Automatic Myanmar Text Summarization System With Semantic Roles

May Thu Naing, Aye Thida
University of Computer Studies, Mandalay
maythu.tgi@gmail.com, ayethida.royal@gmail.com

Abstract

Automatic text summarization is to compress the larger original text into shorter text called as summary. It is the essential researches for Natural Language Processing (NLP). Automatic semantic analysis of natural language text has received much attention by NLP community. In proposed Myanmar text summarization system, automatic summarized text will produce as mail dairy notes for incoming mail system. Anaphora resolution is a key problem in natural language processing, and has correspondingly received a significant amount of attention in the literature. So, automatic pronominal anaphora resolution in Myanmar texts will be used for summary generation system in this paper. This system will identify the semantic relationships, or semantic roles, and fill by constituents of a sentence within a semantic frame using Myanmar Verb Frames. So, verb frame resource with annotated semantic roles will be developed for proposed system.

Keywords: automatic text summarization, natural language processing, semantic roles.

1. Introduction

In the growth of the world-wide web and online text collections, it has become increasingly important to provide improved mechanism for finding information quickly [7]. Presenting the user with summaries of the matching sentences in documents can help the user identify which documents are most relevant to the user's needs.

Text summarization refers to the task of shortening a long text. It can be either be a generic summary, which gives an overall sense. There are two major directions in text summarization: the extractive and the abstractive paradigm [11]. The first approach in creating summaries (most common) is based on identifying important words in texts by using their frequencies, and determining those sentences that contain a bigger number of important words. These sentences are extracted from the original text, and taken to constitute the summary. In this paradigm, the summarization is performed through sentence extraction: the summary is a subset of the sentences in the original text.

An alternative approach is to build a summary consisting of sentences that don't necessarily have to show up in that specific form in the source text. This method can also be applied in the case of very large texts, such as a whole novel, where neither the determination of most significant sentences based on occurrences of frequent words, nor building discourse structures could be of help. In these cases, other methods, mainly expanding a collection of predefined flexible summary could be applied.

In addition to extracts and abstracts, summaries may differ in several other ways. Some of the major types of summary that have been identified include indicative (keywords indicating topics) vs. informative (content-laden); generic (author's perspective) vs. query-oriented (user-specific); background vs. just-the-news; single-document vs. multi-document; neutral vs. evaluative [5]. This makes the study of automated text summarization an exciting area in which to work.

The proposed system is to summary building uses the first method, sentence extraction. However, the novelty of our approach consists in basing the extraction of different sentences from the original text on semantic role analysis, an association which is not yet explored at its full potential. The method provided that named entity and semantic roles extraction modules are available.

The remaining parts of the paper are organized as follows: the natural language processing systems

using semantic roles are presented in section 2, Myanmar verb frame resource is introduced in section 3, section 4 describes the overall of the proposed system and section 5 concludes the paper and identifies future work.

2. Related Works

The natural language processing community has recently experienced a growth of interest in semantic roles, since they describe WHO did WHAT to WHOM, WHEN, WHERE, WHY, HOW etc. for a given situation, and contribute to the construction of meaning. Thus, lexical-semantic resources (LSRs) are the foundation of many NLP tasks such as word sense disambiguation, semantic role labeling, question-answering and information extraction. They are needed on a large scale in different languages.

They [13] were the first to stress the importance of semantic roles in answering complex questions. Their system identified predicate argument structures by merging semantic role information from PropBank and FrameNet. Expected answers are extracted by performing probabilistic inference over the predicate argument structures in conjunction with a domain specific topic model.

They [4] introduce a general framework for answer extraction which exploits semantic role annotations in the FrameNet semantic resource.

There [1] was among the first to investigate the usefulness of semantic roles in summaries. More recently, [Suanmali et al., 2010] used semantic roles and WordNet to compute the semantic similarity between two sentences in order to decide if the sentences are to be kept or not in the summary.

