How to cite this article:

Mohd Noor, N. K., Mohd Noah, S. A., & Ab Aziz, M. J. (2020). Classification of short possessive clitic pronoun *nya* in Malay text to support anaphor candidate determination. *Journal of Information and Communication Technology, 19*(4), 513-532. https://doi.org/10.32890/jict2020.19.4.3

# CLASSIFICATION OF SHORT POSSESSIVE CLITIC PRONOUN *NYA* IN MALAY TEXT TO SUPPORT ANAPHOR CANDIDATE DETERMINATION

**[1]Noor Huzaimi@Karimah Mohd Noor, [2]Shahrul Azman Mohd Noah & [2]Mohd Juzaiddin Ab Aziz**

*[1]Faculty of Computing, Universiti Malaysia Pahang, Malaysia*
*[2]Faculty of Information Science & Technology,*
*Universiti Kebangsaan Malaysia, Malaysia*

*nhuzaimi@ump.edu.my; shahrul, juzaiddin@ukm.edu.my*

## ABSTRACT

Anaphor candidate determination is an important process in anaphora resolution (AR) systems. There are several types of anaphor, one of which is pronominal anaphor. Pronominal anaphor is an anaphor that involves pronouns. In some of the cases, certain pronouns can be used without referring to any situation or entity in a text, and this phenomenon is known as pleonastic. In the case of the Malay language, it usually occurs for the pronoun *nya*. The pleonastic that exists in every text causes a severe problem to the anaphora resolution systems. The process to determine the pleonastic *nya* is not the same as identifying the pleonastic 'it' in the English language, where the syntactic pattern could not be used because the structure of *nya* comes at the end of a word. As an alternative, semantic classes are used to identify the pleonastic itself and the anaphoric *nya*. In this paper, the automatic semantic tag was used to determine the type of *nya*, which at the same time could determine *nya* as an anaphor

**Received:** 10/4/2019 **Revised:** 28/4/2020 **Accepted:** 30/4/2020 **Published:** 20/8/2020

candidate. The new algorithms and MalayAR architecture were proposed. The results of the F-measure showed the detection of clitic *nya* as a separate word achieved a perfect 100% result. In comparison, the clitic *nya* as a pleonastic achieved 88%, clitic *nya* referring to humans achieved 94%, and clitic *nya* referring to non-humans achieved 63%. The results showed that the proposed algorithms were acceptable to solve the issue of the clitic *nya* as pleonastic, human referral as well as non-human referral.

**Keywords:** Anaphora resolution, natural language processing, Malay anaphora resolution, anaphor candidate determination.

## INTRODUCTION

Anaphora can be defined as a linguistic relation between two textual entities. It is determined when a textual entity (the anaphor) refers to another entity in the text that usually occurs before it (the antecedent) (Sukthanker, Poria, Cambria, & Thirunavukarasu, 2019). The process of determining the antecedent of an anaphor is referred to as anaphora resolution (AR). AR is important in most Natural Language Processing (NLP) applications such as question answering systems (Asao, Iida, & Torisawa, 2018), text summarisation (Antunes, Lins, Lima, Oliveira, Riss, Simske, 2018) and text classification (Saqia, Khan, Khan, Khan, Subhan, & Abid, 2018). Researchers have claimed that by implementing AR, the performance of NLP-related applications is significantly improved (Vicedo & Ferrández, 2000; Kabadjov, 2007; Nøklestad, 2009). There are three basic steps involved in AR: the first step is to determine a list of anaphor candidates, while the second step is to determine a list of antecedent candidates. The last step is to resolve the anaphor with its appropriate antecedent. There are many kinds of anaphora such as pronominal anaphora, propositional anaphora, adjectival anaphora, and modal anaphora (King & Lewis, 2018). Despite the variations of anaphora, this paper will only focus on pronominal anaphora, since it receives the most attention in linguistic and philosophical literature. In pronominal anaphora, the anaphor is a pronoun word that refers to the antecedent that can be in the same sentence or can also span in many sentences. For example, in the following sentences, the pronouns *she* and *her* refer back to the word *girl*.

> *When the girl went outside, she put on her hat. But she could still feel the cold.*

Pronominal anaphora poses a challenge in many NLP applications since the anaphor has to fulfil some grammatical agreements with the antecedent.

In the Malay language, determining the anaphor candidates for pronominal anaphora is relatively similar to English, which involves the detection and retrieval of relevant pronouns. However, in the Malay language, there exists a unique word *nya*, which is a potential anaphor. The behaviour of the word *nya* can be loosely equated with the word *it* in English. Nevertheless, it is rather challenging to identify the pronoun *nya* in Malay language since the word is attached to the end of other words such as *bukunya* (his/her book) and *keretanya* (his/her car). Therefore, there is a need to separate the word *nya* from its host word. In English, the word *it* is able to act as a word that does not refer to any entity in some situations. This situation is referred to as pleonastic (Shalom & Herbert, 1994). The same situation applies for the word *nya* where it may or may not refer to any entity in a text. Syntactic patterns are frequently used to identify the pleonastic *it*. In spite of this, such patterns cannot be applied to identify *nya* because the behaviour of *nya* depends on the host words it is attached to. This paper discusses the method to determine the use of pronoun *nya* to ensure that the right selection of anaphor candidate is achieved.

