

Detection and Resolution of Anaphora in Myanmar Text using A Hybrid Approach

Khin Theink Theink Soe

Tin Htar Nwe, Khin Thandar Nwet

{*khintheinktheinksoe, tinhtarnwe, khinthandarnwet*}@ucsy.edu.mm

Abstract

Anaphora Resolution is the process of resolving references to an entity from the set of referring expressions or discourse. Anaphora resolution plays an important role in the natural language understanding. It is needed to be addressed in almost every application dealing with natural language such as language understanding and processing, dialogue interpretations, machine translation, question-answering, information extraction, text summarization, and so on. Anaphora resolution in English and other languages has been well done in early. But Myanmar language has not sufficiently applied. The paper presents a hybrid approach to resolve pronominal anaphora in Myanmar text. The hybrid model adopted the strategy based on the combination of a rule-based and machine learning approach. Rule-based manner with morphological information is used to collect anaphora and possible antecedents. K-Nearest Neighbor (k-NN) approach is used to select the most probable candidate as the antecedent of the anaphor.

Keywords: Anaphora resolution, candidate, antecedent, Natural Language Processing, k-Nearest Neighbors classifier

1. Introduction

Anaphor is commonly a pronoun or referring word pointing back to previous item in the discourse. Discourse is a group of collocated and

related sentences. Antecedent can be noun, noun phrase, verb phrase or clause in this discourse. Anaphora resolution is a process of determining the antecedent of an anaphor and the subsequent replacement of the anaphor by its antecedent. Understanding of natural language is difficult for computers because natural languages are inherently ambiguous. Human beings can easily manage to pick out the intended meaning from the set of possible interpretations. But computer uses only their limited knowledge and inability in complex contextual situations. The resolution of anaphora to their intended referents has been considered one of the most intriguing and challenging problems in Natural Language Processing (NLP). Anaphora resolution helps to fully and correctly understand the text [7]. Myanmar language exhibits the characteristics of an agglutinative language. The work done of anaphora resolution in Myanmar language is not sufficiently studied.

The main goal of this study is to develop an efficient method for anaphora resolution for Myanmar language. There are two main steps in anaphora resolution. First step is the identification of pronouns or noun phrase from the set of referring expressions or discourses. Secondly, antecedent-anaphora relations are identified and then selects the correct and possible antecedent or candidate.

This paper represents a hybrid approach for resolving Myanmar pronominal anaphora. The collection of anaphora and respective possible antecedents was identified in a rule based manner and morphological features. The most probable

candidate was selected by machine learning based on k-Nearest Neighbor (k-NN) approach.

2. Literature Review

Anaphora resolution approaches were heavily based on domain and linguistic knowledge. Many approaches on anaphora resolution are syntax-based approaches include Hobbs algorithm, discourse-based approaches include the Centering approach and hybrid approaches include Lappin and Leass algorithm.

Thit and Aye (2015) presented the anaphora resolution system using Centering Theory for Myanmar Language. This system resolves the extracted pronouns by using centering theory and outputs sentences that replaced in the position pronouns with their referent. It can resolve only pronouns whose antecedents are in the immediately preceding sentence and can't process sentences in complex grammar structure and didn't consider all of Myanmar Prepositions as marker. It obtained f-measure of 42% and 20% in personal pronouns and some demonstrative pronouns respectively [1].

May and Aye (2014) also proposed a system that is based on Hobbs' algorithm for pronominal anaphora resolution in Myanmar language. This system resolved all three types of personal pronouns except for 'it' and has tested for limited set of sentences depend on Earely parser and used the POS tagger in ML2KR. After anaphora resolve, automatic text summarization system is done by changing Myanmar text with English text compactor tools. This system performed on two types of dataset of short stories in Myanmar and basic Myanmar essays and obtains a substantial accuracy rate of 80% [2].

A.Abolohom and N.Omar [3] proposed a hybrid approach that combines a rule-based manner and machine learning base on k-Nearest Neighbor (k-NN) approach to resolve pronominal anaphora in Arabic in 2015. The rule-based filtering module is used to provide the

morphological and syntactic filter and machine learning technique is used for feature extraction and to classify a new input instance among all stored training instances. It used Quranic corpus annotated with antecedent references of pronoun and the Arabic Statistical POS Tagger and it considered the search limit of 17 sentences. They presented that the proposed hybrid approach is completely reasonable and feasible for Arabic pronominal anaphora resolution.

