use KANAUNG Converter to convert Myanmar 3 font. Assume that one news is one document of the same topic. User query options are left hand side of the typical earthquake news features. Segmentation is not consider in this proposed system.

4.4. Extraction Phase

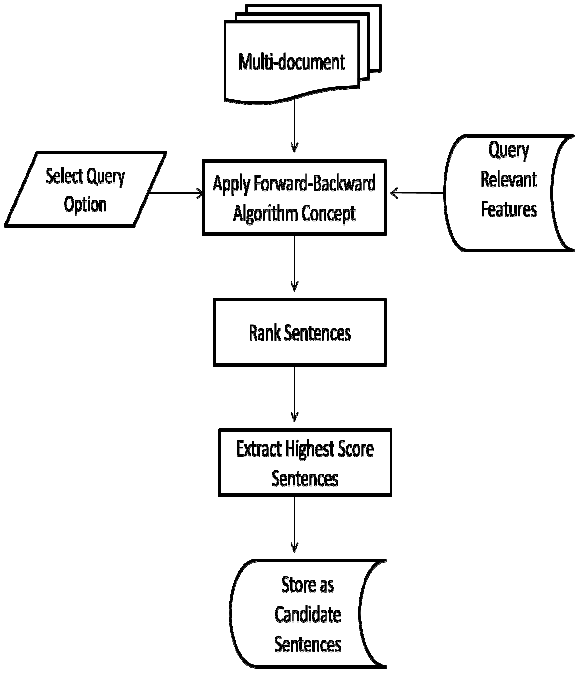

Figure 4. Extraction

In Extraction Phase, the system input is multi-document of the same topic. When the user queries one or more options, the system use Forward-Backward algorithm concept to retrieve the important sentences from the multi-documents according to the user query options and query relevant features . If the user query option is , the system search query relevant features of time ( ) form the multi-document and find the information related to the time. The system selects all the sentences which contain information related to time. After that, the system ranks the sentences with the highest score and store the highest score sentences as candidate sentences.

4.5. Abstraction Phase

In Abstraction Phase, the system use longest matching approach to reduce redundant information. The system also needs to develop some abstraction methods such as substitution, averaging, comparison methods to get the condensation.

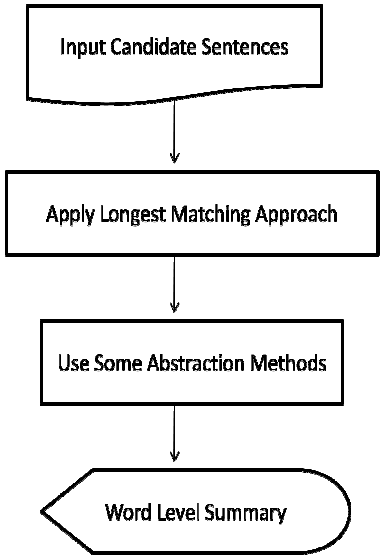

Figure 5. Abstraction

Finally, the proposed system displays the accurate word level summary to the user that is relevant to the query options.

5. Conclusion

This paper proposed query-focused document summarization on Earthquake news written in Myanmar Language. Myanmar Language structure is complex so new methodologies may have to be added to get the accurate word level summary. In this system, both extraction and abstraction will be applied to get the better result. This system can give the user to a word level summary about Earthquake news written in Myanmar Language at a short period of time. The result of sample word level summary according to the user query options is as shown in Figure.

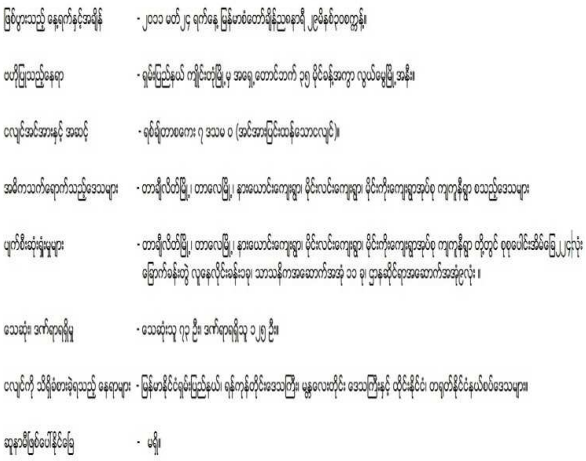
Figure 6. Sample Word Level Summary

References

[1] Aijun Xu, “Query-focused Document Summarization Using Concept Analysis”, Dec 2011.
[2] Chao Shen, Tao Li, Multi-Document Summarization via the Minimum Dominating Set”, 2010.
[3] Hla Hla Htay, Kavi Narayana Murthy, “Myanmar Word Segmentation using Syllable level Longest Matching”, The 6th Workshop on Asian Languae Resources, 2008.
[4] Jayabharathy, Kanmani and Buvana, “An Analytical Framework for Multi-Document Summarization”, May 2011.
[5] May Thu Win Moet Moet Win Moh Moh Than, “Burmese Phrase Segmentation”, Conference on Human Language Technology for Development, Alexandria, Egypt, May 2011.
[6] Pann Yu Mon, Yoshiki Mikami, “Myanmar Language Search Engine”, International Journal of Computer Science Issues (IJCSI), March 2011.
[7] Pinaki Bhaskar, Sivaji Bandyopadhyay, “A Query Focused Multi Document Automatic Summarization”, 2009.
[8] Prashant D. Joshi, M. S. Bewoor, S. H. Patil, “System for Document Summarization Using Graphs in Text Mining”, International Journal of Advances in Engineering & Technology, Sept 2011.
[9] R.Kowsalya, R.Priya, P.Nithiya, “Multi Document Extractive Summarization Based On Word Sequences”, IJCSI International Journal of Computer Science Issues, Vol. 8, Issue 2, March 2011.
[10] Sun Park, “User-focused Automatic Document Summarization using Nonnegative Matrix Factorization and Pseudo Relevance Feedback”, 2011.
[11] Vishal Gupta, Gurpreet Singh Lehal, “A Survey of Text Summarization Extractive Techniques”, Aug 2010.
[12] Wen-tau Yih Joshua Goodman Lucy Vanderwende Hisami Suzuki, “Multi-Document Summarization by Maximizing Informative Content-Words”, 2007.
[13] Xiaojun Wan and Jianwu Yang, “Multi-Document Summarization Using Cluster-Based Link Analysis”,2008.
[14] Yanting LI,Kai CHENG, “Key Sentence Extraction from Single Document based on Triangle Analysis in Dependency Graph”.
[15]http://www.gabormelli.com/RKB/Text_Summarization_Algorithm
[16] http://www.healburma.org/mm/news-in-burmese/1293-earthquake-news-march-25-2011.html
[17] http://myanmarwords.pikay.org/
[18] http://www.ir-facility.org/scoring-and-ranking-techniques-tf-idf-term-weighting-and-cosine-similarity
[19] http://nlp.hivefire.com/articles/share/30940/