In an earlier research (Karimah, Aziz, Noah, & Hamzah, 2011), three instances where the pronoun *nya* is used have been identified: referring to humans, referring to non- humans, or not referring to any situation in a text. It is of crucial importance to correctly identify as to whether the pronoun *nya* refers to human and non-human as it will implicate the correct referral to the antecedent candidate within sentences. Consider for instance the sentences "*Ali memberi makan kucing. Bulunya lebat.*", which are translated to English as "Ali feeds the cat. Its fur is thick.". In this example, Ali (human) and *kucing* (cat) are antecedent candidates. Although humans can easily point out that *nya* refers to *kucing* (cat) and not Ali (human), machines need to understand the semantic class of *bulu* (fur), which refers to animal. Therefore, it can be inferred that *nya* refers to antecedent *kucing*, which is non-human. Such resolution is also important for NLP applications such as machine translations (Tabrizi, Mahmud, Idris, & Tohidi, 2016) in order to translate anaphoric expressions correctly into a target language.

The word *nya*, as in pronoun *it* in the English language, sometimes does not refer to any situation in a text and does not function as the referring expressions; this is usually known as pleonastic (Werlen & Popescu-Belis, 2017; Shalom & Herbert, 1994). This may turn out to be a serious problem for anaphora resolution systems. As a result, such pleonastic must be identified in any AR systems. For example, Aone and Bennett (1996), Kennedy and Boguraev (1996), and Bouzid, Trabelsi, & Zribi (2017) manually removed pleonastic pronouns, while other researchers employed machine learning techniques on the available corpus (Yifan, Musilek, Reformat, & Wyard-Scott, 2009). Anaphora resolution for Malay text can be seen in the works by Fazal Mohamed (2006), Yap (2011), and Xian, Saloot, Ghazali, Bouzekri, Mahmud, and Lukose (2016). However, these works did not specifically address the issues related to the word *nya* in AR. *nya* is widely used in Malay text and the study by Xian et al. (2016) showed that *nya* is the most frequently used pronoun in the Dewan Bahasa dan Pustaka (DBP) corpus. This paper, therefore, aims to further explore the behaviour of the word *nya* and propose algorithms that can automatically classify the type of *nya* in Malay text.

This paper is organised into the following sections. The second section provides an overview of the related research in this area, and different usages of the pronoun *nya* in Malay text are discussed in following section. The fourth section describes the method utilised to determine the different usages of the pronoun *nya* and the retrieval of the pronoun *nya* in an unstructured text. The result and discussion are presented in the fifth section, while the conclusion and suggestion for future work are given in the last section.

## RELATED WORKS

As mentioned earlier, AR is the problem of resolving references to earlier or later items in the discourse. These items are usually noun phrases representing objects in the real world called referents; however, they can also be verb phrases, whole sentences or paragraphs. Works on AR have been extensive for the English language. Conversely, other languages such as Chinese (Zhao, Liu, & Yin, 2017; Kong, Zhang, & Zhou, 2019), Arabic (Hammami & Belguith, 2018), and Hindi (Sikdar, Ekbal & Saha, 2016) have been gaining interests of the NLP research community. In this section, some related works on AR for the English language are provided, which particularly focused on the pleonastic pronouns. This study then focuses on related works for the Malay language.
Resolution of Anaphora Procedure (RAP) is an approach for anaphora resolution proposed by Shalom and Herbert (1994). The system developed by them can identify the third pronoun resolution and pleonastic pronouns, where the pronouns do not have any referent entity. The pronoun *it* is commonly used as the pleonastic pronoun in English (Loaiciga, Guillou, & Hardmeier, 2017; Qiu, Kan, & Chua, 2004; Yaneva, Ha, Evans, & Mitkov, 2018). The pleonastic *it* basically appears with a modal adjective such as "*…it is important to…*" or in its passive participle, *it* comes together with a cognitive *verb* like "*…it is recommended that...*". To identify a pleonastic in a sentence, RAP not only relies on the lists of modal adjectives and cognitive verbs, but also on the syntactic pattern as follows (Qiu et al., 2004; Shalom & Herbert, 1994):
i. it is Modaladj that S,
ii. it is Modaladj (for NP) to VP,
iii. it is Cogv-ed that S,
iv. it seems/appears/means/follows (that) S,
v. NP makes/finds it Modaladj (for NP) to VP,

vi. it is time to VP, and
vii. it is thanks to NP that S,

where Modaladj stands for a modal adjective and Cogv-ed stands for the passive participle of a cognitive verb. The approach is a part of a larger system that aims to resolve the pronominal anaphora. Subsequently, individual evaluations of *it* are not discussed in this paper as only the overall system is evaluated.

The study by Paice and Husk (1987) that used bracketing patterns such as: "*it ..... to*" and "*it ..... who*" was different from the study carried out by Shalom and Herbert (1994). However, both approaches relied on a pattern to determine pleonastic pronouns. Paice and Husk's (1987) approach produced impressive results with 93.9% accuracy in determining pleonastic without using parts-of-speech tagging or parsing.

In another study, Bergsma, Lin, and Goebel (2008) used N-gram by counting the web-scale data to detect pleonastic pronouns. First, the pattern was created by looking at the context of words around *it*. For example, the sentence '*it is able to*' was translated into '_ *is able to*' pattern. The patterns were comprehensive and were represented using several rules for detecting inconsistent verbs and common abbreviation. The patterns were then tested using the Google N-gram data developed by Brants and Franz (2009) to determine words that fill the patterns.

In Malay text, *nya* and *ia* are two common pleonastic pronouns; therefore, determining the use of the word *nya* is non-trivial. There is limited or no research efforts to investigate the pleonastic within the NLP research area, except for the study conducted by Karimah et al. (2011), where experiments were conducted to determine the usage of *nya* by utilising 60 Malay texts in a semi-automatic form. The method involved tagging the part of speech (POS) of the sentence and manually classifying all the words attached to *nya* either as human, animal, position, direction or location. A set of rules were developed to detect *nya* as either anaphor or pleonastic. In the case of anaphor, the rules were able to detect whether *nya* refers to one of the three categories of anaphors, which are human, non-human, and not-referral word.

