Using Anaphora Resolution in Myanmar Summaries

May Thu Naing, Aye Thida
University of Computer Studies, Mandalay
mtn.maythunaing27@gmail.com,ayethida.royal@gmail.com

Abstract

Automatic summarization is the process of reducing a text document with a computer program. A summary retains the most important points of the original document. It is essential for Natural Language Processing (NLP) researches. Automatically summarized text can sometimes result in broken anaphoric references. This is due to the fact that the sentences are extracted without making any deeper linguistic analysis of the text. Anaphoric reference resolution is a common phenomenon in natural language processing, which has correspondingly received a significant amount of attention in the literature. In this paper, we present a Myanmar Pronominal Anaphora Resolution (MPAR) algorithm which bases on Hobbs' (1986) algorithm that works on only the surface syntax of sentences in a given text. This paper also proposes a new summarization method using MPAR algorithm for Myanmar texts. The summarized text using anaphora resolution can achieves more meaningful result than a system not using anaphora resolution.

Keywords*: text summarization, anaphors, pronominal anaphora resolution.*

1. Introduction

Text summarization refers to the task of shortening long text. There are two major directions in text summarization: the extractive and the abstractive paradigm [5]. "The first approach is based on identifying important words in texts by using their frequencies. It determines those sentences that contain a bigger number of the important words [5]." 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 to show up the specific form in the source text [5]." This requires a certain amount of deeper understanding of the text. "This method can also be applied in the case of very large texts, such as a whole novel. It is neither the determination of most significant sentences based on occurrences of frequent words, nor building the discourse structures [5]." In these cases, other methods are 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 and neutral vs. evaluative [3]." A full understanding of the major dimensions of variation, and the types of reasoning required to produce each of them, is still a matter of investigation. This makes the study of automated text summarization an exciting area in which to work.

Information about anaphoric relations could be beneficial for applications such as summarization and segmentation that involve extracting discourse models from text. Many of automatic text summarization systems apply a scoring mechanism to identify the most salient sentences. However, the task result is not always guaranteed to be coherent with each other. It could lead to errors if the selected sentence contains anaphoric expressions. To improve the accuracy of extracting important sentences, it is essential to solve the problem of anaphoric references in advance. Anaphoric dependence is a relation between two linguistic expressions such that the interpretation of one, called anaphora is dependent on the interpretation of the other, called antecedent. The problem of anaphora resolution is to find the antecedent(s) for every anaphora.

The remaining parts of the paper are organized as follows: the relevant works for text summarization and pronominal anaphora resolution in natural language processing are presented in section 2, about Myanmar language introduces in section 3, section 4 proposes

Myanmar pronominal anaphora resolution algorithm, section 5 explain the anaphoric expressions and automatic text summarization and section 6 concludes the paper and identifies future work.

2. Literature Review

Lexical Semantic Resources (LSRs) are the foundation of many NLP tasks such as word sense disambiguation, text summarization, semantic role labeling, question-answering and information extraction [12]. These tasks need on a large scale in different languages. A new method for using anaphoric information in Latent Semantic Analysis (lsa) is proposed and discussed to develop an LSA-based summarizer [8].

Traditionally, anaphora resolution systems rely on syntactic, semantic or pragmatic clues to identify the antecedent of an anaphor. Hobbs' (1986) algorithm is the first syntax-oriented method presented in this research domain [4]. From the result of syntactic tree, they check the number and gender agreement between antecedent candidates and a specified pronoun [4]. P. Rashmi, S. Michael (1998) proposed one of the most important approaches for anaphora resolution in Hindi. They applied a discourse salience ranking to two pronoun resolution algorithms, the BFP and the S-List algorithm [7]. Anaphora Matcher (AM) algorithm is implemented to handle inter-sentential anaphora over a two-sentence context [1]. Dagan, Ido and Alon Itai (1990) introduce a statistical approach, in which the corpus information was used to disambiguate pronouns [2]. Mitkov, Ruslan, Richard and Constantin (2000) proposed a knowledge-poor approach, which can also be applied to different languages (English, Polish, and Arabic) [6]. The main components of this method are so called "antecedent indicators" which are used for assigning scores (2, 1, 0, -1) against each candidate noun phrases [6]. But Myanmar text summarization system using anaphoric information has not been developed.

