Template-Driven Automatic Myanmar Text Summarization using Conditional Random Fields

Win Thuzar Kyaw, Ni Lar Thein, Hla Hla Htay
University of Computer Studies, Yangon
winthuzarkyaw19@gmail.com, nilarthein@gmail.com,hlahlahtay123@gmail.com

Abstract

By providing information as the summary, the reader can save time and can easily absorb the main concepts of the articles which are described in digital form. Therefore, automatic text summarization, the process of compressing the documents into the compact style by means of a computer, plays an important role. In this paper, Template-Driven Automatic Myanmar Text Summarization using Conditional Random Fields (CRFs) is introduced for Myanmar news articles in natural disaster domain collected from official Myanmar Newspaper. CRFs are undirected graphical models which can be used to segment and label natural language text. CRF model is mainly used for information extraction in this work.

Keywords: automatic text summarization, conditional random fields (CRFs)

1. Introduction

Human beings need to spend much time to read and manually capture the crucial parts of the documents because of information overloading in digital age. Thus, the computerized process of distilling the source text and converting into the summary that includes the main content and meaning of the text, automatic text summarization, is essential and being a hot topic to date.

Text summarization system can be distinguished depending on the input, purpose and output factors [Sparck, 1999]. The source of the input may be from a single document or from multiple documents. Relating to the purpose, two types of summaries are discovered based on the writer's opinion (generic summaries) or reader's concern (query-based summaries). Extractive summaries generated as the output from the summarization process contain the main segments of the source text by reusing the same words of the original document. On the other hand, abstractive summaries are created in a new style using additional composition. Moreover, summaries can be described in an indicative way which tells the topic of the content from which the reader can make decision for deep reading of the content whereas summaries in informative way includes all important parts in abridged version of the text.

Although there are text summarizers developed in other languages such as Thai [Jaruskulchai et al., 2003], Korean [Kim et al., 2000], Chinese [Hu et al., 2004], summarizers in Myanmar language cannot be available till now.

The remaining parts of the paper are organized as follows: the work concerning CRFs and template-driven automatic text summarization systems are presented in section 2, the definition of CRFs 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 Work

CRFs have been applied in a variety of Natural Language Processing areas. [Peng and McCallum, 2004] employed CRFs for information extraction to draw out various common fields from the headers and citation of research papers. They evaluated the performance on a standard benchmark data set and found that the error rate of their proposed system is less than that of SVM. In addition, accuracy is more pleasing compared to HMMs. Keywords extraction is used in text mining applications. [Zhang et al., 2008] implemented the automatic keywords extraction by applying CRF model and they discovered that the CRF model got better result in keyword extraction than other machine learning methods such as support vector machine, multiple linear regression model etc. In the field of Part of Speech (POS) tagging, [Patel and Gali, 2008] proposed a POS tagger and chunker based on CRF for Hindi. They measured the accuracy of POS tagger and the performance of the chunker with evaluation script from conll 2000. And the next application of CRF can be seen in Lao word segmentation [Vanthanavong et al., 2011] in which word segmentation task can be assumed as a sequential labeling task. They judged the result of word segmentation as well as name entities segmentation and found out more beneficial result than dictionary-based approach. In the automatic document summarization aspect, a generic single-document sentence extraction framework using CRFs [Shen et al., 2007] was presented. In this paper, they took the summarization task as a sequence labeling problem and labeled the sentences by 1 (the summary sentences) and 0 (non-summary sentences). They evaluated their approach in F1 and ROUGE-2 and their results can improve the performance over the supervised baseline and unsupervised baseline.

Some of the template-driven text summarization systems are SUMMONS [Radev and McKeown, 1998] directed the possibility of fusing information extraction with natural language generation for a summarization system. A very alike system to SUMMONS was RIPTIDES [White et al., 2001]. But it tried to summarize larger document sets and also tried to solve the problem of SUMMONS of the comparing reported numbers of varying specificity by using rules. GISTexter [Harabagiu and Lacatusu, 2002] produced both extracts and abstracts for single and multiple documents with the use of the templates outputted from CICERO IE system.