3. Proposed System

This section presents a hybrid architecture for pronominal anaphora resolution in Myanmar Text. The hybrid approach is based on the combination of rule-based filtering module and machine learning module. Myanmar is a highly inflectional language and is rich in morphology. It has sentence boundary marker. It is a free word order and verb final language, which usually follows the subject-object-verb (SOV) order. Preposition adjunctions can appear in several different places of the sentence. In this study, we can take the datasets of narrative type of author Dagon Taryar. The proper subject-object identification is needed for an efficient anaphora resolution system. The proposed tagset is identified as subject, object, number, gender, line number, and living thing or non-living thing. This system for POS tagging will use Part-of-Speech Tagger and Chunker for Myanmar Language [4]. Pronoun and noun-phrase is selected by using rules for Myanmar text. The most possible candidate in the words defined feature value is selected by using machine learning approach based on k-Nearest Neighbors algorithm. The following subsections discuss in detail the main steps of each component respectively, including pre-processing task, morphological filtering, feature extraction and classification step.

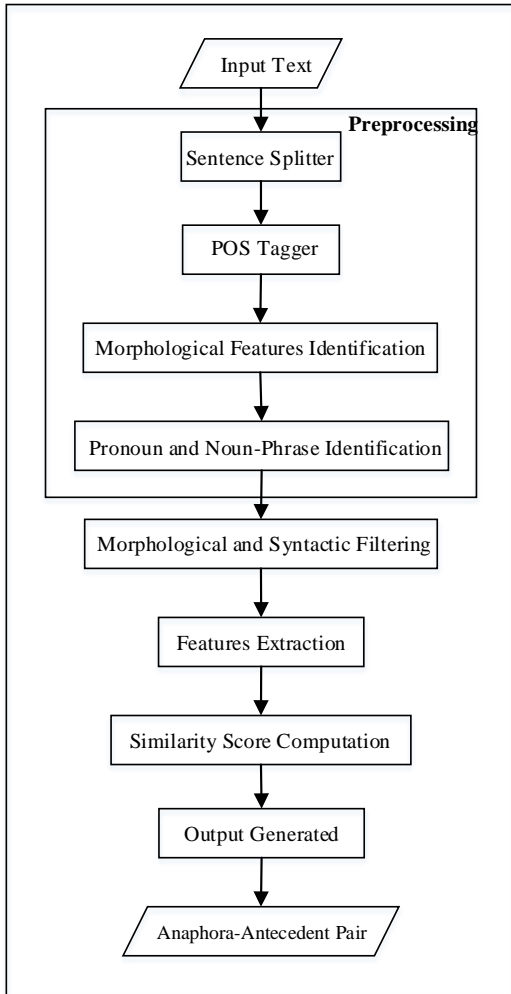


Figure 1. Architecture of proposed system

3.1. Preprocessing

In the pre-processing stage, various NLP techniques are applied. These modules in our system include sentence splitting, POS tagging, features identification for words with morphology and pronoun and noun-phrase identification. We take input text and apply Sentence Splitter in order to generate output expected by POS tagger. In this pronoun resolution system, punctuation mark (။) is used as sentence separator. The output of our

sentence splitter is a sequence of sentences each having a unique sentence number.

Part of Speech (POS) is the process of assigning appropriate linguistic categories (noun, verb, adjective, etc.) of each word in a sentence. POS tagging is one of the most well studied problems in the field of Natural Language Processing (NLP). POS tag for this system according to the Part-Of_Speech Tagger and Chunker for Myanmar Language has included many modules, such as word segmentation module, basic tagging and standard tagging for POS, and chunking module. This tagger uses 14 for basic tags and 9 for standard tags.