Therefore, this research aims to implement a system that is based on Hobbs' (1986) algorithm for pronominal anaphora in Myanmar. A syntactic rule based algorithm is run on manually parsed sentences. Hobbs (1986) tested his algorithm for the pronouns he, she, it. The algorithm is adapted successfully for more languages such as Chinese, which has similar Subject-Verb-Object (SVO) structure and follows a fixed word order. And likely, Myanmar language, which is a free word order. It has difficulties to inherent for the application of Hobbs' algorithm. To apply the syntactic anaphora resolution using Hobbs' (1986) algorithm for Myanmar texts, the Hobbs' (1986) algorithm must be modified to find antecedents for pronouns. Therefore, this research describes the anaphora resolution algorithm for Myanmar text.

3. Myanmar Language

The Myanmar Language is the official language of Myanmar. It is also the native language of the Myanmar and related sub-ethnic groups of Myanmar, as well as some ethnic minorities in Myanmar like the Mon. Myanmar language use tonal and pitch-registers, is a largely monosyllabic and analytic language, with a Subject Object Verb (SOV) word order. The written language uses the Myanmar script, derived from the Old Mon script and ultimately from the Brahmi script.

The language is classified into two categories. One is formal, used in literary works, official publications, radio broadcasts and formal speeches. The other is colloquial, used in daily spoken conversation. This is reflected in the Myanmar words for "language": စာ sa refers to written, literary language, and စကား sa-kar refers to spoken language. Therefore, Myanmar language can be explained as either "maran-ma-sa" (written Myanmar language), or "mran-ma-sa-ka:" (spoke Myanmar language). Much of the differences between formal and colloquial Myanmar language occur in grammatical particles and lexical items [10]. Different particles (to modify nouns and verbs) are used in the literary form from those used in the spoken form. For example, the postposition after nouns is ၌ hnai: the postposition in formal Myanmar language after noun is မှာ hma: in colloquial Myanmar language. In this study, we focus on written Myanmar language.

Example 1: အမေသည်အိမ်၌ရှိသည်။

Mother is at home. (Formal form)

Example 2: အမေအိမ်မှာရှိတယ်

Mother is at home. (Spoken form)

3.1. Introducing Myanmar Pronouns

Anaphoric reference type can be classified into abstract (event) references, where an anaphora refers to an event, or a proposition and concrete (entity)

reference, where the anaphora refers to a concrete entity like noun phrase (person, place) qualifiers.

Pronominal anaphora is the most commonly encountered in general written language. Myanmar language has four categories, which are

1. Personal pronoun
2. Demonstrative pronoun
3. Question pronoun
4. Mathematic pronoun

In the above pronouns, question and mathematic pronouns do not need for summary generation. Therefore, we ignore the resolving these two types of pronoun. A comparison of pronominal anaphora in English and Myanmar are shown in Table 1. In Myanmar language, the usage of "it" is not the same as the English use. Therefore, we consider to resolve personal pronouns that do not include "it, itself, its" in Myanmar.

Most algorithms in the literature resolve the pronouns 'he', 'she', 'it', 'her', 'him', 'his' and 'its' in English whenever they have an antecedent that is a noun phrase. In this study, we reduce the scope of anaphoric phenomena and focus on a sub-problem of anaphora resolution. Therefore, our pronoun resolution system will resolve all three types of personal pronouns except for "it" whenever they have an antecedent that is a noun phrase.

Table 1. Pronominal anaphora in English and Myanmar

Type of Pronoun		Anaphora in English	Anaphora in Myanmar
Personal Pronoun	Nominative	He, she, it, they	He(thu), she(thuma), they(thuto)
	Possessive	His, her, its, their	His(thuei), her(thuma ei), their (thutoe ei)
	Objective	Him, her, it, them	Him (thu ko), her(thumako), them (thutoeko)
Demonstrative Pronoun		this, that, these, those	Ei, the, hto, yin, le' gaung

4. Overview of the System

For preprocessing of proposed systems, 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 the proposed system, word segmentation is performed with Myanmar Word Segmenter (MWS) [9]. MWS generated a score, to be used for later processing, for every possible sentence segmentation for a phrase. MWS can handle unknown word cases as it does not depend on a lexicon.

