A COMPARATIVE STUDY OF MACHINE LEARNING ALGORITHMS FOR ANAPHORA RESOLUTION OF MYANMAR LANGUAGE

Khin Theink Theink Soe

University of Computer Studies, Yangon
Yangon, Myanmar
khintheinktheinksoe@ucsy.edu.mm
https://www.ucsy.edu.mm

Khin Mar Soe

University of Computer Studies, Yangon
Yangon, Myanmar
khinmarsoe@ucsy.edu.mm
https://www.ucsy.edu.mm

Abstract

Anaphora resolution (AR) is the task of identify noun phrase that refers to the same entity as earlier and later items in the set of referring expressions or discourse. It is one of the more prolific areas of research in the Natural Language Processing (NLP) community and has correspondingly received a significant amount of attention in the literature. Anaphora resolution can support to improve the accuracy in almost every NLP application namely machine translation, text summarization, information extraction, dialogue interpretation, question-answering, etc. Myanmar language needs to be sufficiently applied in anaphora detection and resolution. This paper focuses on the comparison of Myanmar anaphora resolution system using machine learning algorithms. Four popular machine learning algorithms, Decision Trees, Support Vector Machines (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN) algorithms were used to assess the feature sets. On novel datasets from Myanmar, several comparative experiments were conducted. The results were then discussed, and conclusions were taken. The experimental results show that our method produces acceptable results.

***Keywords*: Anaphora Resolution; Candidate; Antecedent.**

1. Introduction

Anaphoric word, such as pronouns, pro-verbs, and definite noun phrases are used in human languages to avoid repetitive words and repetitions. It is important to define relevant entities, especially in discourse analysis, referred to by anaphoric expressions in NLP. Computers find it challenging to comprehend natural language due to its inherent ambiguity. One of the fundamental goals of NLP is to build computer systems that can understand natural language. Human beings can effortlessly discern the intended meaning from a range of potential interpretations. However, computers depend solely on their constrained knowledge and struggle in complex contextual scenarios. An anaphor is usually a pronoun or referring word that points back to a previous item in the discourse. Discourse is a group of collocated and related sentences [Abolohom et al. (2021)]. Antecedent is the entity to which anaphor refer. Anaphor and antecedent are two fundamental concepts from the problem of anaphora resolution. Pronoun resolution, also known as anaphora resolution, is the challenge of locating references to a pronoun in earlier statements. Nouns, noun phrases, verb phrases, and/or clauses can be used as references. Anaphora resolution's primary goal is to identify a pronoun's correct antecedent within the collection of referring expressions. A pronoun's antecedent can be determined by a variety of qualities, including verb predicates for plural and event pronouns, grammatical relations for person pronouns, and number gender agreement features. The search range of a specific method has been specified, in which all noun phrases (NPs) preceding an anaphora are identified as candidates for antecedents, and many anaphora resolution factors are used to track the correct antecedents. The features of anaphora and their corresponding antecedents were determined using morphological characteristics and rule-based methodology.

Anaphora resolution is a complicated problem in the NLP and has attracted the attention of many researchers. Myanmar language exhibits the characteristics of an agglutinative language. The work done of anaphora resolution in Myanmar language is not sufficiently studied. The complexity of the problem is based on the following. Preprocessing of texts typically includes part of speech tagging and morphological analysis, detection of noun phrases and syntactic parsing. Through of these steps can be automated with a certain degree of precision, which affects the outcome as errors are accumulated from precious steps. Anaphora is a pervasive phenomenon in natural language, essential for nearly every NLP application. Extensive work in this area is based on various theoretical approaches. In this paper, the most probable candidate of Myanmar anaphora is implemented by using popular machine learning algorithms based on Decision Trees, Support Vector Machines (SVM), Naïve Bayes, and K-Nearest Neighbors (KNN) algorithms.

2. Related Works

Over the 1960s, the computational models of anaphora resolution were developed, applying the theories with knowledge of syntactic, commonsense, and discourse for anaphora. It is defined as theoretically-oriented and ambitious work regarding the types of anaphora treated. Such approaches were focused heavily on domain and linguistic knowledge. In the 1990s, with the need to become more stable, language-independent and inexpensive NLP systems, the researchers were encouraged to increase the availability of annotated entities and corpora, which impelled new anaphora resolution systems to grow. A powerful resource is provided with annotated corpora, containing morphological semantic and syntactic information from co-occurrence rule derivation to training machine learning algorithms or statistical approaches. Three common fields of anaphora resolution can be identified: statistical, machine learning and syntax-based approaches.