Nevertheless, from the linguistic community, two research works of Fazal Mohamed (2006) and Yap (2011) are worth mentioning. Fazal Mohamed (2006) researched on the clitic pronoun *nya* and focused only on the verb host word. In his study, he used syntactic analysis, which included the movement hypothesis and in-situ hypothesis, in order to identify *nya* as a clitic word that is attached to transitive and semi-transitive verbs. He did not, however, mention how to classify the clitic *nya* referring to whom. He claimed that the movement analysis was able to explain enclitisation of the clitic *nya* in a verb phrase. The work of Yap (2011) established whether the clitic *nya* is a referral word or otherwise. Again, in this study, the author did not focus on whether the clitic *nya* is referring to humans or non-humans. He used robust grammaticalisation patterns that were consistent with typological observations reported elsewhere, including a robust nominal/pronominal, nominalise, and stance marker development.

Another important work was by Xian et al. (2016) that used the maximum entropy model and Random Forest classifier for pronominal AR in 100 articles of which half of them were news articles. However, the work did not specifically address the behaviour of *nya* and semantically classify the words attached to *nya*. Classifying the words attached to *nya* is important in AR in order to remove pleonastic words. Furthermore, it is important to categorise the word (or host word) attached to *nya* either as human or non-human due to fact that in AR, only humans will be resolute.

## PRONOUN *nya* IN MALAY TEXT

In Malay texts, the usage of the pronoun *nya* is mostly used to represent humans, events, and organisations. It is often used as a third-person pronoun or to show the concept of belonging. The pronoun *nya* always appear at the end of another word, such as *abang**nya*** (his/her brother). However, there are cases where *nya* that appears in a word is actually part of a word and does not act as the pronoun *nya* as in the word *tanya* (ask). Sentence 1 shows the word *nya* attached to the word *wang* (money) as a third-person pronoun that refers to *Ali* where *Ali* is a proper name for humans.

Sentence 1:
***Ali** hendak membeli buku, tetapi wang**nya** tertinggal di rumah.*
**Ali** wanted to buy a book, but he left his **money** at home.

Nevertheless, there are some cases of *nya* that do not reflect the third-person pronoun for humans, but refers to animals, events, and objects. The example of this case is shown in Sentence 2.

Sentence 2:
***Kucing saya** bernama Putih. Bulu**nya** panjang dan tebal.*
**My cat's** name is Putih. **It** has long and thick fur.

Sentence 2 shows that word *nya* is likely used as a pronoun for the cat. Conversely, this is against what has been stated by Nik Safiah, Farid, Hashim, and Abdul Hamid (2008) and Asmah (2009), where the usage of the word *nya* is only as a third-person pronoun. However, based on the present research, there are various usages in texts where *nya* refers to other than humans. Therefore, in such cases, the term non-human referral word will be used to refer to *nya* that is not referring to humans. Apart from that, Nik Safiah et al. (2008) demonstrated two situations whereby the word *nya* does not refer to any object. The first situation is where it is used to stress an important issue highlighted in a sentence. The following example illustrates such a situation. The word *sesungguhnya* in the sentence stresses that the issue mentioned in Sentence 3 is a serious condition.

Sentence 3:
*Lapisan ozon sesungguh**nya** semakin menipis.*
Indeed, the ozone layer is thinning.

The other situation is where the word *nya* changes the POS tag of a word into a noun as shown in the following example, Sentence 4. The word *laju* in the sentence is originally an adjective. However, by adding the word *nya* at the end of the word *laju*, it becomes a noun.

Sentence 4:
*Laju**nya** ialah 100 km sejam*
**The speed** is 100 km per hour

The method to determine the different usages of *nya* is needed in order to process the Malay text for supporting anaphor resolution. Some approaches used in linguistic are by tagging of phrases (i.e. verb, noun, and adjective) that come together with the phrase *nya* and cross bridging of the sentence or word itself to look at the semantic sentence. However, there is no formal discussion on how to solve these cases in computational linguistic.

In a different situation, the word *nya* is part of the word itself. Some examples are shown in Table 1.

**Table 1**

Words end with nya that represent the word itself

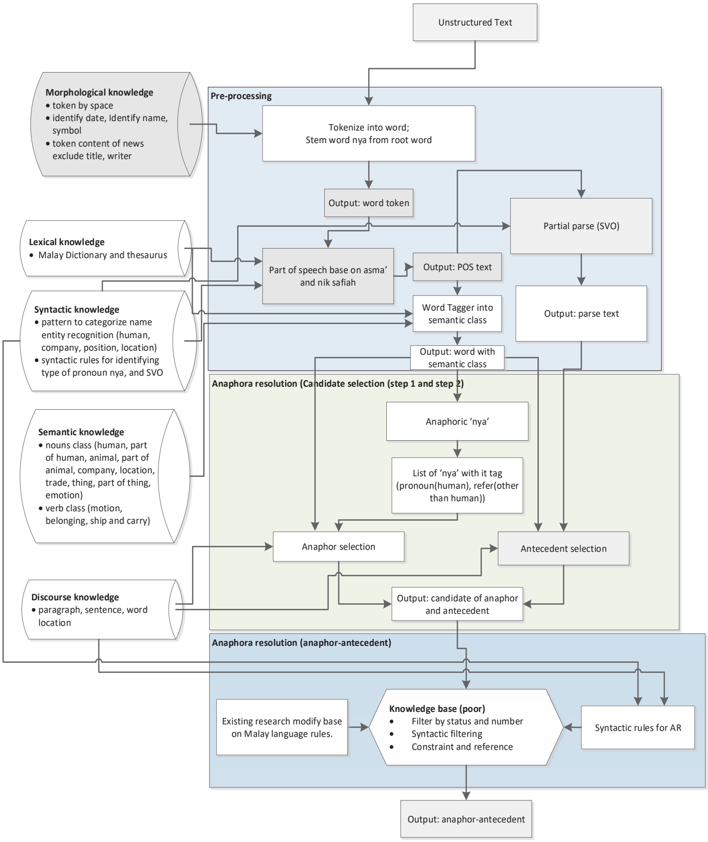
| Word Ends With *nya* | Formal Translation |
|---|---|
| Ta*nya* | Ask |
| Ha*nya* | Only |
| Pu*nya* | Have/has |
| Empu*nya* | Owner |

In determining whether the word *nya* in the text is either part of the complete word itself or not, a gazetteer or dictionary that contains such words is required. If the phrase that has the word *nya* does not exist in the dictionary, then the process to separate the word *nya* from the original word is executed.