In addition, it provides the morphological features (number, gender, line number, semantic type, pronoun type, grammatical role) that are assigned to each word. The pronoun numbers are classified as singular(S), plural(P) and common(C). Gender information is classified as feminine(F), masculine(M), neutral(N) and unknown(U). Line number is needed to defined search limit. We also basically need to define semantic type as living or non-living things because we will mainly find the living things for candidates of anaphora. This study resolves personal, reflexive and possessive types of pronouns. Noun and noun-phrase is also defined with grammatical role such as subject, object, and unknown. And then, we create a list that include the pronouns and noun-phrases with morphological features to match the pairs of pronominal anaphora and its antecedent candidates. We give an example for the preprocessing steps of Myanmar text document such as number, gender, line number, semantic type, pronoun type, grammatical role: **For Example:**

(1) Input text

- 1) ဒေါ်ငြိမ်းခင် အတွက်မှာလည်း ငါ ဒီမိန်းမကို
ဘာမှကြောက်စရာမရှိဘူး ဟု သူ့ကိုယ်သူပြောမိလေသည်။

- 2) မေအစို့မှာလည်း သူ့သမီးနှစ်ယောက် အကြိုက်လိုက်၍ မြောက်ပင့်နေလိုက်လျှင် ဟန်ကျမည်ဖြစ်၏။
- 3) ဆယ့်လေးနှစ်နှင့် ဆယ့်ငါးနှစ် သမီးနှစ်ယောက်မှာ ဖြူဖြူကျစ်ကျစ်ကလေးများ ဖြစ်ကြ၏။
- 4) ပထမ မေ့ကို သူတို့ရှက်နေကြလေသည်။
- 5) နောက် မေကသူတို့၏ဗိုလ်ခွေးကလေးနှစ်ကောင်ကို ချစ်သည်ဟန် ပြလိုက်သဖြင့် သဘောကျသွားကြလေသည်။

- သူတို့@PRNR.Person (P, N, 4, Living, PRNR.Person, U)
- 5) နောက်@NN.Common (C, U, 5, Non-Living, NN, U)
မေ@NN.Person (S, F, 5, Living, NN, S)
သူတို့@PRNR.Person (P, N, 5, Living, PRNR.Person, U)
ဗိုလ်ခွေးကလေး@NN.Common (S, U, 5, Living, NN, U)
နှစ်@NN.Cardinal (P, U, 5, Non-Living, NN, U)
ကောင်@Part.Type (C, U, 5, Living, Part.Type, U)
ဗိုလ်ခွေးကလေးနှစ်ကောင်@NNR.Person (P, U, 5, Living, NN, U)

(2) Feature Extraction

This process is to define the POS tagging and extract features for input text.

- 1) ဒေါ်ငြိမ်းခင်@NN.Person (S, F, 1, Living, NN, S)
ဒီ@PRN.Distobj (C, U, 1, Non-Living, PRN.Distobj, U)
မိန်းမ@NN.Person (S, F, 1, Living, NN, O)
ဘာ@PRN.Question (C, U, 1, Non-Living, PRN.Question, U)
သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U)
ကိုယ်သူ@NN.Body (S, N, 1, Non-Living, NN, U)
- 2) မေ့@NNP.Possessive (S, F, 2, Living, NN, S)
သူ့@PRN.Possessive (S, N, 2, Living, PRN.Possessive, U)
သမီး@ NN.Person (S, F, 2, Living, NN, U)
နှစ်@NN.Cardinal (P, U, 2, Non-Living, NN, U)
ယောက်@Part.Type (C, U, 2, Living, Part.Type, U)
သမီးနှစ်ယောက်@NNPR.Person (P, F, 2, Living, NN, U)
- 3) သမီး@NN.Person (S, F, 3, Living, NN, U)
နှစ်@NN.Cardinal (P, U, 3, Non-Living, NN, U)
ယောက်@Part.Type (C, U, 3, Living, Part.Type, U)
သမီးနှစ်ယောက်@NNPR.Person (P, F, 3, Living, NN, U)
- 4) ပထမ@NN.Ordinal (S, U, 4, Non-Living, NN, U)
မေ့@ NNP.Possessive (S, F, 4, Living, NN, O)

(3) Pronoun and Noun-Phrase Identification

In this step, we will get a list included pronoun and noun-phrase pair after preprocessing.