In 2015, Thit and Aye have been presented an anaphora resolution system for the Myanmar language using Centering Theory. They use the testing set that contains 230 utterances having 4025 words and 178 anaphora. These testing date sentences are collected from biography of famous author in Myanmar and simple sentences of Myanmar Grammar book. This system can solve correctly 42% in personal pronoun and 20% in demonstrative pronoun. It ignored some factors such as number, gender, recent frequency and animistic features that affect accuracy [Lwin et al. (2015)].

For Anaphora resolution of Myanmar Language, Hobbs' algorithm is presented by May and Aye in 2014. With the exception of "it," it handled all three forms of personal pronouns and tested a small sample of sentences that relied on the Earely parser and the POS tagger in ML2KR. The resolving data are used in the automatic text summarization system and changed Myanmar text with English text compactor tools. The dataset of short stories and basic essay in Myanmar are tested and a substantial accuracy rate obtains as 80% [Naing et al. (2014)].

For the first time, the researchers Soon, Ng, and Lim achieved outcomes that were on par with non-learning strategies. Not only did they resolve pronouns, but all definite descriptors as well. To generate feature vectors, they employed training data from a tiny annotated corpus. Following that, a machine learning algorithm that builds classifiers using a decision tree technique called C5, a variant of C4.5 (Quinlan 1993), was given these training instances. Their end-to-end system, which combines sentence segmentation, part-of-speech tagging, morphological processing, noun phrase recognition, and semantic class determination, is a noteworthy feature. They obtained 58.6% recall, 67.3% precision, and 62.6% F-measure using 12 characteristics. Using the training and test corpora from MUC-6 and MUC-7, they assessed their system. They quantify each feature's contribution to system performance and determine that alias, string match, and appositive are the features that have the biggest impacts. Pronouns are the subject of very few features, aside from the general number and gender agreement features. Additionally, pronouns are not discussed in the mistake analysis; instead, it solely addresses noun phrase resolution. The absence of syntactic features and salience metrics for pronoun resolution suggests that a large portion of the poor performance must be attributed to pronoun resolution errors [Ram et al. (2013)].

In order to resolve pronominal anaphora in Arabic, A. Abolohom and N. Omar suggested a hybrid technique in 2015. It combines a rule-based method with machine learning based on the k-Nearest Neighbor (k-NN) approach. The morphological and syntactic filter is provided by the rule-based filtering module, while feature extraction and new input instance classification are accomplished using machine learning techniques. It took into account the search limit of 17 sentences and employed the Arabic Statistical POS Tagger together with the Quranic corpus annotated with antecedent references of pronouns. They demonstrated that, in terms of Arabic pronominal anaphora resolution, the hybrid approach that has been suggested is both acceptable and workable [Abolohom et al. (2015)].

3. Anaphora Resolution for Myanmar Language

A member of the Sino-Tibetan language family's Lolo-Burmese branch is Burmese, also known as Myanmar. Since it is the official language of Myanmar (Burma), it is mostly spoken there. An estimated 33 million people in Myanmar and the neighboring countries speak Burmese. It is written in the Burmese script, which is a syllabic writing system similar to Thai that represents sounds with circles and other shapes. The languages of Shan, Karen, and Mon are likewise closely linked to Burma.

Burmese is spoken in two registers: high and low. Formal speeches, newspapers, radio, and literature all use the high register. Comic books, television, casual literature, and daily conversation all use the low register. Certain

Burmese writers have advocated for the adoption of the low register and suggested doing away with the high register since the 1960s. The name Myanmar (မြန်မာ [mjəmà]) is the high register name for the country, and Burma (ဗမာ [bəmà]) is the low register name.

The Burmese or Myanmar script evolved from the Mon script, which was derived from a southern Indian script in the 8th century. The earliest known inscriptions in the Burmese script date back to the 11th century. Myanmar language is basically ordered as subject-object-verb (SOV) and there are 33 consonants: beginning with "က" and ending with "အ". Although there are only eight Part-of-Speech classes are classified for English language, Myanmar language has nine Part-of-Speech classes. These are Noun ("နာမ်"), Pronoun ("နာမ်စား"), Verb ("ကြိယာ"), Adverb ("ကြိယာဝိသေသန"), Adjective ("နာမဝိသေသန"), Conjunction ("သမ္ဗန္ဓ"), Postpositional Marker ("ဝိဘတ်"), Interjection ("အာမေဍိတ်") and Particles ("ပစ္စည်း"). Users of Myanmar Language usually use space as they see fit, some write with no space at all. There is no fixed rule for using space to segment. This characteristic makes parsing and other NLP related processing of Myanmar more complicated than English ones.