3. Conditional Random Fields

A Conditional Random Field (CRF), a variant of Markov Random Network can be viewed as an undirected graphical model. It combines classification and graphical modeling for segmenting and labeling sequential data. Therefore, it has been widely used in many natural language processing tasks. Because CRF is simply a conditional probability distribution, it is able to solve the problem of complex independencies, the main difficulty of Hidden Markov Models (HMMs) that define the joint probability distribution. Moreover, CRF avoids label bias problem which is a restriction of Maximum Entropy Markov Models (MEMMs). Let $D\ (D_1, D_2, \ldots, D_n)$ be the observation sequential data and $L\ (L_1, L_2, \ldots, L_n)$ be the labels. A linear chain CRF can be defined as follows:

$$p(L|D) = \frac{1}{Z_D} exp\left(\sum_i \sum_j \lambda_j f_j(l_{i-1}, l_i, D, i)\right)$$

in which Z_D is a normalization factor which can be defined as

$$Z_D = \sum_L exp\left(\sum_i \sum_j \lambda_j f_j(l_{i-1}, l_i, D, i)\right)$$

and $f_j(l_{i-1}, l_i, D, i)$ is a feature function and λ_j is the weight for feature f_j.

4. Overview of the proposed system

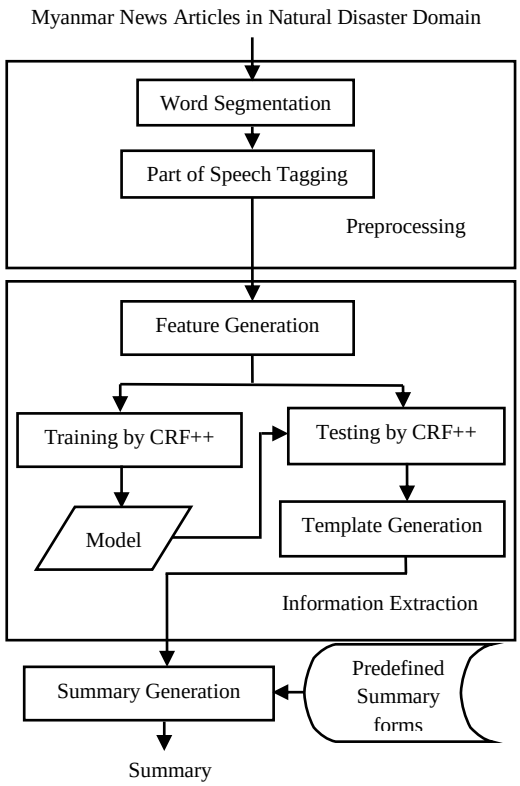

Figure: Overview of the Automatic Myanmar Text Summarization System

Table 1. Sample Input of the Automatic Myanmar Text Summarization System

Input Landslide News Article

Deadly landslide strikes at Papua New Guinea, 60 killed
PORT MORESBY, 25 January
A deadly landslide struck in Southern islands of Papua New Guinea on 24 January killed about 60 residents, authorities said on 25 January.
The disaster caused near Tari in a Southern mountainous region around 7:00 am on 24 January.
That place situates on the north-west of Papua New Guinea and it is also an area of liquefied natural gas fields.
Papua New Guinea officials arrived at the scene and set up relief operations.

ပါပူဝါနယူးဂီနီ၌ မြေပြိုမှုဖစ်ပွား၊ ၆၀ သေဆုံး
ပို့မိုရက်ဘိုင် ဇန်နဝါရီ ၂၅
ပါပူဝါနယူးဂီနီနိုင်ငံတောင်ပိုင်း ကုန်းမငြိဒေသတွင် ဇန်နဝါရီ ၂၄ ရက်က မြေပြိုမှုဖစ်ပွားရာ ဒေသခံပြည်သူ ၆၀ ခန့် သေဆုံးခဲ့ကောင်း ပါပူဝါနယူးဂီနီအရာရှိများက ဇန်နဝါရီ ၂၅ ရက်တွင် ပေဂြကြားသည်။
ဖစ်ပွားခဲ့
ဖစ်ပွားမှုမှာ တောင်ပိုင်းကုန်းမငြိဒေသရှိ တာရီမိုဒ်တွင် ဇန်နဝါရီ ၂၄ ရက် နံနက် ၇ နာရီခန့်ကဖစ်ပွားခဲ့ခြင်းဖစ်သည်။
ထိုဒေသရောက်ရှိနေသော ပါပူဝါနယူးဂီနီမိုဒ်တော်၏ အနောက်မြောက်ပိုင်းတွင် တည်ရှိပြီး ရေနံနှင့်သဘာဝဓာတ်ငွေ့ကုမ္ပဏီတစ်ခုက စီမံကိန်းတစ်ရပ်အကောင်အထည်ဖော်နေသော ဒေသတစ်ခုလည်း ဖစ်သည်။
ကယ်ဆယ်ရေးအဖွဲ့များက မြေပြိုမှု ဖစ်ပွားသည့် နေရာသို့ သွားရောက်ကာ ကယ်ဆယ်ရေး လုပ်ငန်းများ ဆောင်ရွက်ပေးလျက်ရှိသည်။