- [{သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), ဒေါ်ငြိမ်းခင်@NN.Person (S, F, 1, Living, NN, S)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), မိန်းမ@NN.Person (S, F, 1, Living, NN, O)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), မေ့@NNP.Possessive (S, F, 2, Living, NN, S)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), သမီး@ NN.Person (S, F, 2, Living, NN, U)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), သမီးနှစ်ယောက်@NNPR.Person (P, F, 2, Living, NN, U)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), သမီး@NN.Person (S, F, 3, Living, NN, U)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), {သမီးနှစ်ယောက်@NNPR.Person (P, F, 3, Living, NN, U)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), မေ့@ NNP.Possessive (S, F, 4, Living, NN, O)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), မေ@NN.Person (S, F, 5, Living, NN, S)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), ဗိုလ်ခွေးကလေး@NN.Common (S, U, 5, Living, NN, U)}, {သူ့@PRN.Person (S, N, 1, Living, PRN. Person, U), ဗိုလ်ခွေးကလေးနှစ်ကောင်@NNR.Person (P, U, 5, Living, NN, U)}]

In example, we will get 11 pairs for pronoun “သူ” in line number 1.

{သူ@PRN.Person (S, N, 1, Living, PRN. Person, U),
 ဗိုလ်ခွေးကလေး@NN.Common (S, U, 5, Living, NN,
 U)}

3.2. Morphological and Syntactic Filter

In the morphological filtering step, we can reduce the large size of pairs of pronoun and a potential antecedent. For each anaphora and noun-phrase obtained from the created list of anaphora and noun-phrases, we find agreement features like line number, semantic type, number, and gender. If the noun-phrase is exactly non-living thing, this word is not thought in selecting the anaphora’s candidates. Moreover, when we found a pronoun from the Myanmar discourse, we will search the noun-phrases from the list that match morphological features including sentences search limit. If the number of pronoun is singular, the number of noun-phrase also must be singular, and the gender of pronoun is the same with of noun-phrase. But the pronouns used the Myanmar novels have some ambiguous situations to choose gender, example., “သူမ” refers only to feminine in gender but “သူ” refers to both feminine and masculine. Therefore, we must think all types of situation. After morphological and syntactic filtering with line number, number, and gender, we will get as:

[{သူ@PRN.Person (S, N, 1, Living, PRN. Person, U),
 ဒေါ်ငြိမ်းခင်@NN.Person (S, F, 1, Living, NN, S)},
 {သူ@PRN.Person (S, N, 1, Living, PRN. Person, U),
 မိန်းမ@NN.Person (S, F, 1, Living, NN, O)},
 {သူ@PRN.Person (S, N, 1, Living, PRN. Person,
 U), မေ့@NNP.Possessive (S, F, 2, Living, NN, S)},
 {သူ@PRN.Person (S, N, 1, Living, PRN. Person, U),
 သမီး@ NN.Person (S, F, 2, Living, NN, U)},
 {သူ@PRN.Person (S, N, 1, Living, PRN. Person, U),
 သမီး@NN.Person (S, F, 3, Living, NN, U)},
 {သူ@PRN.Person (S, N, 1, Living, PRN. Person,
 U), မေ့@ NNP.Possessive (S, F, 4, Living, NN, O)},
 {သူ@PRN.Person (S, N, 1, Living, PRN. Person,
 U), မေ@NN.Person (S, F, 5, Living, NN, S)},

After filtering, it will leave 8 pairs for pronoun “သူ” in line number 1 to find the antecedent of anaphora.

3.3. Features Extraction

Features are allocated scores and can then be ranked by their scores. Those features with the highest scores can be selected for inclusion in the training dataset, whereas those remaining can be ignored. Features extraction is an extremely important task in Natural Language Processing (NLP). Machine Learning (ML) techniques mainly required data to be represented as a feature vector of attribute and value pairs. These features were applied on the models for classification [3]. The features in this system are grouped in three categories: features of the pronominal anaphora, features of the antecedent candidate, and features the relationship between both. The used feature set composed of 7 features. They are:

F1 – Candi-Line-No: The line number of noun-phrase must be included in the search limit line numbers for each anaphora.

F2 – Ana-Ante-Number: Its possible values are 0, 1, 2. If the anaphora and the antecedent agree in number, the value is 2, otherwise the value is 0. If the noun phrase is unknown, the value is 1.

F3 – Ana-Ante-Gender: It has two types of value (1,2). When the anaphora and antecedent match in gender, the value is 2, otherwise the value is 1. Myanmar language has some ambiguous situation for usage of pronoun.

F4 – Same-Sent: If the anaphora and the antecedent candidate are in the same sentence the value is 0, otherwise the value is 1.

F5 – Dist: The sentence distance between the anaphora and the antecedent candidate.

F6 – Candi-subj-NP: If the candidate is the subject of the sentence, the value is 2, else 1.