A language relationship between two textual entities is known as an anaphora, and it is established when one textual entity (the anaphor) alludes to another, generally earlier textual thing (the antecedent). Finding an anaphora's antecedent is the process of solving an anaphora. For example,

ဦးဘသည် ရုံးသို့သွားသည်။
သူသည် လာမည့်နှစ်တွင် အသက် ၆၀ ပြည့်မည်။

"He- သူ" in the second sentence is anaphor and "U Ba- ဦးဘ" is antecedent for this anaphor. Some sentences have two or more nouns and noun phrase that may be ambiguous for each anaphora. Therefore, this situation can be satisfied with the anaphora resolution.

The process of anaphora resolution consists of three main steps.

(1) Anaphors are selected from the given discourse or group of words.
(2) Possible candidates for an anaphor that selected to resolve are identified within the sentence search limit. The sentence search limit is the search scope, the space where it is likely to find the correct antecedent.
(3) The last is to decide which the antecedent of the given anaphor may be the noun or noun phrase of the possible candidate list.

When the possible antecedent for this pronoun is determined, it considers the specific features for each anaphor, each antecedent and the relation between this anaphora and antecedent. The feature set has usually included the features such as number agreement, gender agreement, grammatical relations, the search scope and repetitive, and so on.

4. Types of Myanmar Pronoun

There are four types of pronouns for Myanmar language. There are personal pronouns, demonstrative pronouns, interrogative pronouns and mathematic pronouns.

(1) Personal Pronoun can be a certain person, thing, or group, and may take on various forms depending on singular or plural number. They may take various forms depending on case, gender or formality such as feminine, masculine, or neuter. A personal pronoun represents as follows:

- 1st Person, the person who speaks
  e.g. I, we, me, us
- 2nd Person, the person who is spoken to
  e.g. you
- 3rd Person, a person or a thing which is spoken about
  e.g. she, he, they, it, her, him, them

Note that the word "you" has the same meaning whether it is used as a subject or an object, single or plural.

(2) Demonstrative Pronoun is used to substitute a noun, noun phrase, activity, or situation, already mentioned in conversation or written work. This is encouraged to prevent repetition, believing that using a demonstrative pronoun won't lead to misunderstanding. The most common use of demonstrative pronouns in Myanmar language are "ဤ၊ သည် - this", "ထို - that", "၎င်း၊ ယင်း - it". These pronouns are usually used to refer inanimate objects and mostly used with the noun phrase in the form demonstrative pronouns to demonstrate the previous object.

(3) Interrogative pronoun is used to build asking questions and it represent a person or a thing what we are asking the question about. They are also known as Wh-words in English. These pronouns can be used as relative pronouns, which may be found in questions or indirect questions. In many cases these pronouns don't have

antecedents, the word that referred by the pronoun. The interrogative pronouns in Myanmar language are similar in using to English language. The most typical interrogative pronouns in Myanmar are “ဘာ - what”, “ဘယ်သူ - who”, “ဘယ်သူ၏၊ မည်သူ၏ - whose”, and so on.

(4) Mathematical pronoun Mathematical pronoun is the word that refers to the number of things according to the grammar rule of Myanmar. Articles in English language such as (a, an, the) are same with some of the mathematic pronouns in Myanmar. “အချို့ - some”, “အားလုံး - all” and “အနည်းငယ် - any” are the examples of mathematic pronouns for Myanmar language.

5. Forms of Anaphora

Various types of anaphors are categorized based on form of the anaphor.

5.1. *Pronominal anaphora*

Pronominal anaphora is a popular anaphora and may consist of sentences which have both singular and plural pronouns of personal, possessive, reflexive, relative, and demonstrative. For example,

"အောင်အောင်သည် သူ၏အခန်းထဲသို့ ပြေးသွားသည်။"

“Aung Aung ran into his room.”

In this sentence, the anaphoric expression is “သူ၏- his” refers to “အောင်အောင် - Aung Aung”.

"အောင်အောင်သည် သူ့အခန်းထဲသို့ ပြေးသွားသည်။"

The difference between these two Myanmar sentences is the antecedent for the reflexive pronoun “his”, the meaning “သူ့” is usually same the meaning of “သူ၏” as the nature of Myanmar language.

5.2. *Noun anaphora*

The anaphoric relation between a name and a noun phrase is the noun anaphora. For example,

"ဒေါ် မြင့်မြင့်သည် ကျွန်ုပ်တို့ကိုအင်္ဂလိပ်စာသင်သည်။ ဆရာမသည် တပည့်တွေအပေါ်အလွန်စိတ်ရှည်သည်။"

“Daw Myint Myint taught us English. The teacher is very patient with the students.”

