ISBN 978-981-14-1455-8
Proceedings of 2019 The 11th International Conference on Future Computer and Communication (ICFCC 2019)
Yangon, Myanmar, February 27-March 1, 2019, pp. 90-94

Anaphora Resolution for Myanmar Text Using K-Nearest Neighbor Algorithm

Khin Theink Theink Soe [1+], Tin Htar Nwe [2] and Khin Thandar Nwet [3]

[1,2,3] Natural Language Processing Lab, University of Computer Studies, Yangon

Abstract. Anaphora resolution which most commonly appears as pronoun resolution is the problem of resolving references to earlier or later items in the discourse. Anaphora resolution is an active area of research, such as text mining, text summarization, dialogue interpretations, information extraction, and so on. Anaphora resolution in English and other European languages has been well done in early. But Myanmar Language has not sufficiently applied. This paper presents Myanmar anaphora resolution system by using rule-based part of speech tagging 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, machine learning, k-nearest neighbour, morphological features

1. Introduction

Anaphora resolution is the process of resolving references to an entity from the set of referring expressions or discourse. Anaphor is commonly a pronoun or referring word pointing back to previous item in the discourse. It is a pervasive phenomenon in natural language communication. Anaphors typically refer back to other constituents in the same sentence, or to constituents in earlier utterances in the discourse. Discourse is a group of collocated and related sentences. Antecedent can be noun, noun phrase, verb phrase or clause in the discourse. Anaphora resolution is a process of determining the antecedent of an anaphor and the subsequent replacement of the anaphor by its antecedent. Computer cannot easily understand the natural language 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 helps to fully and correctly understand the text. Myanmar language exhibits the characteristics of an agglutinative language.

This study will develop an efficient method for anaphora resolution for Myanmar language, mainly pronoun resolution for Myanmar grammar. There are two main steps in this anaphora resolution system. First step is the identification of pronouns or noun phrase from the set of referring expressions or discourses of Myanmar novels and short story. Secondly, antecedent-anaphora relations are identified and then selects the correct and possible antecedent or candidate for the pronoun.

Data for this system is collected from Myanmar novel "May" narrative type of author "Dagon Taryar". 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.

This paper presents anaphora resolution system for resolving Myanmar pronominal anaphora. The collection of anaphora and respective possible antecedents was identified in a rule-based manner and

[+] Corresponding author.
E-mail address: khintheinktheinksoe@ucsy.edu.mm

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 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) proposed a system 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 obtains a substantial accuracy rate of 80% [2].

A. Abolohom and N. Omar proposed a hybrid approach that combines a rule-based manner and machine learning based 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].

An anaphora resolution system for Russian language [4] uses the formal rules to obtain a number of features, the Extra Trees machine learning algorithm and the Balance Cascade algorithm for working with imbalanced learning sets and a neural network algorithm SyntaxNet to analyze the syntactic context. This presents the imbalance issue of the anaphora nature, and also solves by using The Balance Cascade algorithm.

3. Proposed System

This section presents the architecture for pronominal anaphora resolution system for Myanmar Text. This system uses the 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, "May" novel. This dataset presents a greater challenge due to a mixture of nominal and pronominal anaphors and a greater range of confounders. The proper subject-object identification is needed for an efficient anaphora resolution system. The proposed training data set 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 [5]. 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.

As shown in figure 1, this system has two main parts. Firstly, preprocessing step has collected raw data, morphological feature identification and pronoun and noun phrase identification using rule-based system. The next step, every pair of pronoun and noun-phrase computes to get similarity score using k-Nearest Neighbor algorithm. Then, it will output the most similarity anaphora and antecedent pair.