3. Semantic Role Resources

One of the goals of the FrameNet project is to design a linguistic ontology that can be used for the automatic processing of semantic information. The associated hierarchy contains an extensive semantic analysis of verbs, nouns, adjectives and situations in which they are used, called frames. The basic assumption on which the frames are built is that each word evokes a particular situation with specific participants [3]. The word that evokes a particular frame is called target word or predicate and can be an adjective, noun or verb. The participant entities are defined using semantic roles and they are called frame elements. The Berkeley FrameNet database consists of more than 7000 English lexical items, together with example sentences annotated with semantic roles [2].

PropBank is a bank of propositions. A “proposition” is on the basic structure of a sentence [3], and is a set of relationships between nouns and verbs, without tense, negation, aspect and modal modifiers. Arguments which belong to propositions are annotated by PropBank with numbered role labels (Arg0 to Arg5) and modifiers are annotated with specific ArgM (Argument Modifiers) role labels. Each verb occurrence in the corpus receives also a sense number, which corresponds to a roleset in the frame file of such verb. A frame file may present several rolesets, depending on how many senses the verb may assume. In the roleset, the numbered arguments are “translated” into verb specific role descriptions. Arg0 of the verb “sell”, for example, is described as “seller”. Thus, human annotators may easily identify the arguments and assign them the appropriate role label. There is currently no frame semantic representation of Myanmar.

Myanmar verb frame files will be built together with example sentences annotated with semantic roles following PropBank guidelines. But, this system could not reproduce the same experience of PropBank. This proposed system interested in designing Myanmar Verb Frame files in relatively independent modules to facilitate the collaborative construction of this resource. Once PropBank guidelines and PropBank frames files are available for consultation, it is design to adopt a different approach: instead of firstly building frames files and Annotator´s Guidelines, Myanmar Verb Frame is start by annotating a corpus using English frames files and guidelines as model. Therefore, unlike PropBank, in this first phase it annotated only semantic role labels and not verb senses. In this way, the proposed system identified language-specific aspects of SRL for Myanmar language, and generated a corpus that used as base to build frames files.

The overall pipeline architecture of the proposed system is presented in Figure 1.

4. Overview of the Proposed System

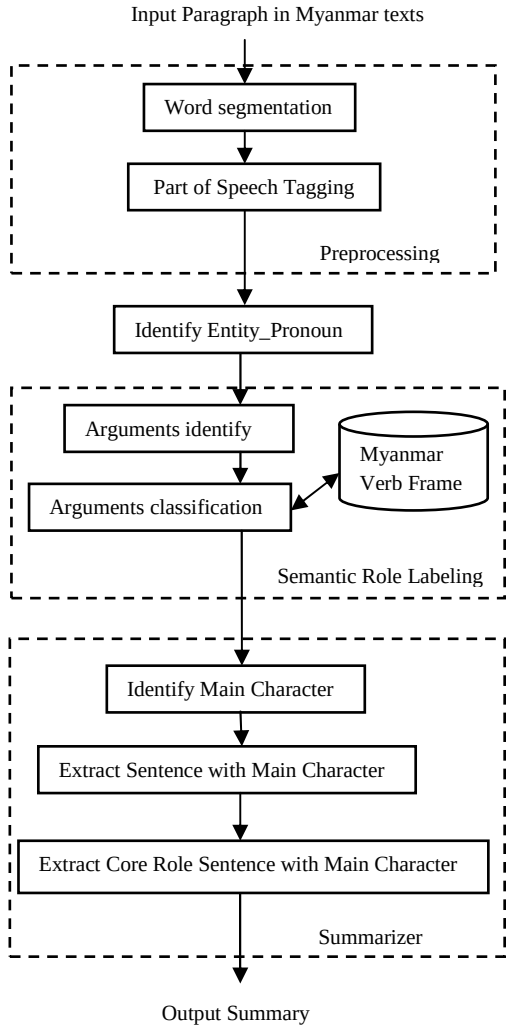

Figure 1. Overview of the Automatic Myanmar Text Summarization System with Semantic Roles