The aforementioned situations show the various implications of the word *nya*. Therefore, the inability to correctly identify the usage of *nya* has a significant implication in the process of anaphora resolutions for the Malay language. The following sections describe the approach utilised in determining *nya* and an anaphor candidate and proposes the use of semantic classes in assisting such detection.

## ANAPHOR CANDIDATE DETERMINATION METHOD

Based on the behaviour of *nya* discussed in the previous section, there are two important questions that need to be addressed. The first question is how to tokenise host words and pronoun *nya* such as the word *kereta***nya** (his/her car). The second question is how to determine whether *nya* represents pronoun or pleonastic. Figure 1 shows the proposed method for dealing with these two issues.



Unstructured Text
Morphological knowledge
• token by space
• identify date, Identify name, symbol
• token content of news exclude title, writer
Pre-processing
Tokenize into word;
Stem word nya from root word
Output: word token
Partial parse (SVO)
Lexical knowledge
• Malay Dictionary and thesaurus
Part of speech base on asma' and nik safiah
Output: POS text
Word Tagger into semantic class
Output: parse text
Syntactic knowledge
• pattern to categorize name entity recognition (human, company, position, location)
• syntactic rules for identifying type of pronoun nya, and SVO
Output: word with semantic class
Anaphora resolution (Candidate selection (step 1 and step 2)
Anaphoric 'nya'
Semantic knowledge
• nouns class (human, part of human, animal, part of animal, company, location, trade, thing, part of thing, emotion)
• verb class (motion, belonging, ship and carry)
List of 'nya' with it tag (pronoun(human), refer(other than human))
Anaphor selection
Antecedent selection
Discourse knowledge
• paragraph, sentence, word location
Output: candidate of anaphor and antecedent
Anaphora resolution (anaphor-antecedent)
Existing research modify base on Malay language rules.
Knowledge base (poor)
• Filter by status and number
• Syntactic filtering
• Constraint and reference
Syntactic rules for AR
Output: anaphor-antecedent

**Figure 1.** *Malay AR architecture.*

As can be seen in Figure 1, the process to determine the type of word *nya* occurs at the Pre-Processing and Anaphora Resolution for Candidate Selection stages, involving three types of knowledge base (KB), namely morphological knowledge (MK), lexical knowledge (LK), and semantic knowledge (SK). MK and LK are used when determining the word *nya* either as the word itself or as pronoun and pleonastic candidate, while SK is used to determine whether the word *nya* is human_referrer, non-human_referrer or as pleonastic.

MK is used to tokenise sentence into words, after which the phrase *nya* is determined whether it is a separate token. Once the phrase *nya* is identified as a separate token, the process to separate the phrase *nya* from its host word is carried out using MK and LK. The process is needed to ensure that the phrase *nya* in a sentence does not belong to the word itself, such as in the words *tanya* (ask), *hanya* (only), and *punya* (belonging), and therefore should not be categorised as a pronoun or pleonastic. Appropriate algorithms are needed during the tokenising process to separate the host word from *nya*. In this case, a list of index words retrieved from a Malay dictionary is used. Algorithm 1 shows the algorithm on how to determine and separate the host phrase from the phrase *nya*.

**Algorithm 1: Algorithm to separate word *nya* from its host word**

***Declaration 1:***
*A(x) is a set of words that use the word nya as its own word*
*x is a list of words that use the word nya as its own word*
*z is nya*
*i is a host word*
*y is a word that needs to be processed*
*prn is nya as a separate word from phrase*

```
Input: $y = i \cup z$
If $y \in A(x)$ then
    $y = i \cup z$
If else i.e. $y \notin A(x)$ then
        $y=i$
        $prn=z$
```

Let us consider the input word ***adik**nya* (her/his brother), which is checked with the element in A(x). If the input word is listed in the dictionary, then the phrase *nya* is declared as part of a word. Otherwise, the host word and phrase *nya* will be separated into different words. In the case of ***adik**nya*, the word does not match any of the words in the list of A(x), then *nya* will be separated from the host word *adik*. At the same time, the host word is labelled with POS using the lexical database adapted from Dewan Bahasa dan Pustaka (2010). The next stage is to determine the semantic class of the host word. The semantic class is divided based on the type of word as the word can be classified as a type noun, verb or function word. The semantic classification of the host word is described in Tables 2 to 5. The semantic class is determined by looking at the meaning of the word and compare it with common words associated with each semantic class. In this case, a Malay dictionary that contains examples of sentences for each word is used (as shown in Figure 2). For an example, the list of common words associated with the semantic class human is *orang* (people), *penduduk* (residents), and *masyarakat* (community). If a match is found in the list, then the host word is classified as the corresponding semantic class. However, the dictionary also has examples of sentences with elements that are difficult to classify into a specific class. Therefore, to filter out these cases, another word index, called the filter_index_word, is used. The filter_index_word, among others, consists of symbols such as '~', as well as words such as *laluan* (path) and *atau*. (or). For example, the following excerpt (Figure 2) from the dictionary shows the meanings and usage examples of the word *laluan* (which can mean road or space for walking). In the examples, the word *laluan* can be summarised to mean: 1) road; 2) space for walking or crossing; and 3) movement such as movements for car and engine or even clock.