For the next step of the preprocessing stage, which is Part of Speech (POS) Tagging, rule based POS tagging of Myanmar language [9] is used. POS tagging uses the context-free grammer (CFG) as rules, which starts parsing the sentence by means of a left to right parsing structure to define the POS of each word.

In this research, the anaphoric system is a basic rule-based system, focusing on named entity anaphoric relations. To solve anaphoric references for Myanmar language, a rule-based system creates an anaphoric link between the pronoun and its antecedent based on Hobbs (1986) algorithm [4]. The structure of Myanmar language is Subject-Object-Verb (SOV) structure and free word order. So, Hobbs (1986) algorithm is not adapted for Myanmar language. Thus, we propose a new syntax based approach developed to solve anaphoric resolution for pronominal reference for Myanmar texts based on Hobbs (1986) algorithm. This algorithm is the Myanmar Pronominal Anaphora Resolution (MPAR) algorithm, which will be explained in section 5. After implementing, the MPAR algorithm will be used in summarization application that is popular for NLP research.

5. Myanmar Pronominal Anaphora Resolution (MPAR) Algorithm

With nominative subject pronouns and possessive pronouns are in complimentary distribution when it comes to expressing relationships. Anaphors are able to find the antecedent in a local domain. Possessive pronouns look for antecedent farther domain. Nominative case is an absolute criterion for subject status in English. But, the role of subject and object in Myanmar are found to have significant impact on anaphora resolution. Most algorithms in the literature resolve the pronouns 'he', 'she', 'it', 'her', 'him', 'his' and 'its' in English. However, MPAR algorithm that is

based on Hobbs' (1986) algorithm can resolve all personal pronouns that include "they", all possessive in Myanmar texts. The following algorithm Fig 1. shows how to resolve anaphora in Myanmar.

Begin

Input: Parse tree of each sentence in Paragraph

Output: Pronoun Resolution

Step 1: Start with NP node of the last parse tree which includes in pronoun

NP, Pronoun ∈ NP;

Step 2: Go up the tree

If (NP is found) then X:= NP;

else if (VP is found) then X:=VP;

else if (highest S is found) then

{ X:=S;

Go to Step 6.}

Step 3: If (X is NP) then Call *funAnti*(X);

Step 4: If (X is VP) then Call *funAnti*(X);

Step 5: Go to Step 2.

Step 6: Call *funAnti*(X);

Step 7: Go to previous parse tree.

X:=Root node of previous parse tee;

Call *funAnti*(X).

If (X is VP) then Go to Step 4.

If (X is NP) then Go to Step 3.

End

funAnti(X)

Begin

Step 1: Do BFS under X.

Step 2: If (Noun in NP –NOM or Noun in

NP –OBJ is found) then

Anti:= Noun Under NP –NOM or NP –OBJ

Else Continue on BFS.

End

Where,

NP =Noun Phrase

X =variable for node

VP =Verb Phrase.

BFS =Bread first search

NP –NOM =Noun phrase of Nominative

NP –OBJ =Noun phrase of Object

Anti =variable for antecedent

Figure 1. MPAR algorithm

6. Anaphoric Expressions and Automatic Text Summarization

The problem with anaphors when performing automatic text summarization by extraction can be shown in the following example.

Example Paragraph:

မောင်လှနှင့်မောင်မျိုးတို့သည်ကျောင်းရှေ့တွင်စာကျက်နေကြသည်။သူတို့သည်ပထမနှစ်ကျောင်းသားများဖြစ်ကြသည်။မြမြသည်သူတို့၏သူငယ်ချင်းဖြစ်သည်။ကျော်ကျော်သည်လည်းသူတို့၏သူငယ်ချင်းဖြစ်သည်။ထို့ကြောင့်မြမြနှင့်ကျော်ကျော်တို့သည်လည်းသူငယ်ချင်းများဖြစ်ကသည်။

Mg Hla and Mg Myo is studying in front of the university. They are first year students. Mya Mya is their friend. Kyaw Kyaw is also their friend. Therefore, Mya Mya and Kyaw Kyaw are also their friends.