4.1. Sample Input of the Automatic Myanmar Text Summarization System

The 27th Sea game was hosted by Myanmar in 2013. It is held by the union of 11 Asian countries every 2 year. Athletes from 11 Asian countries participate in this Sea game. There are many different items in this sea game. The winners are awarded gold as 1st prize, silver as 2nd prize and bronze as 3rd prize in different competitions. Myanmar athletes who got identified prizes are given as bonus from organizations of Myanmar. The 28th Sea game will be held by Singapore in 2015. The sea game will be hosted again by Myanmar in the next 20 years.

မြန်မာနိုင်ငံသည် ၂၇ကြိမ်မြောက်ဆီးဂိမ်းအားကစားပွဲတော်ကြီးကိုအိမ်ရှင်အနေဖြင့်လက်ခံကျင်းပခဲ့သည်။ ထိုပွဲတော်ကို အာဆီယံနိုင်ငံများစုပေါင်းပြီး
၂နှစ်တစ်ကြိမ်ကျင်းပသည်။ထိုအားကစားပွဲတော်ကြီးတွင်အာဆီယံ၁၁နိုင်ငံမှအားကစားသမားများ
လာရောက်ယှဉ်ပြိုင်ကြသည်။ထိုပွဲတော်တွင်အားကစားပြိုင်ပွဲပေါင်းများစွာပါဝင်သည်။ပထမဆုအဖြစ်ရွှေဒုတိယဆုအဖြစ်ငွေ၊တတိယဆုအဖြစ်ကြေးတို့အနိုင်ရသူများအားဆုပေးသည်။ထိုအပြင်မြန်မာနိုင်ငံအနေဖြင့်ဆုရသောမြန်မာအားကစားသမားများအတွက်အပိုဆုကြေးများထပ်မံချီးမြှင့်ခဲ့သည်။၂၈ကြိမ်မြောက်ဆီးဂိမ်းအားကစားပွဲတော်ကြီးကို၂၀၁၅ခုနှစ်တွင်စင်ကာပူနိုင်ငံ၌ကျင်းပမည်ဖြစ်သည်။မြန်မာနိုင်ငံအနေဖြင့်နောက်နှစ်ပေါင်း၂၀အကြာတွင်ဆီးဂိမ်းအားကစားပွဲတော်ကိုပြန်လည်ကျင်းပခွင့်ရရှိပါမည်။

.

4.2. Preprocessing

For preprocessing of proposed system in Figure 1, word segmentation is the first stage. Without a word segmentation solution, no NLP application (such as Part-of-Speech (POS) tagging and translation) can be developed. Words can be combined to form phrases, clauses and sentences. Thus, in proposed system, word segmentation is performed with Myanmar Word Segmenter [12]. This segmenter generated all possible segmented sentence of phase pattern as their score for further processing. It can handle the unkown case of word and it does not only depend on lexicon.
Example of word segmentation for one sentence in the input paragraph,

ထို ပွဲတော် တွင် အားကစား ပြိုင်ပွဲ များစွာ ပါဝင်သည်

For the next step of the preprocessing stage which is Part of Speech (POS) Tagging, rule based POS tagging of Myanmar language. [12] is used. This tagging used the context-free grammer (CFG) as rules which parsing is start with sentence and left to right parsing structure to define the POS of each word.
Example of POS tagging for above word segmentated sentence,

```
0:[S]
  0:[NP_PLACE]
    0:[PRON_PPN]
        0:ထိုပွဲတော်
  0:[PREP_PLACE]
        0:တွင်
  0:[NP]
    0:[N]
        0:အားကစားပြိုင်ပွဲများစွာ
  0:[VEND]
    0:[V]
        0:ပါဝင်သည်
```

4.3. Identify Entity_Pronoun

In order to identify the semantic role a specific entity express, the entity must be first identified in the text. This is the task of named entity recognition (NER). As a preprocessing module for our summary generation system in Figure 1, this system will be used a Named Entity Recognition [14] system for Myanmar language. This NER system will identify the boundary of candidate NE that composes of many words by using NER algorithm with Rules and Tag Set files. It displays NEs and correctly classifies the class of these NEs. Therefore, after identifying named entities and their types (person, organization, place, etc.), a simple anaphora resolution method, based on a set of reference rules, is applied to input text, in order to link all entities to their pronoun.