1. tempat lalu lintas (utk orang, kenderaan, dll): belanja membuat ~;
2. tempat yg harus dilalui atau dilintasi supaya sampai ke tempat (ruang) tertentu, laluan: tolong tunjukkan ~ ke rumah datuk penghulu; ~ laut laluan yg diikuti di laut utk pergi ke sesuatu tempat.
3. gerakan (kereta, enjin, dll): kereta itu sangat laju ~nya; jam itu cepat~nya;".

*Figure 2*. *Example of content used in a dictionary.*

The meaning of the word *laluan* contains the phrase '*utk orang*' (for people) indicating that the word *laluan* is used for humans. Intuitively, it can be concluded that *laluan* is a 'thing' used by people. With the filter_index_word, the phrase "*utk orang*" will be seen as an indicator signifying that the referred word is among words that do not refer to the human class. Consequently, *laluan* is not classified as a human semantic class.

After determining the class of the host word, next is the process of classifying the type of *nya*. The process, which involves SK, is performed using the algorithm shown in Algorithm 2. In this case, each word will be classified according to the correct POS. In this study's approach, the POS used are verb, noun, adjective, and function words. Function words are words that are used to perform certain functions in building sentences, clauses, and phrases. In other words, function words are the words used to make sentences grammatically correct. Pronouns, determiners, prepositions, and auxiliary verbs are examples of function words. The POS is used to determine the different types of usage of the word *nya*. In the example of the word ***adik**nya* (his/her younger brother/sister), the host word *adik* is classified as a noun and belongs to the human class. Therefore, *nya* will be tagged as a pronoun that refers to humans. The approach or guide used to determine human referrer based on the POS tags of noun, verb and function words is shown in Tables 2, 3, and 4, respectively.

**Algorithm 2: Algorithm for determining type of word *nya***

```
Declaration 2:
x set of tag use {Noun, Verb, Function_word}
k is a word that has the word nya at the end
z is nya
c is a semantic class
Cnh is a set of class for x that has Noun tag and refers to human
Cvh is a set of class for x that has Verb tag and refers to human
Ctgsh is a set of class for x that has Function_word tag and refers to human
Cnn-h is a set of class for x that has Noun tag and refers to non-human
Cvn-h is a set of class for x that has Verb tag and refers to non-human
 If $k \cup z$
 If $k(x) = NOUN$ and $k(c) \in c(Cnh)$
    $or\ k(x) = VERB$ and $k(c) \in c(Cvh)$
    $or\ k(x) = Ktgs$ and $k(c) \in c(Ctgs)$
      $z$ = human referral
 else if $k(x) = NOUN\ and\ k(c) \in c(Cnn - h)$
    $or\ k(x) = VERB\ and\ k(c) \in c(Cvn - h)$
      $z$ = non-human
 else
      $z$ = pleonastic
```

Table 2 shows the categories of nouns that relate *nya* as a human referrer, which includes *humans*, *part of humans*, *object*, *event*, *company*, *action*, *emotion*, and *animal*.

**Table 2**

Categories of nya attached to nouns referring to humans

| Categories | Description | Example |
|---|---|---|
| human/ part of humans | All words related to humans including part of humans. | ***rakan****nya* (his/her friend) |
| object | All things including food and drink related but not involved with part of a thing. | ***rumah****nya* (his/her house) |

(continued)

| Categories | Description | Example |
|---|---|---|
| event | Something that relates to an event. | ***perkahwinan**nya* (his/her marriage) |
| company | All words related to a company or an institution. | ***kementerian**nya* (his/her ministry) |
| action | Word that shows the action taken by someone. | ***penghakiman**nya* (his/her judgment) |
| emotion | Word that shows the condition expressed by someone. | ***kemarahan**nya* (his/her anger) |
| animal | All thing related to animals but not involved with part of the animal. | ***kucing**nya* (his/her cat) |

For example, assume the word ***anak**nya* (her/his child) shown as an example in Sentence 5. *Anak* is a noun that matches the category of humans; and therefore, *nya* is categorised as a pronoun that refers to humans.

Sentence 5:
*Ahmad pergi ke taman bersama **anak**nya.*
Ahmad went to a park with his child.

Table 3 describes four categories that illustrate *nya* **attached to a verb as referring to humans. As an example, consider the sentence '***Jawatan itu telah lama dipegangnya.***' (The position has long been held by him). In this case, the word** ***dipegangnya*** **is a verb and it refers to the semantic class position. Therefore,** ***nya*** **in this situation is a pronoun that refers to humans.**

**Table 3**

Categories nya attached to verbs referring to humans

| Categories | Description | Example |
|---|---|---|
| Event | The verb that shows the event | *Kes pembunuhan itu,*<br>*tidak **mengaitkannya** sebagai suspek*<br>(The murder case does not c onsider as the suspect.) |

(continued)

| Categories | Description | Example |
|---|---|---|
| Motion | Word that shows movement | *Ketika dalam perjalanan ke masjid, secara tiba-tiba kereta itu* ***melanggarnya****.* (While on the way to the mosque, he was suddenly hit by the car.) |
| Position | Word that shows position | *Jawatan pangarah* ***dipegangnya*** *sejak tahun lalu.* (He **holds the** director position since last year.) |

In order to assign a semantic class for those words tagged with the function words, the words must be categorised either as location or direction as shown in Table 4.