In this example, the word “ဒေါ် မြင့်မြင့် – Daw Myint Myint” is refered by the word “ဆရာမ - The teacher” as mentions noun anaphora.

5.3. *Verb anaphora*

The verb anaphora occurs when the verb anaphor is followed by the verb or verb phrase (VP). For example,

“မမသည် စိတ်ညစ်ချိန်တွင် လုပ်လေ့ရှိသည့်အတိုင်း ခန်းဆီးများကို လဲလှယ်ခဲ့သည်။"

“Ma Ma changed her urinals, as she always did when she was upset.”

In the above sentence, the word “did” refer to an action Ma Ma, “changed her urinals”. In Myanmar, the word “လုပ်လေ့ရှိ” means the word “ခန်းဆီးများကိုလဲလှယ်”.

5.4. *Cardinal/ Ordinal*

The cardinal numbers apply to the number of something or quantity of something. The ordinal numbers apply to the numeral positions of something. The anaphor is a cardinal like “one, two, three” and an ordinal like “first, second, third”. For example,

"ကျွန်ုပ်တွင် ကြောင်ငါးကောင်ရှိသည်။ နှစ်ကောင်မှာ အထီးဖြစ်ပြီး၊ သုံးကောင်မှာ အမဖြစ်သည်။"

“I have five cats. The two is male and the three is female.”

In this example sentences, cardinals such as “နှစ်ကောင် - the two” and “သုံးကောင် - the three” refer to “ကြောင်ငါးကောင် – five cats”.

5.5. *Adverb anaphora*

In the following example, the anaphor “ထိုနေရာ - there” refers to the antecedent “ပဲခူးမြို့ - Bago”.

“ကျွန်မ၏ဇာတိမှာပဲခူးမြို့ဖြစ်၍ ထိုနေရာတွင် အထက်တန်းအောင်ခဲ့သည်။”

“I was born in Bago and graduated there.”

5.6. *Ontology based anaphora*

The ontology-based anaphora is the most complicated anaphor category, where the anaphor refers to any real-world information that has not been discussed anywhere in the discourse before. For example,

“မြန်မာနိုင်ငံသည် စက်ရုံအချို့နှင့်အရေးမကြီးသည့် လက်လီဆိုင်များကို ကိုဗစ် 19 ကြောင့် ပိတ်ခဲ့ရသည်။ ဤအကျိုးဆက်ကြောင့် အလုပ်သမား ၅၀၀၀၀ ခန့် အလုပ်လက်မဲ့ ဖြစ်ခဲ့ကြသည်။”

“Myanmar has been closing some factories and nonessential retail stores for Covid 19. Because of this effect, about 50,000 workers lost their jobs.”

The anaphoric word "ဤ - this" lacks an explicit relation to world knowledge, requiring context such as "this effect is for Covid 19" to resolve its reference.

5.7. *Zero anaphora*

Zero anaphora is also known as 'invisible' anaphora, and does not contain an explicit word or phrase in this form of anaphora. For example,

“မလှသည်ပန်းသီးများကို ခူးယူနေသည်။ ခူးခဲ့သော ပန်းသီးများသည် အလွန်လတ်ဆတ် ချိုမြသည်။”

“Ma Hla is picking apples. The apples picked are very fresh and sweet.”

In this example, the pronoun “သူမ - she” was omitted but we can understand that the subject of “ခူးခဲ့သောပန်းသီးများ - the apples picked” was “မလှ – Ma Hla”.

5.8. *Pleonastic anaphora*

The pleonastic pronoun "it" in English is considered a serious problem. The problem is that "it" does not refer to something in the sentences "It is raining." or "it is ten o'clock." but is just an expression. This type of anaphor is called pleonastic. The use of “it” in Myanmar language is not the same as in English. Similar use of pronoun can rarely be found in Myanmar written text, but it can be found in dialogue style writing. For example,

“အချိန်တန်ပြီ။”

“It is right time.”

Pleonastic anaphora is usually mentioned as the form that has no subject in Myanmar language.

5.9. *Intra-sentential anaphora and Inter-sentential anaphora*

Anaphora is also categorized by its position in sentences as intra-sentential or inter-sentential anaphora. It is called intra-sentential anaphora if the anaphor and its antecedent occur in the same sentence and if they occur in the separate sentence, it is called inter-sentential anaphora. For example,

“ဦးကျော်သည် သူ၏သားများကို မြန်မာသူရဲကောင်းများ အကြောင်းပြောပြလေ့ရှိသည်။”

“U Kyaw used to tell his sons about Burmese heroes.”