News articles for two years about natural disasters are collected from Myanma Ah Lin Myanmar Newspaper for training and testing data.

Fortunately, the type of natural disaster can be determined in the heading of the news article. Thus, we can easily search what kind of disaster is only by keyword matching with the headline.

4.1. Word Segmentation

Although English language clearly delimits the words by spaces, Myanmar language similar to other Asian languages lacks spaces between words. Therefore, word segmentation is an important preprocessing task for natural language processing applications such as machine translation, information retrieval and text categorization and so on. For the automatic Myanmar text summarization, the first preprocessing stage, word segmentation is performed with the use of Myanmar Word Segmenter [Pa, 2009].In her research, a combination approach of word juncture model and bigram model is used in which the former model reflects the affinity of a pair of known words in forming another words, especially an unknown word and the latter model calculate the probability of forming two words together. The performance is evaluated with precision, recall and F-measure. If the unknown words are zero, the precision is 99% and if the unknown words increase to 2000, precision decreases to 71.5%.

The following is an example of word segmentation of a sentence from the inputted news article.

Sample input sentence

Sentence in English:

A deadly landslide struck in Southern islands of Papua New Guinea on 24 January killed about 60 residents, authorities said on 25 January.

Sentence in Myanmar:

ပါပူဝါနယူးဂီနီနိုင်ငံတောင်ပိုင်း ကုန်းမငြိဒေသတွင် ဇန်နဝါရီ ၂၄ ရက်က မေပြိုမှုဖစ်ပွားရာ ဒေသခံပညိသူ ၆၀ ခန့် သေဆုံးခဲ့ကေငြိး ပါပူဝါနယူးဂီနီအရာရှိများက ဇန်နဝါရီ ၂၅ ရက်တွင် ပေဂြကြားသည်။

After word segmentation phase,

ပါပူဝါနယူးဂီနီ_နိုင်ငံ_တောင်ပိုင်း_ကုန်းမငြိ_ဒေသ_တွင်_ဇန်နဝါရီ_၂၄_ရက်_က_မေပြိုမှု_ဖစ်ပွား_ရာ_ဒေသခံ_ပညိသူ_၆၀_ခန့်_သေဆုံးခဲ့ကေငြိး_ပါပူဝါနယူးဂီနီ_အရာရှိများ_က_ဇန်နဝါရီ_၂၅_ရက်_တွင်_ပေဂြကြားသည်_။__

Each word is separated by an underscore.

4.2. Parts of Speech (POS) Tagging

For the next step of the preprocessing stage which is Part of Speech (POS) Tagging, Myanmar POS Tagger [Zin, 2010] is used. This tagger deals with a combination of supervised and rule-based learning by using pre-tagged corpus. To find the tag for unknown words, Hidden Markov Model (HMM) and bigram model are used. She evaluated the performance by increasing unknown tags number 250 ranging from 0 to 2000 unknown tags. And it is mentioned that every increase in 250 unknown tags, precision decreases between 1 and 3%.She compared the accuracy of POS tagger using HMM with Rule Based Approach (94.56%) to POS tagger only using Rule Based Approach (89.82%).

An example of the Part of Speech (POS) Tagging of the above segmented sentence is as follows:

After POS tagging phase,

ပါပူဝါနယူးဂီနီ_JJ နိုင်ငံ_JJ တောင်ပိုင်း_NN ကုန်းမငြိ_NN ဒေသ_NN တွင်_LTPRP ဇန်နဝါရီ_NN ၂၄_SYM ရက်_NN က_SPRP မေပြိုမှု_NN ဖစ်ပွား_V ရာ_CS ဒေသခံ_JJ ပညိသူ_NN ၆၀_SYM ခန့်_NN သေဆုံးခဲ့ကေငြိး_NN ပါပူဝါနယူးဂီနီ_JJ အရာရှိများ_NNS က_SPRP ဇန်နဝါရီ_NN ၂၅_SYM ရက်_NN တွင်_LTPRP ပေဂြကြားသည်_V ။_SM

Each word is followed by its corresponding POS tag. In the above format, word and its POS tag are divided by an underscore and words are cut by space.