F7 – Freq: If the candidate has been repeated more than one in search limit, the value is 2, else 1.

3.4. k-Nearest Neighbor Algorithm

K-nearest neighbor is a method to perform the classification of objects based on the learning data that were located closest to the object. K-NN algorithm is a type of instance-based learning, or lazy learning, where the function is only approximated locally and all computation is deferred until classification. The k -NN algorithm is among the simplest of all machine learning algorithms. k-NN algorithm is used in many application areas such as data mining, statistical pattern recognition, image processing. It is a method to perform the classification of objects based on the learning data that were located closest to the object. Learning data projected into many dimensional space, where each dimension represents the features of the data. The space is divided into sections based on the learning data classification. K value is best for this algorithm depends on the data, in general, the value k may reduce the effect of noise on the classification, but it makes the boundaries between each classification becomes more blurred. We should specify a positive integer for k to get the correct antecedent [5].

The k-nearest-neighbor classifier is commonly based on the Euclidean distance between a test sample and the specified training samples. The Euclidean distance between the points (a_1, a_2, \dots, a_p) and (b_1, b_2, \dots, b_p) is defined as

$$d(\mathbf{a}, \mathbf{b}) = \sqrt{\sum_{p=1}^n (a_p - b_p)^2} \quad (1)$$

where $d(\mathbf{a}, \mathbf{b})$ is the distance between pronoun and noun-phrase pairs from the list, n is the number of feature extracted. K value of the KNN algorithm is a factor which indicates a required number of words from the collection which is closest to the selected word [6].

4. Data used for Experiment

Data for this system is collected from Myanmar novel “May” (မေ) narrative type of author “Dagon Taryar”. The proposed system is relied on supervised learning. Myanmar language has four categories such as personal pronoun, demonstrative pronoun, question pronoun and mathematic pronoun. Among them, we will resolve personal pronoun. The training set consists of about 800 pronouns for nominative, objective, possessive and reflexive types that include in personal pronoun type.

5. Conclusion

This paper proposed a hybrid approach for Myanmar Anaphora Resolution system. This approach leverages the strengths of both rule-based and machine approach. The result of this study will be helpful to create a better anaphora resolution tool for Myanmar Language.

Every anaphora resolution system needs correct Part of Speech tagging for the document. Anaphora resolution system is widely used in summarization, machine translation, question-answering system, and so on. This system can resolve four pronouns: nominative pronoun, objective pronoun, possessive pronouns and demonstrative pronouns. The proposed anaphora resolution system based on the combination of rule-based and machine learning approach is completely reasonable and feasible for Myanmar Language. The future work will be directed to resolves for all pronoun types. And then, a hybrid

approach will be created by using other machine learning approach such as Random Forest, Support Vector Machine Algorithm, and so on.

References

- [1] Thit Lwin and Dr.Aye Thida, “Myanmar Anaphora Resolution based on Centering Theory”, UCSM, 2015.
- [2] May Thu Naing and Dr.Aye Thida, “Pronominal Anaphora Resolution Algorithm in Myanmar Text”, International Journal of Advanced Research in Computer Engineering & Technology (IJARCET) , August 2014, p.2795-2800.
- [3] Abdullatif Abolohom and Nazlia Omar, “A Hybrid Approach to Pronominal Anaphora Resolution in Arabic”, Journal of Computer Science 2015, May 11, jcssp.2015.764.771.
- [4] Phyu Hnin Myint, Tin Myat Htwe and Ni Lar Thein, “Lexicalized HMM-based Part-of-Speech Tagger for Myanmar Language”, Proceedings of the Tenth International Conference on Computer Application (ICCA 2012), Yangon, Myanmar, February 28-29, 2012.
- [5] M. Bramer, Principles of Data Mining, Landon: Springer, 2007.
- [6] Amri Danades, Devie Pratama, Dian Anggraini and Diny Anggriani, “Comparison of Accuracy Level K-Nearest Neighbor Algorithm and Support Vector Machine Algorithm in Classification Water Quality Status”, 2016 IEEE 6th International Conference on System Engineering and Technology (ICSET), Bandung, Indonesia, October 3-4, 2016.
- [7] Kalyani P. Kamune and Avinash Agrawal, “Hybrid Approach to Pronominal Anaphora Resolution in English Newspaper Text”, I.J. Intelligent Systems and Applications, 2015, 02, 56-54.