**Table 4**

Classes that are used to determine nya as a human referrer for function word

| Classes | Description | Example |
|---|---|---|
| Direction | Shows direction such as *oleh* (from), *terhadap* (against), *untuk* (for), *kepada* (to), *daripada* (from) | *Surat itu ditulis* ***olehnya****.* (The letter was written **by him**.) |
| Location | Shows location such as *atas* (on) | *Kejadian itu memberi kesan yang besar ke* ***atasnya****.* (*The incident was a major impact* ***on him****.*) |

Basically, the index word can be used for determining semantic class for the direction and location classes because the list of the words is not huge. For example, the sentence, *Ali menerima bungkusan itu daripadanya* (Ali received the parcel from him/her). In this case, *nya* refers to someone outside the context. However, *nya* is classified as human referrer because of the word that comes with it (*daripada*) is classified as a *function word* and labelled as location.

The categories that are used to determine non-human referrer for words that are attached to *nya* and classified as noun and verb are shown in Table 5. Words that do not belong to any of these categories and other categories described in Tables 2, 3, and 4 are considered as pleonastic.

**Table 5**

Categories for determining nya as non-human referrer for verb and noun

| Word type | Categories | Description | Example |
|---|---|---|---|
| noun | Part of a thing | Anything that shows part of a thing, such as *pintu* (door), *tombol* (knob) | ***Pintu****nya* (its door) |
| | Part of an animal | Anything that shows part of an animal, such as *sayap*(wings) and *bulu* (feather) | ***Sayap****nya* (its wings) |
| | Position | Anything that is related to position | ***Timbalan****nya* (his/her deputy) |
| | ownership | Anything that shows belonging to a position, company, committee | ***Pihak****nya* (his/her side) |
| verb | transaction | Action that shows transaction between two people | *Buku tersebut telah dihantar melalui syarikat penghantar dan mereka telah* ***menghantarnya*** *semalam*.<br>(The book was sent via a courier company and they have **sent it** yesterday.) |

As stated in the discussion by Asmah (2009), human referrer function words are subdivided into sub-classes based on the meaning of words that came together with the word *nya*. The names of the classes have been finalised using statistical and semantic role approaches, as discussed by Daniel and Daniel (2002).

## MALAY AR EVALUATION

The experiments involved 66 local news text collections and consisted of 382 *nya* words. The usage of word *nya* in the data set is approximately 61.88% of the total of the third-person pronoun. The evaluation aims to evaluate the ability of the proposed approach to detect the type of *nya* either pleonastic or anaphor as to include those that refer to humans or non-humans.

The result of the detected and retrieved categories usage of *nya* is benchmarked against human classification. This study used the standard precision (*P*), recall (*R*), and *F*-measure metrics to compare and analyse the result, which can be defined by Equations 1, 2, and 3 as follows:

$$Precision\ (P) = \frac{S}{S+N} \tag{1}$$

$$Recall\ (R) = \frac{S}{S+I} \tag{2}$$

$$F-measure = 2\frac{P \times R}{P+R} \tag{3}$$

where *S* represents the relevant word *nya* that has been resolved into appropriate categories, *N* is a set of retrieved irrelevant word *nya*, and *I* is the non-retrieved relevant word *nya*. The aim of the evaluation is to measure the performance of retrieving the word *nya* and to automatically classify the type of *nya* usage. Table 6 shows the result for the total retrieved *nya* occurring in the 66 Malay news texts, whereas Table 7 illustrates the results in terms of precision, recall, and F-measure.

**Table 6**

Results of the experiment

| Elements | Total number in the corpus (benchmark) | Total number correctly classified | Total number incorrectly classified | Total number missed (not detected) |
|---|---|---|---|---|
| *nya* as a separate word | 385 | 383 | 0 | 2 |
| *nya* as pleonastic | 74 | 72 | 18 | 2 |
| *nya* referring to human | 277 | 255 | 10 | 22 |
| *nya* referring to non-human | 34 | 19 | 7 | 15 |

**Table 7**

The precision, recall, and f-measure for each type of word nya

| Elements | Precision (%) | Recall (%) | F-measure (%) |
|---|---|---|---|
| *nya* as a separate word | 100% | 99% | 100% |
| *nya* as pleonastic | 80% | 97% | 88% |
| *nya* referring to human | 96% | 92% | 94% |
| *nya* referring to non-human | 73% | 56% | 63% |

The results in Table 7 showed that the proposed algorithm could perfectly recognise the word *nya* as a separate element achieving 100% precision and 99% recall values. Furthermore, the proposed approach displayed good capability for classifying the usage of *nya* as either referring to human or pleonastic. As can be seen, the F-measures for classifying each of this category were 94%, and 88%, respectively. However, for *nya* referring to non-human, only 63% was achieved.

The results of Tables 6 and 7 indicated how MK could separate the word *nya* with 99% accuracy. MK that used the index word performed better in recognising the word *nya*. However, in some cases, the use of MK by looking at how the word was constructed needs to be considered, such as in determining *nya* in the word *ditanya (asked)*. The word consists of the word '*di*' and the word '*tanya*'; therefore, the word '*ditanya*' needs to be separated into '*di*' and '*tanya*' before being checked into the index data. Out of the 285 words with *nya*, two were unable to be correctly identified.

In terms of pleonastic *nya*, the 97% recall value showed the ability of the proposed rule-based methods to successfully retrieve the 72 clitic *nya*. However, the 80% precision values indicated that some of the retrieved clitic *nya* were not relevant. To be more accurate, out of the 90 total retrieved clitic *nya*, 18 were not relevant. The 18 non-relevant clitic *nya* were host words that could not be classified to any semantic class and were not tagged as a verb, noun or function word. This is due to the limitation of the dictionary, as some words were not tagged accordingly.