The word anaphor “သူ၏ - his” and the word antecedent for this anaphor “ဦးကျော် – U Kyaw” are in the same sentence.

“မမသည် ပန်းများကို နှစ်သက်သည်။ သူမအကြိုက်ဆုံးမှာ နှင်းဆီနီ ဖြစ်သည်။”

“Ma Ma loves flowers. Her favorite is the red rose.”

Antecedent “မမ – Ma Ma” is in a different preceding sentence from the anaphor “သူမ - her”.

Anaphora can relate to both forward and backward reference in general, according to certain linguists. Cataphora is the name given to the forward anaphora. Words or phrases that allude to something discussed later in the conversation are said to be cataphoric. Anaphora can relate to both forward and backward reference in general, according to certain linguists. Cataphora is the name given to the forward anaphora. Words or phrases that allude to something discussed later in the conversation are said to be cataphoric. “Because he was very cold, Aung Aung put on his coat.” The pronoun “he" is a cataphor, it points to the right toward its postcedent “Aung Aung”.

6. Methodology

This section describes a thorough exploration of the architecture of the proposed technique in Myanmar anaphora resolution, detailing the functionality of each component within the Myanmar anaphora resolution model. The system aims to resolve pronominal anaphora referring to personal pronouns such as first person, second person, third person, possessive, and reflexive types. The process encompasses the following stages: data collection, pre-processing, anaphora identification, noun phrase identification, feature extraction, and anaphora resolution.

6.1. *Data Collection*

The narrative type of Myanmar novel "May" written by the author "Dagon Tayar" is used as the data set including 3511 formal sentences in this study. This data set has included about 1780 pronouns and 80 noun phrases. A quantitative measure for this Myanmar anaphora resolution system is presented in Table 1.

Types of Pronouns	Count	%
Personal Pronoun	910	51
Possessive Pronoun	451	25
Reflexive Pronoun	9	0.5

Table 1. Quantitative Measures of Myanmar Data Set

6.2. *Pre-processing module*

In preprocessing stage of a Myanmar pronoun resolution system, the input text will be split into formal sentences. All Myanmar sentences will be end with the punctuation mark (။) as sentences separator. And then, the input sentences will be defined a unique sentence number.

Part of Speech (POS) tagging is the process of giving each word in the corpus an opposite part of speech or word category. The part-of-speech tagger and chunker for the Myanmar language are parsed into each of these sentences and applied to the sentence. Verbs, nouns, adjectives, adverbs, prepositions, determiners, conjunctions, pronouns, and particles are among these categories. These groups are characterized by a particular set of morpho-syntactic traits.

6.3. *Anaphora Identification*

The identification of anaphora is carried out by referring to their grammatical parts of speech. Nominative, objective, reflexive, and possessive pronouns were marked based on to their occurrence within a sentence.

The words with the part of speech tags such as '@PRN.Person#', '@PRN.Possessive#', '@PRN.Reflexive#', '@PRNR.Person#', '@PRNR.Possessive#' and '@PRNR.Reflexive#' have been extracted as anaphora to resolve in the proposed system. The tag 'PRN' refers to singular pronoun and 'PRNR' also refers plural pronoun.

6.4. *Noun Phrase Identification*

In the system, the noun phrases with the POS tag such as '@NNP.Livingthing#', '@NNP.Possessive#', '@NN.Person#', '@NN.Possessive#', '@NNPR.Livingthing#', '@NNPR.Possessive#', '@NNR.Person#' and '@NN.Group#' are extracted as the possible candidates for each anaphor in the discourse.

Example sentences with pronouns for anaphora and noun phrases for candidates are described as follow:

ဦးဘေမောင်@NNP.Livingthing#မှာ@PPM.Subject# သူ@PRN.Person#ပြော@VB.Common#သည်မှာ@RB.State#မှန်ကန်ကြောင်း@NN.VBConvertEnd#ထောက်ခံသည့်အလား@JJ.VBConvert#ကြောင်နက်ကလေး@NN.Animal#ကို@PPM.Object#လက်ညှိုး@NN.Body#ညွှန်ပြ@VB.Common#လေ@Part.Support#သည်@SF.Declarative#

လာလာ@VB.Compound#၊@SYM.Common#ပူစီပူစီ@NN.Common#ဟု@CC.Part#သူ@PRN.Person#သည် @PPM.Subject#လက်@NN.Body#ကို@PPM.Object#လှုပ်@VB.Common#၍@CC.Sent#ခေါ်@VB.Common# လေ@Part.Support#သည်@SF.Declarative#

In the above sentences, many anaphora and many noun phrases for candidates are included and then the probable candidate for each anaphor is in the same sentence or another sentence. Based on a few manually extracted rules, we created this module for noun-phrase identification. A linguistic relation when there is a correlation between components and linguistic forms and which is established by grammatical rules is known as the grammatical relation. The primary factors that establish the appropriate antecedent are the grammatical relations. At this point, every sentence in the text is examined to determine the subject and object grammatical correlations for each non-plural noun phrase.