In this paper, the anaphoric system is a basic rule-based one, focusing on named entity anaphoric relations. To solve anaphoric references for Myanmar language, a rule-based system create an anaphoric link between the pronoun and its antecedent based on Hobbs algorithm [6]. Hobbs algorithm is adapted successfully for those languages (eg, Chinese) which have similar Subject-Verb-Object (SVO) structure and follow a fixed word order. But the structure of Myanmar language is Subject-Object-Verb (SOV) structure and free word order. So, Hobbs algorithm does not adapt for Myanmar language. Thus, a new syntax_based approach will be developed to solve anaphoric resolution for pronominal reference for Myanmar texts based on Hobbs algorithm.

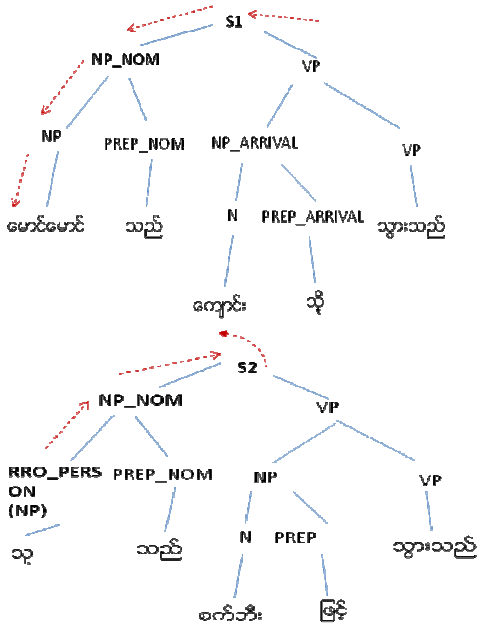

Figure 2. Example of Identify Entity_Pronoun in Myanmar Sentence based on Hobbs Algorithm.

1. Start search at PRO_PERSON (NP) in S2.
2. NP_NOM is first NP we encounter. Call this node X and call the path used to reach it p.
3. Traverse all branches below node X to the left of path p in left to right, breadth-first search. No NP node between it and X.
4. From node X, go up the tree. S2 is first node we encounter. Call this new node X, and call the path traversed to reach it p.
5. If X is highest S node in the sentence, traverse trees of previous sentence S1 in left to right. When NP node မောင်မောင် is encountered, proposed as antecedent of သူ.
6. Result: မောင်မောင် for သူ.

The following is the named entity anaphoric relations from the inputted paragraphs.

မြန်မာနိုင်ငံသည် ၂၇ကြိမ်မြောက် ***ဆီးဂိမ်းအားကစားပွဲတော်ကြီး***ကို အိမ်ရှင်အနေဖြင့်လက်ခံကျင်းပခဲ့သည်။***ထိုပွဲတော်*** (***ဆီးဂိမ်းအားကစားပွဲတော်ကြီး***)ကို အာဆီယံနိုင်ငံများစုပေါင်းပြီး ၂နှစ်တစ်ကြိမ်ကျင်းပသည်။***ထိုအားကစားပွဲတော်ကြီး(ဆီးဂိမ်းအားကစားပွဲတော်ကြီး)***တွင်အာဆီယံ၁၁နိုင်ငံမှအားကစားသမားများလာရောက်ယှဉ်ပြိုင်ကြသည်။***ထိုပွဲတော်(ဆီးဂိမ်းအားကစားပွဲတော်ကြီး)တွင်***အားကစားပြိုင်ပွဲပေါင်းများစွာပါဝင်သည်။ပထမဆုအဖြစ်ရွေးခုတိယဆုအဖြစ် ငွေ၊တတိယဆုအဖြစ်ကြေးစသဖြင့်ပြိုင်ပွဲတွင်အနိုင်ရသူများကိုဆုပေးသည်။ထိုအပြင်မြန်မာနိုင်ငံအနေ ဖြင့်ဆုရသောမြန်မာအားကစားသမားများအတွက် အပိုဆုကြေးများ