Until now, there is no online Myanmar automatic text summarization available. Therefore, we use English text compactor tools [11] to reduce the text for summarization. Then, we translate this summarized text to Myanmar language. This tool does not use anaphoric information to retrieve summaries. Therefore, it decides to extract only the following two sentences.

Example summarization output without using anaphoric information:

ကျော်ကျော်သည်လည်းသူတို့၏သူငယ်ချင်းဖြစ်သည်။ ထို့ကြောင့်မြမြနှင့်ကျော်ကျော်တို့သည်လည်းသူငယ်ချင်းများဖြစ်ကသည်။

Kyaw Kyaw is also their friend. Therefore, Mya Mya and Kyaw Kyaw are also their friends.

But it will be hard, if not impossible, to tell whom 'သူတို့၏' (their) is. In the worst cases 'သူတို့' (their) can appear to refer to some completely different person. Therefore, summarized text without anaphoric resolution cannot catch the main meaning of the paragraph.

If the anaphoric relations in the summarized texts are resolved with MPAR algorithm, the extract summarized sentences will be unambiguously intelligible and give a meaningful text translation.

Example of summarization with MPAR algorithm:

ကျော်ကျော်သည်လည်းမောင်လှနှင့်မောင်မျိုးတို့၏ သူငယ်ချင်းဖြစ်သည်။ထို့ကြောင့်မြမြနှင့်ကျော်ကျော်တို့သည်လည်းသူငယ်ချင်းများဖြစ်ကြသည်။

Kyaw Kyaw is also Mg Hla and Mg Myo's friend. Therefore, Mya Mya and Kyaw Kyaw are also their friends.

6. Conclusion

The purpose of the present work shows how to use pronominal anaphora resolution in summarization. This paper presents the implementation of pronominal anaphora resolution algorithm for Myanmar written language by taking into account the free words order and grammatical role in pronoun resolution in Myanmar text. The role of subject and object in Myanmar written language are found to have significant impact on anaphora resolution for possessive pronouns. MPAR algorithm was tested for a limited set of sentences depending on Earely Parser [9]. This algorithm can apply to many NLP applications such as information retrieval and questioning/answering.

References

[1] Denber, Michel., "Automatic resolution of anaphora in English," Technical report, Eastman Kodak Co, 1998.

[2] Dagan, Ido and Alon Itai, "Automatic processing of large corpora for the resolution of anaphora references," In Proceedings of the 13th International Conference on Computational Linguistics (COLING'90), Vol. III, 1990.

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

[4] J.Hobbs, "Resolving pronoun references. In Readings in natural language processing", Morgan Kaufmann Publishers Inc, 1986.

[5] Mani Inderjeet, "Automatic Summarization", John Benjamins Pub Co; 2001.

[6] Mitkov, Ruslan, E. Richard and O. Constantin, "A new fully automatic version of Mitkov's knowledge-poor pronoun resolution method," In Proceedings of CICLing- 2000, Mexico City, Mexico.

[7] P. Rashmi ,S. Michael, "Discourse salience and pronoun resolution in hindi", U. Penn Working Papers in Linguistics, 2000.

[8] S. Josef, P. Massimo, A.K Mijail, J. Karel, "Two uses o f Anaphora resolution in Summarization", Preprint submitted to Elsevier Science, 7 December 2006.

[9] Soe Lai Phyue, 2012 "Lexical Analyzer for Myanmar Language", in Proceedings of the 10th International Conference on Computer Applications (ICCA2012), Yangon, Myanmar.

[10] Thet Thet Zin, Khin Mar Soe and Ni Lar Thein, "Myanmar Phrases Translation Model with Morphological Analysis of Statistical Myanmar to English Translation System", University of Computer Studies, Yangon,Myanmar.

[11] http://textcompactor.com/

[12] G Francopoulo, "LMF Lexical Markup Framework", 2013.