Some of the results for the word *nya* as human referrer were over retrieved (i.e. about ten), while some other relevant words were not retrieved, which were 22. Nevertheless, the achievement of 96% and 92% of the precision and recall values, respectively, indicated that the proposed approach was reliable in performing such a classification task.

In terms of *nya* that refers to non-human, out of the 34 benchmarks, 19 were correctly retrieved and classified and 7 were wrongly classified. This resulted in 73% and 56% of precision and recall, respectively. This occurred because of the supporting tool that was used to label the semantic class content as a considerable amount of information could create difficulties to process.

The results showed that the proposed algorithm and enhancement of semantic class label could determine the type of *nya* before going with an anaphora resolution. The automatic semantic class label could determine the usage type of clitic *nya*. However, in some host words, there was a mislabelling of semantic class, and some of them had more than one semantic class. This happened because the lexical used not only contained the meaning of the word but also a few sentence examples. Therefore, constructing the semantic labelling became more challenging.

## CONCLUSION

The pleonastic *nya* in the Malay language cannot be determined using a syntactical pattern similar to determining the pleonastic *it* in English. As the pronoun *nya* always occur at the end of another word, the semantic class tag for the host word that comes with the word *nya* is used. The classes are identified based on the meaning of the word itself, and they are filtered using a pairing word such as a symbol. The challenges of using this approach included some of the words have been wrongly tagged and are relevant in more than two classes. However, the result shows that the approach is acceptable when determining the pleonastic *nya* and *nya* as a human referrer. Among the limitations of the proposed approach are: 1) the reliance on the dictionary in order to identify whether *nya* is part of the word or a separate element; and 2) the identification of a semantic class that uses a certain parameter, therefore neglecting the actual meaning of the word within the context of the sentences. Future work should focus on determining the appropriate antecedent of each type of the word *nya* using factors such as status, class label, and distance. Furthermore, the identification of semantic class can be improved by employing methods from the machine learning field (Ayala, Hernandez, Ruiz, & Toro, 2019).

## ACKNOWLEDGEMENT

This research received no specific grant from any funding agency in the public, commercial, or not-for profit sectors.

## REFERENCES

Antunes, J., Lins, R. D., Lima, R., Oliveira, H., Riss, M. & Simske, S. J. (2018). Automatic cohesive summarization with pronominal anaphora resolution. *Computer Speech & Language*, *52*, 141–164. https://doi.org/10.1016/j.csl.2018.05.004

Aone, C., & Bennett, S. W. (1996). *Applying machine learning to anaphora resolution*. Berlin, Heidelberg: Springer. https://doi.org/10.1007/3-540-60925-3_55

Asao, Y., Iida, R., Torisawa, K. (2018, May). Annotating zero anaphora for question answering. In *Proceedings of the Eleventh International Conference on Language Resources and Evaluation*, Miyazaki, Japan (pp. 3523–3528). https://www.aclweb.org/anthology/L18-1556

Asmah, H. O. (2009). *Nahu Melayu Mutakhir* (5th ed.). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Ayala, D., Hernandez, I., Ruiz, D., & Toro, M. (2019). TAPON: A two-phase machine learning approach for semantic labelling. *Knowledge Based Systems*, *163*(1), 931–943.

Bergsma, S., Lin, D., & Goebel, R. (2008, June). Distributional identification of non-referential pronouns. In *Proceedings of ACL-08: HLT*, Columbus, Ohio (pp. 10–18). https://www.aclweb.org/anthology/P08-1002

Bouzid, S. M., Trabelsi, F. B. F., & Zribi, C. B. O. (2017, November). *How to combine salience factors for arabic pronoun anaphora resolution.* In *2017 IEEE/ACS 14th International Conference on Computer Systems and Applications (AICCSA)*, Hammamet, Tunisia (pp. 929–936). https://10.1109/AICCSA.2017.83

Brants, T, and Franz, A. (2009). *Web 1T 5-gram, 10 European Languages Version 1 LDC2009T25*. Web Download. Philadelphia: Linguistic Data Consortium, 2009. https://catalog.ldc.upenn.edu/LDC2009T25

Daniel, G., & Daniel, J. (2002). Automatic labeling of semantic roles. *Computational Linguistics*, *28*(3), 245–288. http://dx.doi.org/10.1162/089120102760275983

Dewan Bahasa dan Pustaka. (Ed.) (2010). *Tesaurus Bahasa Melayu Dewan Edisi Baharu* (Ed. baharu). Kuala Lumpur: Dawama Sdn. Bhd.

Fazal Mohamed, M. S. (2006). Pengklitikan enklitik "-nya" pada kata kerja: Aplikasi teori kuasaan dan tambatan (Chomsky, 1986). *Jurnal Bahasa*, *6*(3), 363–384.

Hammami S. M., & Belguith L. H. (2018). Arabic pronominal anaphora resolution based on new set of features. In Gelbukh A. (Eds), *Computational Linguistics and Intelligent Text Processing. CICLing 2016. Lecture Notes in Computer Science*, vol 9623. Springer, Cham. https://doi.org/10.1007/978-3-319-75477-2_38

Kabadjov, M. A. (2010). *Anaphora resolution and discourse-new classification: A comprehensive evaluation.* Berlin: VDM Verlag Dr. Müller.