ပ်မံချီးမြှင့်ခဲ့သည်။၂၈ ကြိမ်မြောက်
ဆီးဂိမ်းအားကစားပွဲတော်ကို၂၀၁၅ခုနှစ်တွင်စင်ကာပူနိုင်ငံတွင်ကျင်းပ
မည်။မြန်မာနိုင်ငံအနေဖြင့်နောက်နှစ်ပေါင်း၂၀အကြာတွင်ဆီးဂိမ်းအား
ကစားပွဲတော်ကို ပြန်လည်ကျင်းပခွင့် ရရှိပါမည်။

4.4. Semantic Role Labeling Model

The third step in the proposed system is semantic role labeling. Firstly, in arguments identify stage, lemma of verb in sample input sentences is identified as predicate.

ကျင်းပ(held)□ ယှဉ်ပြိုင်(race)၊ပါဝင်(include)၊

ဆုပေး(award)□ချီးမြှင့်၊ဖြစ်(be)၊ရရှိ(get)

For role semantics to become relevant for language technology, robust and accurate methods for automatic semantic role labeling (SRL) are needed. Automatic Labeling of Semantic Roles, arguments for a predicate (verb) within a sentence and tag them with appropriate semantic roles. Role assignment has generally been modeled as a classification task.

Secondly, for semantic role assignment or argument classification stage, support vector machine (SVM) classifier used to assign semantically meaningful labels to arguments corresponding to predicates. SVM classifier takes a set of features for syntactic structure in the input sentence as input. These features are syntactic information (path from predicate to constituent, phrasal type of constituent) and lexical information (head word of the constituent, predicate). SVM determines proper labels for semantic arguments of a given predicate using Myanmar Verb Frame resource and annotated training corpus.

The following is Myanmar verb frame files based on PropBank for sample input predicates.

Frame file for ကျင်းပ(hold)

Roles:

Arg0-holder
Arg1-thing held
Arg2-with whom

မြန်မာနိုင်ငံသည်**(Arg-0)**၂၇ကြိမ်မြောက်
ဆီးဂိမ်းအားကစားပွဲတော်ကြီးကို**(Arg-1)**
အိမ်ရှင်အနေဖြင့်လက်ခံကျင်းပခဲ့သည်။ထိုပွဲတော်
(ဆီးဂိမ်းအားကစားပွဲတော်ကြီး)ကို**(Arg-1)**
အာဆီယံနိုင်ငံများ**(Arg-0)**စုပေါင်းပြီး၂နှစ် တစ်ကြိမ်
ကျင်းပသည်။

Frame file for ယှဉ်ပြိုင်(compete)

Roles:

Arg0-competitor
Arg1-opponent
Arg2-prize

ထိုအားကစားပွဲတော်ကြီး(ဆီးဂိမ်းအားကစားပွဲတော်ကြီး)
တွင်(**ArgM-place**)အာဆီယံဒေသနိုင်ငံမှ
အားကစားသမားများ**(Arg-0)**လာရောက်
ယှဉ်ပြိုင်ကြသည်။

Frame file for ပါဝင်(include)

Roles:

Arg0-agent, entity causing some grouping
Arg1-thing being included
Arg2-group

ထိုပွဲတော်(ဆီးဂိမ်းအားကစားပွဲတော်ကြီး)တွင်**(Arg2)**အ
ားကစားပြိုင်ပွဲပေါင်းများစွာ**(Arg1)**ပါဝင်သည်။