6.5. *Feature extraction*

Features extraction is mainly essential task in NLP. In this stage, all anaphora and antecedents are identified with the morphological feature extraction and the most possible candidate are evaluated with the machine learning algorithms. Feature vector of each attribute and value pairs of the training data are required in the machine learning approach. In this study, three categories are used to classify the aspects: pronominal anaphora features, antecedent features, and features relating to the relationship between the anaphora and the antecedent. Each pair of anaphora and antecedent is represented as the following feature vector: (F1: value1, F2: value2,….., Fm: value m)

Features are identified in the following.

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: The anaphora and the antecedent agree in number, or not (singular or plural).

F3 - Ana-Ante-Gender: The anaphora and antecedent match in gender, or not.

F4 - Same-Sent: The anaphora and the antecedent are in the same sentence, or not.

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

F6 - Ana-Candi-Pos-NP: The anaphora and candidate noun phrase are the position of subject, object, or other.

F7 - Freq: The candidate has been repeated more than one in search limit, or not.

The first part of features extraction is identified the values of F1, F2, F3, and F6 for all anaphora and candidates. The output for this stage all anaphora is described with line number. After selecting each anaphor, the features values for F4, F5 and F7 will be specified based on the relation between this selected anaphor and all candidates that include within sentence search limit 17.

6.6. *Classification Methods*

Four machine learning techniques are applied in this paper to solve the anaphora resolution challenge. The K-Nearest Neighbors (KNN), Decision Trees, Support Vector Machines (SVM) and Naive Bayes are introduced in the following subsections.

K-Nearest Neighbour (kNN): A technique for classifying objects based on the learning data that were closest to the item is called K-Nearest Neighbor (kNN). The instance-based learning involves approximating the function locally and deferring all calculations until the classification stage.

For classification tasks, the kNN algorithm assigns a class label based on the majority rule, selecting the label most frequently found among the neighboring data points. In other words, the classification output is determined by the mode of the nearest neighbors. The kNN algorithm works as a supervised learning algorithm, meaning it is fed training datasets it memorizes. It relies on this labeled input data to learn a function that produces an appropriate output when given new unlabeled data.

Learning data is projected into a multi-dimensional space, with each dimension representing a different feature of the data. The optimal K value for this algorithm depends on the data. Generally, a higher K value can reduce the effect of noise on classification, but it also makes the boundaries between classifications more blurred. A positive integer should be specified for K to accurately determine the correct antecedent. The kNN classifier typically relies on the Euclidean distance between a test sample and the specified training samples. The Euclidean distance between the points (a1, a2, ……, ap) and (b1, b2, ……, bp) is defined as

$$d(a,b) = \sqrt{\sum_{p=1}^{n} (\mathrm{a}_p - b_p)^2} \tag{1}$$

where d(a,b) is the separation between noun-phrase and pronoun-phrase pairings in the list, where n is the number of extracted features. The kNN algorithm's K value is a factor that shows how many words from the collection must be the closest to the chosen word.

Decision tree: a popular machine learning algorithm for classification and regression methods is decision tree. It operates by dividing the data into subsets according to the values of input features, forming a tree-like structure of decisions. The C4.5 based on the type of target variables is used in the system and was modified to suit our objectives. The process starts at the root node with the entire set of examples. If all examples belong to the same class, no further division is needed, and the node becomes a leaf node representing that class. If all examples have identical feature values but different class values, division is not possible, and this is treated as data noise; the node becomes a leaf node representing the majority class. Otherwise, the attribute that best divides the current set of examples, based on the highest information gain, is chosen. A child node is created for each value of this feature, and examples with that value are transmitted to the corresponding subset. For example, if the target attribute takes on c different values, then the entropy of S relative to this c classification is expressed as

$$Entropy(S) = \sum_{i=1}^{c} (-p_i \log_2 P_i) \tag{2}$$

We can define information gain as a measure of an attribute's efficacy in categorizing the training data, given that entropy is a measure of impurity in a set of training datasets.