Karimah, M. N. N., Aziz, M. J. A., Noah, S. A. M., & Hamzah, M. P. (2011, June). "nya" as anaphoric word: A proposed solution. In *2011 International Conference on Semantic Technology and Information*

*Retrieval (STAIR)*, Putrajaya, Malaysia (pp. 249–254). https:// 10.1109/STAIR.2011.5995797

Kennedy, C., & Boguraev, B. (1996). Anaphora for everyone: Pronominal anaphora resolution without a parser. In *Proceedings of the 16th conference on Computational Linguistics - Volume 1*, Copenhagen, Denmark (pp. 113–118). https://www.aclweb.org/anthology/C96-1021

King, J. C., & Lewis, K. S. (2018). *"Anaphora", The Stanford encyclopedia of philosophy* (Fall 2018 Edition), Edward N. Zalta (Ed.) https://plato.stanford.edu/archives/fall2018/entries/anaphora/

Kong, F., Zhang, M., & Zhou, D. (2019). Chinese Zero Pronoun Resolution: A Chain-to-chain Approach. *ACM Transactions on Asian and Low-Resource Language Information Processing, 19*(1), 1–21. https://doi.org/10.1145/3321129

Loaiciga, S., Guillou, L., & Hardmeier, C. (2017, September). What is it? Disambiguating the different readings of the pronoun 'it'. In *Proceedings of the 2017 Conference on Empirical Methods in Natural Language Processing*, Copenhagen, Denmark, (pp. 1336–1342). https://doi.org/10.18653/v1/D17-1137

Nik Safiah, K., Farid, M. O., Hashim, H. M., & Abdul Hamid, M. (2008). *Tatabahasa Dewan* (3rd ed.). Kuala Lumpur: Dewan Bahasa dan Pustaka.

Nøklestad, A. (2009). *A machine learning approach to anaphora resolution including named entity recognition, PP attachment disambiguation, and animacy detection.* (Doctoral dissertation). University of Oslo, Norway. https://www.duo.uio.no/bitstream/handle/10852/26326/397_Noeklestad_17x24.pdf?sequence=1&isAllowed=y

Paice, C. D., & Husk, G. D. (1987). Towards the automatic recognition of anaphoric features in English text: The impersonal pronoun "it". *Computer Speech and Language, 2*(2), 109–132. https://doi.org/10.1016/0885-2308(87)90003-9

Qiu, L., Kan, M.-Y., & Chua, T.-S. (2004, June). A public reference implementation of the RAP anaphora resolution algorithm. In *Proceedings of the Language Resources and Evaluation Conference 2004 (LREC 04)*, Lisbon, Portugal. https://www.aclweb.org/anthology/L04-1506/

Saqia, B., Khan, K., Khan, A., Khan, W., Subhan, F., & Abid, M. (2018). Impact of anaphora resolution on opinion target identification. *International Journal of Advanced Computer Science and Applications*, *9*(6), 230–236. https://dx.doi.org/10.14569/IJACSA.2018.090633

Shalom, L., & Herbert, J. L. (1994). An algorithm for pronominal anaphora resolution. *Computational linguistics*, *20*(4), 535–561.

Sukthanker, R., Poria, S., Cambria, E., & Thirunavukarasu, R. (2019). Anaphora and coreference resolution: A review. *Information Fusion*, *59*, 139–162. https://doi.org/10.1016/j.inffus.2020.01.010

Sikdar, U. K., Ekbal, A. & Saha, S. (2016). A generalized framework for anaphora resolution in Indian languages. *Knowledge-Based Systems*, *109*(1), 147–159. https://doi.org/10.1016/j.knosys.2016.06.033

Tabrizi, A. A., Mahmud, R., Idris, N., & Tohidi, H. (2016). A rule-based approach for pronoun extraction and pronoun mapping in pronominal anaphora resolution of Quran English translations. *Malaysian Journal of Computer Science*, *29*(3), 207–224. https://doi.org/10.22452/mjcs.vol29no3.4

Vicedo, J. L., & Ferrández, A. (2000, October). Importance of pronominal anaphora resolution in question answering systems. In *Proceedings of the 38th Annual Meeting of the Association for Computational Linguistics*, Hong Kong (pp. 555–562). https://10.3115/1075218.1075288

Werlen, L. M., & Popescu-Belis, A. (2017, September). Validation of an automatic metric for the accuracy of pronoun translation (APT). In *Proceedings of the Third Workshop on Discourse in Machine Translation*, Copenhagen, Denmark (pp. 17–25). https://www.aclweb.org/anthology/W17-4802/

Xian, B. C. M., Saloot, M. A., Ghazali, A. S. M., Bouzekri, K., Mahmud, R. and Lukose, D. (2016, June). Benchmarking Mi-AR: Malay anaphora resolution. In *International Conference on Optoelectronics and Image Processing (ICOIP)*, Warsaw, Poland (pp. 59–69). https://doi.org/10.1109/OPTIP.2016.7528520

Yaneva, V., Ha, L. A., Evans, R., & Mitkov, R. (2018, June). Classifying referential and non-referential it using gaze. In P*roceedings of the Conference on Empirical Methods in Natural Language Processing*, Brussels, Belgium. http://dx.doi.org/10.18653/v1/D18-1528

Yap, F. H. (2011). Referential and non-referential uses of nominalization constructions in Malay. In F. H. Yap, K. Grunow-Hårsta, & J. Wrona (Eds.), *Nominalization in Asian Languages*. John Benjamins Publishing Company. https://doi.org/10.1075/tsl.96.22yap

Yifan, Y. L., Musilek, P., Reformat, M., & Wyard-Scott, L. (2009). Identification of pleonastic it using the web. *Journal of Artificial Intelligence Research*, *34*(2009), 339–389. https://10.1613/jair.2622

Zhao Y., Liu J., & Yin C. (2018). Chinese anaphora resolution based on adaptive forest. In Park J., Loia V., Yi G., Sung Y. (Eds.), *Advances in Computer Science and Ubiquitous Computing. CUTE 2017, CSA 2017. Lecture Notes in Electrical Engineerin*g, vol. 474. Singapore: Springer. https://doi.org/10.1007/978-981-10-7605-3_79