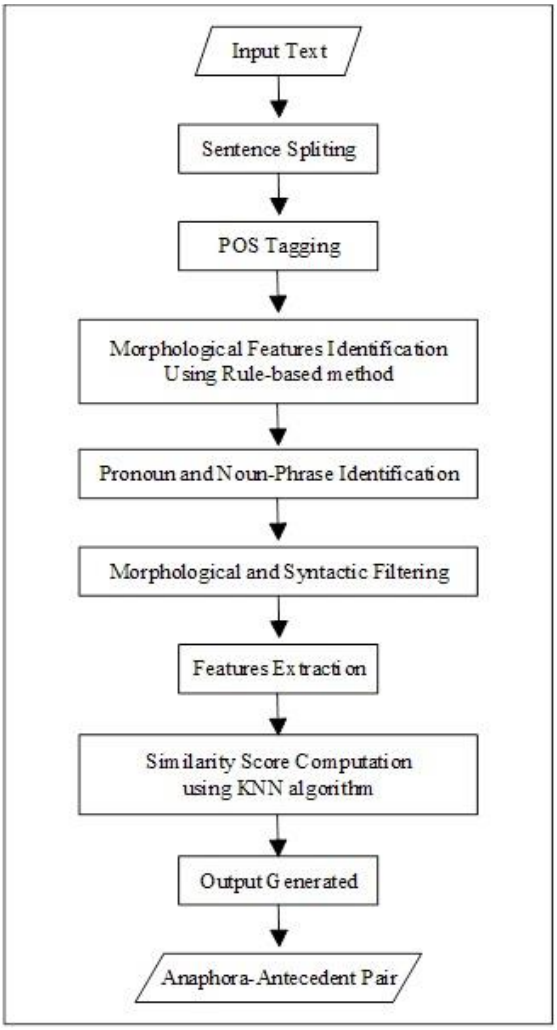

Fig. 1. Architecture of proposed system

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 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 of each word in a sentence, such as noun, verb, adjective and so on. 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, basic and standard tagging, and chunking module. This tagger uses 14 for basic tags and 9 for standard tags [5].

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 includes the pronouns and noun-phrases with morphological features to match the pairs of pronominal anaphora and its antecedent candidates.

3.2. Morphological and Syntactic Filter

In the morphological filtering step, we can reduce the large size of pairs (pronoun and a potential

antecedent). For each anaphora and noun-phrase from the created list, 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. 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.

3.3. Features Extraction

Features are allocated scores and can 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. Our features are grouped into 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 value is 1. Myanmar language has some ambiguous situations 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 where the function is only approximated locally and all computation is deferred until classification. It 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 spaces, where each dimension represents the features of the data. 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 [6]. 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 (a1, a2, ……, ap) and (b1, b2, ……, bp) is defined as

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

where d(a,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.

4. Data Used for Experiment

Data for this system is collected from Myanmar novel "May" (ဧဝ) narrative type of author "Dagon Taryar". Myanmar language has four categories such as personal pronoun, demonstrative pronoun, question pronoun and mathematic pronoun. Among them, we will resolve personal pronoun. There are 19 chapters in

this novel and we tested 19 datasets of each chapter. Each chapter has 250 personal pronouns to take anaphora resolution. Therefore, we will detect and resolve about 4000 pronouns.

5. Experiment Results

Anaphora Resolution was regarding as a classification problem. Precision, recall and F-measure were used for measuring the performance. The different values are obtained from the different datasets in Table 1.

$$\text{Precision} = \frac{TP}{TP+FP} \quad (2)$$

$$\text{Recall} = \frac{TP}{TP+FN} \quad (3)$$

$$\text{F-measure} = \frac{2xPrecisionxRecall}{Precision+Recall} \quad (4)$$

TABLE I: Comparison of Results

Data Set	Precision	Recall	F-measure
Dataset 1	0.8153	0.5672	0.669
Dataset 2	0.5924	0.8977	0.7138
Dataset 3	0.6801	0.7984	0.7345

6. Conclusion

This paper proposed Myanmar Anaphora Resolution System using Rule-based and K-Nearest Neighbor Algorithm. 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.

7. References

[1] T. Lwin, A. Thida. Myanmar Anaphora Resolution based on Centering Theory. UCSM, 2015. M.

[2] M. T. Naing, A.Thida. Pronominal Anaphora Resolution Algorithm in Myanmar Text. International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), August 2014, p.2795-2800.

[3] A. Abolohom, N. Omar. A Hybrid Approach to Pronominal Anaphora Resolution in Arabic. Journal of Computer Science 2015, May 11, jcssp. 2015.764.771.

[4] A. Kozlova, A. Svischev, O.Gureenkovo, T. Batura. A hybrid approach for anaphora resolution in the Russian language. 2017 Siberian Symposium on Data Science and Engineering, April 12-13, 2017.

[5] P. H. Myint, T. M. Htwe, N. Thein. Lexicalized HMM-based Part-of-Speech Tagger for Myanmar Language. ICCA 2012, Yangon, Myanmar, February 28-29, 2012.

[6] M. Bramer. Principles of Data Mining, Landon: Springer, 2007.