Frame file for ဆုပေး(award)

Roles:

Arg0-giver
Arg1-award
Arg2-benefactive

ပထမဆုအဖြစ်ရွှေ၊ဒုတိယဆုအဖြစ်ငွေ၊တတိယဆုအဖြစ်
ကြေးတို့ကို**(Arg1)**အနိုင်ရသူများအား **(Arg2)**
ဆုပေးသည်။

Frame file for ချီးမြှင့်(confer)

Roles:

Arg0-giver
Arg1-gift
Arg2-given to

ထိုအပြင်မြန်မာနိုင်ငံအနေဖြင့်**(Arg0)**ဆုရသော
မြန်မာအားကစားသမားများကို**(Arg2)**အပိုဆုကြေးများ(
Arg1) ထပ်မံချီးမြှင့်ခဲ့သည်။
၂၈ကြိမ်မြောက်ဆီးဂိမ်းအားကစားပွဲတော်ကို**(Arg1)**
၂၀၁၅ခုနှစ်တွင်(**ArgM-time**) စင်ကာပူနိုင်ငံ၌**(Arg0)**
ကျင်းပမည်။

Frame file for ရရှိ(get)

Roles:

Arg0- receiver
Arg1- thing gotten
Arg2- giver
Arg3- price paid, in-exchange-for
Arg4- benefactive

မြန်မာနိုင်ငံသည်**(Arg0)**နောက်နှစ်ပေါင်း ၂၀အကြာတွင်(**ArgM-time)**ဆီးဂိမ်းအားကစားပွဲတော်ကို **(Arg1)** ပြန်လည်ကျင်းပခွင့် **(ArgM)** ရရှိပါမည်။

4.5. Summarizer

The fourth step, summarizer, in Figure 1 consists of three components.

The first stage of the summarizer is to identify main character in input paragraph. The main character is considered to be the named entity having the higher number of occurrences in the text.

Table 1. Frequency of characters in sample input paragraph

Character	Frequency
မြန်မာနိုင်ငံ	3
ဆီးဂိမ်းအားကစားပွဲတော်	6
အိမ်ရှင်နိုင်ငံ	2
အာဆီယံနိုင်ငံ	2
မြန်မာအားကစားသမားများ	1
စင်ကာပူနိုင်ငံ	1

In this table, ဆီးဂိမ်းအားကစားပွဲတော် (sea game) has maximum frequency. Thus, this system chose ဆီးဂိမ်းအားကစားပွဲတော် (sea game) as main character.The second stage is to extract sentences with main character.

မြန်မာနိုင်ငံတွင် ၂၇ကြိမ်မြောက်ဆီးဂိမ်းအားကစားပွဲတော်ကြီးကိုအိမ်ရှင်နိုင်ငံအဖြစ်လက်ခံကျင်းပခဲ့သည်။
ထိုပွဲတော်မှာအာဆီယံနိုင်ငံများစုပေါင်းပြီး ၂နှစ်တစ်ကြိမ်ကျင်းပသည်။
ထိုအားကစားပွဲတော်ကြီးတွင်အာဆီယံ၁၁နိုင်ငံမှအားကစားသမားများလာရောက်ယှဉ်ပြိုင်ကြသည်။ထိုပွဲတော်တွင်အားကစားပြိုင်ပွဲပေါင်းများစွာပါဝင်သည်။၂၈ကြိမ်မြောက်ဆီးဂိမ်းအားကစားပွဲတော်ကို ၂၀၁၅ခုနှစ်တွင်စင်ကာပူနိုင်ငံတွ
င်ကျင်းပလိမ့်မည်။မြန်မာနိုင်ငံအနေဖြင့်နောက်နှစ်ပေါင်း ၂၀အကြာတွင်ဆီးဂိမ်းအားကစားပွဲတော်ကိုအိမ်ရှင်နိုင်ငံအဖြစ် ပြန်လည် ကျင်းပခွင့် ရရှိပါမည်။

The third stage is to extract obligatory arguments that are Arg0_Arg4 semantic roles sentences with specific character.