4.3. Template Generation using Conditional Random Fields (CRFs)

To fill the values of the template, it needs to extract the important information from the source. This task is assumed as labeling the text and performed by CRF model. To accomplish this function, CRF++ tool which is an open source implementation of Conditional Random Fields (CRFs) for segmenting/labeling sequential data is applied.

As this tool is customizable, it needs to specify train, test and feature template files in advance. In the training phase, train file and feature template file are utilized to produce a model file which is further used for testing. The sample train data file for CRF model is described Table 2.

Table 2. Train data file for CRF

```
ပါပူဝါနယူးဂီနီ JJ B-Loc
နိုင်ငံ JJ I-Loc
တောင်ပိုင်း NN I-Loc
ကုန်းမငြို့ NN I-Loc
ဒေသ NN I-Loc
တွင် LTPRP O
ဇန်နဝါရီ NN B-Date
၂၄ SYM I-Date
ရက် NN I-Date
က SPRP O
မေပြိုမှု NN O
ဖစ်ပွား V O
ရာ CS O
ဒေသခံ JJ B-DeathScore
ပညြ်သူ NN I-DeathScore
၆၀ SYM I-DeathScore
ခန့် NN I-DeathScore
သေဆုံးခဲ့ကောငြ်း NN O
ပါပူဝါနယူးဂီနီ JJ O
အရာရှိများ NNS O
က SPRP O
ဇန်နဝါရီ NN O
၂၅ SYM O
ရက် NN O
တွင် LTPRP O
ပေဂြကဲြသည် V O
။ SM O
```

In the table, the first column describes the word, the second column is its part of speech and the last column is true answer tag represented in IOB2 format. In table 2, B-Loc means that the word ပါပူဝါနယူးဂီနီ is the beginning word for answer tag 'Location', I-Loc represents the word နိုင်ငံ is intermediate word for the tag corresponding to 'Location' and O refers that the word does not include in the answer tag. By analyzing the news articles, tag sets for the system are depicted in Table 3.

For the features, the word itself and the neighboring words, POS tags, and the type of information tags for the words are employed. The features used for training and testing need to describe in template file. One template is depicted as one line in the template file. The format of the template that will be used to specify a token in the input data is %x[row,col]. For example, if the current token is 'ဇန်နဝါရီ NN B-Date', then the feature of the template %x[0,0] is ဇန်နဝါရီ.

In the testing phase, the model file and test file are used to predict the tag for information extraction. Both the training and test files must be the same format.

The template generated from the template generation phase is as follows:

Loc : ပါပူဝါနယူးဂီနီနိုင်ငံတောင်ပိုင်း ကုန်းမငြို့ဒေသ
Date : ဇန်နဝါရီ ၂၄ ရက်
DeathScore: ဒေသခံပညြ်သူ ၆၀ ခန့်

4.4. Summary Generation

Summary Generation is only a form-filling phase. Summary forms for respective news articles (Earthquake, Landslide, Flooding, and so on) are predefined. When the template from the previous step is received, the blanks in the summary forms are completed with the

Table 3. Tag Sets of the respective natural disasters for CRF model to extract required information

Earthquake	Volcanic Eruption	Tornado	Landslide	Strom	Flooding	Fire
Loc Date Name Magnitude Epicenter Depth DeathScore Injuries HLoss PDamage	Loc Date Name Frequency Smoking Warning	Loc Date Path DeathScore Injuries HLoss PDamage	Loc Date Area DeathScore Injuries HLoss PDamage	Loc Date Name Movement Type DeathScore Injuries HLoss PDamage	Loc Date Rainfall DeathScore Injuries HLoss PDamage	Loc Date BurntArea Smoking DeathScore Injuries HLoss PDamage

corresponding slots of the template produced from the previous step.

Here is the sample summary form of Landslide disaster news article.