$$Gain(S,A) = Entropy(S) - \sum \frac{|S_v|}{|S|} Entropy(S_v) \tag{3}$$

Support Vector Machines: one efficient way to solve this anaphora resolution is Support Vector Machines (SVMs). The SVM will find a hyperplane that best separates the positive pairs from the negative pairs in the feature space. In a two-dimensional space, a hyperplane is a line that separates the data into different classes. In higher dimensions, it is a plane or hyperplane that does this separation. The goal of SVM is to find the optimal hyperplane that maximizes the margin between the different classes. The margin is defined as the distance between the hyperplane and the nearest data points from each class, which are called support vectors. A larger margin is associated with better generalization, meaning the model is less likely to overfit. These are the data points that are closest to the hyperplane and influence its position and orientation. Only these points are used in defining the hyperplane, which makes SVM efficient. For a set of training examples (xi,yi), where xi is the feature vector and yi is the class label:

- Find the hyperplane w · x + b = 0 that maximizes the margin.
- The optimization problem is:

$$\min_{w,b} \frac{1}{2} \| w \|^2 + C \sum_{i=1}^{n} \varepsilon_i \tag{4}$$

Subject to:

$$y_i(w.x_i + b) \geq 1 - \varepsilon_i \text{ and } \varepsilon_i \geq 0 \tag{5}$$

- ξi are slack variables that allow for misclassification.

Naive Bayes Classifier: Naive Bayes can be a simple yet effective approach for anaphora resolution when combined with careful feature engineering and enough annotated data. While it has limitations due to its independence assumption, its probabilistic nature and efficiency make it a valuable tool in natural language processing tasks. Bayes' theorem provides a way to update the probability estimate for a hypothesis as more evidence or information becomes available. It is stated as:

$$P(C \mid X) = \frac{P(X \mid C).P(C)}{P(X)} \tag{6}$$

P(C|X) is the posterior probability of class C given feature vector X. P(X|C) is the likelihood of feature vector X given class C. P(C) is the prior probability of class C. P(X) is the evidence or marginal likelihood of feature vector X.

7. Evaluation

The Quranic corpus was used as the subject of experiments to resolve Arabic anaphora. We used 10-fold cross-validation to assess each method in order to gauge the overall performance of our system. Two metrics that are frequently used to determine the worth of outcomes are precision and recall. Recall is a measure of completeness, whereas precision is a measure of accuracy. The total of pairs that were mistakenly identified as coreferent—true positives and false positives—and the total of pairs that were correctly categorized as negatives—true positives and false negatives. Scores for memory and precision are sometimes added together to create a single measurement known as the F-measure, which computes recall and precision. The following describes these metrics:

$$precision = \frac{TP}{(TP + FP)} \tag{7}$$

$$recall = \frac{TP}{(TP + FN)} \tag{8}$$

$$F_1 = \frac{2 * \text{recall} * \text{precision}}{(recall + precision)} \tag{9}$$

8. Comparison of Experiment Results

The Myanmar novel data set was used to test the classifiers (kNN, Decision Tree, SVM, and Naïve Bayes) for overall anaphora resolution performance. Several experiments were conducted to empirically compare seven different features. In each primary experiment, a set of features was applied with nearly all other features using one of the four classification techniques. This experiment examined several feature types and investigated how they affected the classification approach's performance. The objective was to create a more accurate classification process by effectively integrating several feature sets and classification algorithms.

Method	Precision	Recall	F-measure
kNN	48 %	37 %	42 %
Decision Trees	47 %	37 %	41 %
SVM	53 %	41 %	46%
Naïve Bayes	48 %	37 %	42 %

Table 2. Performance Comparison

Using various feature sets, the precision, recall, and F-measure of each model were used to evaluate its performance. Although the F-measures of kNN and Naïve Bayes classifiers are the same value, their performance of each pronoun type are varied as below:

Type of Classifier	Accuracy	Nominative & Objective	Reflexive	Possessive
KNN	Precision	46 %	44 %	52 %
	Recall	30 %	0.29 %	17 %
	F-measure	36 %	0.57 %	26 %
C4.5	Precision	48 %	33 %	47 %
	Recall	32 %	0.22 %	15 %
	F-measure	38 %	44 %	23 %
SVM	Precision	55 %	44 %	47 %
	Recall	37 %	0.29 %	15 %
	F-measure	44 %	58 %	27 %
Naive Bayes	Precision	50 %	44 %	45 %
	Recall	33 %	0.29 %	15 %
	F-measure	40 %	58 %	23 %

Table 3. Performance of Classifiers based on Pronoun Types

Though the system has solved that problem by using phrase segmentation, it still cannot correctly tag the entity of the sentence since the phrase segmentation of this system did not consider all of Myanmar prepositions as marker. The system may increase in performance if the tagger could correctly tag the phrases and recognize the name entities. Therefore, the POS tagger which can recognize the name entity is very essential for all anaphora resolution systems.