မြန်မာနိုင်ငံသည် ၂၇ကြိမ်မြောက်ဆီးဂိမ်းအားကစားပွဲတော်ကြီးကိုအိမ်ရှင်နိုင်ငံအဖြစ်လက်ခံကျင်းပခဲ့သည်။၂၈ကြိမ်မြောက်ဆီးဂိမ်းအားကစားပွဲတော်ကို ၂၀၁၅ခုနှစ်တွင်စင်ကာပူနိုင်ငံတွင်ကျင်းပလိမ့်မည်။မြန်မာနိုင်ငံအနေဖြင့်နောက်နှစ်ပေါင်း ၂၀အကြာတွင်ဆီးဂိမ်းအားကစားပွဲတော်ကိုအိမ်ရှင်နိုင်ငံအဖြစ် ပြန်လည် ကျင်းပခွင့် ရရှိပါမည်။

5. Conclusion and Future Work

In this paper, a summary generation system based on semantic roles is presented. The main work includes word segmentation, POS tagging, identifying entity_pronoun, and semantic role labeling and summary generation. For resolving entities, syntax-based method is introduced to create anaphoric link for Myanmar language. In SRL, arguments identification and arguments classification are developed using Myanmar Verb Frame Resource. The main components of the summary generation are to identifying named entities, marking semantic roles, and selecting the sentences of the text to be kept in the summary. As the future work, the system needs to implement as automatic text summarization system via mobile phone.

References

[1] Christopherson and Steven L, “Effects of knowledge of semantic roles on summarizing written prose”, Contemporary Educational Psychology, Vol 6(1), Jan 1981, 59-65.

[2] C. F. Baker, C. J. Fillmore and J.B. Lowe, “The Berkeley FrameNet project”, In Proceedings of the COLING-ACL, Montreal, Canada, 1998.

[3] C.J. Fillmore, “The case for case”, In Universals in Linguistic Theory, 1968.

[4] D. Shen, M. Lapata, “Using Semantic Roles to Improve Question Answering”, In Proceedings of

the Conference on Empirical Methods in Natural Language Processing (EMNLP '07), 2007.

[5] Eduard Hovy and Chin-Yew Lin, “Automated Text Summarization and The Summarist System”, Information Sciences Institute of the University of Southern California.

[6] J.R. Hobbs, “Resolving Pronoun References”, Dept. of Computer Science, City College, New York, Augest 1977.

[7] J. Goldstein,, V. Mittal, J. Carbonell and M. Kantrowitz “Multi-Document Summarization By Sentence Extraction”.

[8] L. Suanmali, N. Salim and M.S. Binwahlan, “SRL-GSM: A Hybrid Approach based on Semantic Role Labeling and General Statistic Method for Text Summarization”, Journal of Applied Sciences 10(3) : 166-173.

[9] M. Palmer, D. Gildea and P. Kingsbury, “The Proposition Bank: An Annotated Corpus of SemanticRoles”, Computational Linguistics, March, 2005, pp. 71-105.

[10] M. Lluis, X. Carreras, K. C. Litkowski, and S. Stevenson, “Semantic role labeling: An introduction to the Special Issue”, Computational Linguistics, 34(2):145-159, 2008.

[11] Mani Inderjeet, “Automatic Summarization”, John Benjamins Pub Co; 2001.

[12] S.L. Phyu, “Development of Lexico-Conceptual Knowledge Resources and Syntax Analyzer for Myanmar Language”, Ph.D Thesis, University of Computer Studies, Mandalay, Myanmar, 2013.

[13] S. Narayanan, S. Harabagiu, “Question answering based on semantic structures”, In Proceedings of the 19th COLING, 2004,pp. 184–191.

[14] Thida Myint, “Machine Transliteration Methods for Myanmar Names”, University of Computer Studies, Mandalay, Myanmar, 2013.