___ (Location) ___တွင် ____ (Date) _____က မေပြီ[မူကောင့် ____ (DeathScroe) _____ သေဆုံးခဲ့ကောင်း သိရသည်။

Proposed Output Summary

It is reported that a deadly landslide struck in Southern islands of Papua New Guinea on 24 January killed about 60 residents.

ပါပူဝါနယူးဂီနီနိုင်ငံတောင်ပိုင်း တွင် ဇန်နဝါရီ ၂၄ ရက် က မေပြီ[မူကောင့် ဒေသခံပြည်သူ ၆၀ ခန့် သေဆုံးခဲ့ကောင်း သိရသည်။

5. Conclusion and Future Work

In this paper, informative user-focused single document Myanmar Text Summarization based on CRFs is presented. The main work includes word segmentation, POS tagging, template generation and summary generation. In template generation, it needs to extract the information from the input text to fill the values of the slots of the template. To do so, Conditional Random Field (CRF) model is used by considering information extraction as a labeling task. One thing to pay attention in the summary output is to detect only the final number and to exclude explorations as the application is automatic text summarization. As the future work, the system needs to extend as the multi-document text summarization system.

References

[1] Harabagiu,S.M, and F. Lacatusu, "Generating single and multi-document summaries with GISTEXTER", *In Workshop on Text Summarization (In Conjunction with the ACL 2002 and including the DARPA/NIST sponsored DUC 2002 Meeting on Text Summarization)*, Philadelphia, July 2002, pp. 11-12.

[2] Hu,P., T.T.He, D.H.Ji, "Chinese Text Summarization Based on Thematic Area Detection", *In Proceedings of the Workshop on Text Summarization Branches Out at ACL' 04*, Jul 2004, pp. 112-119

[3] Jaruskulchai,C., C. Kruengkrai, "A Practical Text Summarizer by Paragraph Extraction for Thai", *Proceedings of the Sixth International Workshop Information Retrieval with Asian Languages*, July 2003, pp. 9-16

[4] Kim, J.H., J.H.Kim, D.Hwang, "Korean text summarization using an aggregate similarity", *In Proceedings of IRAL'2000*, pp.111-118

[5] Pa,W.W., *Myanmar Word Segmentation using Hybrid Approach*, Ph.D Thesis, University of Computer Studies, Yangon, Myanmar, January 2009.

[6] Patel,C., K.Gali , "Part-Of-Speech Tagging for Gujarati Using Conditional Random Fields", *Proceedings of the IJCNLP-08 Workshop on NLP for Less Privileged Languages*, Hyderabad, India, January 2008, pp. 117–122.

[7] Peng,F., A.McCallum, "Accurate Information Extraction from Research Papers using Conditional Random Field", *In: Information Processing & Management, 42(4)* **,***In: Human Language Technology, Conference of the North American Chapter of the Association for Computational Linguistics HLT-NAACL (2004)*, Boston, Massachusetts , May 2-7, pp. 329-336

[8] Radev, D. R and McKeown, K, "Generating natural language summaries from multiple on-line sources", *Computational Linguistics, 1998*, 24(3):469-500.

[9] Shen, D., J.Sun, H.Li, Q.Yang, Z.Chen, "Document Summarization using Conditional Random Fields", *In Proceedings of the Twentieth International Joint Conference on Artificial Intelligence (IJCAI'07)*, Hyderabad, India, January 6-12, 2007.

[10] Sparck Jones,K., Automatic summarizing: factors and directions, In Inderjeet Mani and Mark Marbury, editors, *Advances in Automatic Text Summarization*, MIT Press, 1999.

[11] Vanthanavong, S., H. Choochart, "LaoWS: Lao Word Segmentation Based on Conditional Random Fields", *Conference on Human Language Technology for Development*, Alexandria, Egypt, 2-5 May 2011

[12] White, M., T. Korelsky, C.Cardie, V.Ng, D.Pierce, and K.Wagstaff, "Multi-document summarization via information extraction", *In Proceedings of the First International Conference on Human Language Technology Research*, 2001.

[13] Zhang, C., H.Wang, Y.Liu, D.Wu, Y.Liao, B.Wang, "Automatic Keyword Extraction from Documents Using Conditional Random Fields", *Journal of Computational Information Systems4:3(2008)* 1169-1180, March 2008

[14] Zin, K.K., *Myanmar Language Tagging based on Part of Speech Tagger*, Ph.D Thesis, University of Computer Studies, Yangon, Myanmar, October 2010.