Every model has a distinct effect, as indicated by the findings of both the combined approach and the individual classifiers. Table 2. shows that the best result is obtained using Support vector machine (SVM) have F-measure over 46%. The resolution of pronoun in Myanmar language by using morphological analysis and machine learning algorithms is relatively low in success rate because the real Myanmar novel sentences are used and do not train to get efficient results by converting sentences.

9. Conclusion

In order to improve the Myanmar Anaphora resolution system, this research suggested comparing four machine learning algorithms: KNN, Decision Tree, SVN, and Naive Bayes. It is predicated on a number of linguistic and computational characteristics. The results of the studies demonstrated that a few features affect the model's performance. The findings show that, in comparison to the other feature sets, the candidate-relevant attributes significantly impact the model's performance. The findings of this study will be useful in developing an improved tool for Myanmar language anaphora resolution.

Future efforts will focus on finding resolves for every kind of pronoun. After that, we want to broaden our research and switch from supervised to deep learning techniques. Furthermore, if we consider other kinds of anaphors to be lexical anaphors, that will be quite fascinating.

Acknowledgments

I am profoundly grateful to many individuals and organizations whose support and contributions were instrumental in the completion of this research. Words cannot express my deep gratitude for the immense patience and insightful criticism of my professor and committee chair. Their generous knowledge and support were crucial in enabling me to embark on this journey.

I am also thankful for my coworkers and classmates, whose moral support, late-night feedback sessions, and help with editing were invaluable. Additionally, I extend my appreciation to the university's librarians, research assistants, and study participants for their inspiration and influence. Lastly, I want to express my sincere gratitude to my family for their unwavering love and support over my academic career.

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.

[8] Abdullatif Abolohom, Nazlia Omar, Sebastião Pais, João Cordeiro, “A Comparative Study of Linguistic and Computational Features Based on a Machine Learning for Arabic Anaphora Resolution”, 5th International Conference on AI in Computational Linguistics, 2021.

[9] Ram, R. Vijay Sundar, and Sobha Lalitha Devi. (2013) “Pronominal resolution in tamil using tree crfs.” In International Conference on Asian Language Processing, pp. 197–200.

[10] Wohiduzzaman, Kazi, and Sabir Ismail. (2018) “Recommendation System for Bangla News Article with Anaphora Resolution.”In Proceedings of the the International Conference on Electrical Engineering and Information & Communication Technology (iCEEiCT), pp. 177–182.

[11] Protopopova, E. V., A. A. Bodrova, S. A. Volskaya, I. V. Krylova, A. S. Chuchunkov, S. V. Alexeeva, V. V. Bocharov, and D. V. Granovsky. (2014) “Anaphoric annotation and corpus-based anaphora resolution: an experiment.”In Proceedings of 20the International Conference on Computational Linguistics Dialog, pp. 562–570.

Authors Profile

Ms. Khin Theink Theink Soe
Contact Information:

- **Email:** khintheinktheinksoe@ucsy.edu.mm
- **Phone:** (+959) 740948266
- **Address:** No.(4) Main Road , Shwe Pyi Thar Township, Yangon, Myanmar

Education:

- **Master of Computer Science**
  University of Computer Studies, Maubin
- **Bachelor of Computer Science**
  University of Computer Studies, Maubin

Publications:

- Khin Theink Theink Soe, Tin Htar Nwe, and Khin Thandar Nwet, Detection and Resolution of Anaphora in Myanmar Text using A Hybrid Approach, ICCA 2018.
- Khin Theink Theink Soe, Tin Htar New, and Khin Thandar Nwet, Anaphora Resolution for Myanmar Text Using K-Nearest Neighbor Algorithm, ISBN: 978-981-14-1455-8, DOI: 10.18178/wcse.2019.03.016, ICFCC 2019.

Professor, Dr. Khin Mar Soe
Contact Information:

- **Email:** khinmarsoe@ucsy.edu.mm
- **Address:** No.(4) Main Road , Shwe Pyi Thar Township, Yangon, Myanmar

Education:

- B.C.Sc.(hons.) , M.C.Sc. , Ph.D(IT) , DAC (India)