Abstract—Anaphora resolution (AR) is a process to identify the appropriate antecedent with its anaphor which occur before the anaphor. AR able to improve most of the NLP applications such as question answering, short answer examination system and information extraction. Most of AR systems are deal with English language. Thus, in 1990’s the research on AR has been applied for other language, such as Arabic, Chinese, Hindi and Norwegian. There are however limited or no effort in dealing with Malay text. The AR systems for one language cannot be simply adapted to use in other languages. This is due to the fact that different languages have different set of rules relating to syntax and semantic to respective language. This paper proposed a model for resolving anaphora phenomena in Malay text. The model consists of three elements consisting of anaphora resolution process, syntactic knowledge process and semantic-world knowledge process. The elements are defined based on the observable fact occurring in Malay language.

Keywords—model of anaphora resolution, poor-knowledge anaphora resolution, Malay text, natural language processing

I. INTRODUCTION

ANAPhORA resolution is a phenomenon in which the occurrence of reference for the word (anaphor) to its entity (antecedent), which occurs before the anaphor. There are various types of anaphora depending on its location in a document. According to [1] and [2] several type of anaphor has been identified. Some of them are pronominal anaphora, nouns anaphora, nouns phrase anaphora and zero anaphora. Pronominal anaphora occurred on pronoun anaphor, where the resolution process is to identify the appropriate antecedent that referred by the pronoun. Noun anaphora is occurs when non-lexical pro-form point back to its entities. The example of noun anaphora are ‘one’, ‘yang’, and ‘other’. Noun phrase anaphora is the common noun or proper noun as anaphor. Basically this type of anaphora needs extra knowledge during search appropriate antecedent, such as semantic knowledge and world knowledge. The zero anaphora is anaphor does not appear in the sentence because they do not present into any word or phrase. The most popular is pronominal anaphora [1, 3-5]. The other types of anaphora apart from pronominal anaphora required the world-knowledge (WK) and common sense reasoning in order to be solved. Some views this as impractical as the WK requires huge knowledge base and set of inference rules [6].

Anaphora resolution is crucial in many current information retrieval applications such as question-answering, semantic search, classification and clustering, and ontology learning. Most of the earlier research on anaphora resolution systems deal with the English language [2]. However, beginning 1990’s, anaphora resolution has been applied to other languages; such as Japanese, French, Chinese, Hindi, Arabic, Spanish and Norwegian. Yet, as to date effort on anaphora resolution for Malay language is little or none.

The research on anaphora resolution has been actively conducted in mid 70s. This is indicated from the study by Hobbs [7]. However, research in this area gained little attention in the early 80s but bounced back at the end of the 80s. Beginning 1990s the research on anaphora resolution not only concern on the theoretical aspects but started to involve the computational linguistic field. JavaRAP and Mitkov’s Anaphora Resolution System are among the examples that portray this trend.

Earlier knowledge-based approach relies on hand-crafted knowledge develop by linguists. In this case, the anaphora resolution is solve by using the linguistic knowledge that represent syntax, semantic and discourse algorithm. All the preprocessing is conducted manually and the input is normally assumed to be perfect that has been check or modified by the expert. Most of these systems are linguistic theoretical test (manually). The existent of pre-processing...
tools and the difficulties to adapt the knowledge based system into other language leads the researchers to change their focus into heuristic based system (poor-knowledge). These systems offer the simplicity and robustness design that give performance comparable with the knowledge based system. Some example using these approaches are the Mitkov’s original approach (MOA), RAP and ARN. Table 1 provides a brief comparison of these systems which was made available in [5].

<table>
<thead>
<tr>
<th>TABLE I. COMPARISON OF THE EXISTING SYSTEMS</th>
</tr>
</thead>
<tbody>
<tr>
<td>Element</td>
</tr>
<tr>
<td>Result (Success rate)</td>
</tr>
<tr>
<td>Preprocessing</td>
</tr>
<tr>
<td>Aim</td>
</tr>
<tr>
<td>Language</td>
</tr>
</tbody>
</table>

This paper discusses main issues in anaphora resolution for Malay language with the intentions of proposing solution model for resolving it. The research will focus on into two types of anaphora; pronominal and noun phrase. This paper is organized into the following section. Section II discussed about the issues of anaphora resolution for Malay text. And the proposed model solution is provided in Section III. We conclude our paper in section IV.

II. ISSUES IN ANAPHORA RESOLUTION FOR MALAY TEXT

Despite the exponential growth of electronic documents in Malay language, there is little or no research on Anaphora Resolution for Malay language that has been done. Thus, one of the interesting are in Malys language technology. Anaphora resolution has the capability to improve most of the NLP applications such as question answering, short answer examination system, information extraction. Consider the example 1-1.

Example 1-1:
Nasi adalah merupakan menu yang berkhasiat. Makanan ini mengandungi kabohidrat dan juga glukosa.
Rice is a healthy menu. The food contains carbohydrate as well as glucose.

Consider the application of a question answering systems that make use of the sentence in Example 1-1. If user wants to know the mineral content rice by submitting a question in Malay such as “Apakah kandungan nasi”. This will be difficult to trace in the sentence since the information needed is in the following sentence. In this example, the word “makanan ini” refers to “nasi”. Anaphora resolution can be used to resolve the similar words that refer to which word in the same paragraph. Let’s used the example 1-1 for short answer examination system. Let say the answer scheme is as provided in the example 1-2. In this case, zero anaphora exit. The accuracy for similarity is less if the user provides the answer as in example 1-2. On the other hand, if anaphora resolution is being considered, the example 1-1 is similar with the answer scheme on example 1-2 event though the word of “makanan ini” is not available in the answer scheme. In the case of information extraction system, if the information about “nasi” is needed only “menu yang berkhasiat” will be retrieved. However, if the anaphora resolution is used, the information retrieval will include the content of “nasi”. In this case, the phrase “mengandungi kabohidrat dan juga glukosa” will be extracted which indicates that carbohydrate and glucose as the content of rice.

III. KNOWLEDGE-POOR APPROACH

Knowledge-poor is one of the initial approach proposed by Mitkov [1] in resolving anaphora. He has applied the approach on Polish and Arabic language with minimum modification or without any modification for both languages and the result still in acceptable in terms of success rates. However, Holen [5] found that the anaphora resolution systems cannot freely be adapted into other languages. Many rules must be added in order to satisfy with the knowledge of each language used. For example, certain rules in Malay language are similar with English especially in morphological knowledge. However, most of Malay language structures are dissimilar English structures. This can be shows the different of basic structure or pattern for English and Malay language. Where there is one basic sentence pattern for English and 4 basic sentences pattern for Malay language.

Mitkov [6] identified in his model, for syntactic module the anaphor and antecedent must exact match in terms of number, gender and person. However, pronouns in Malay language are dissimilar with English language or other language such as Arabic, Japanese and Chinese, where it cannot be used to identify gender. This is because some of pronoun such as dia, ia, nya, baginda and beliau could refer to both female or male. Gender in Malay language can only be differentiated by using nouns phrase. Determination of nouns phrase used is accordance with the pronoun that reflects the use of appropriate status. Therefore the noun recognition is needed. Among the noun recognition required is the difference between the animate vs. inanimate, human

Example 1-2
Nasi adalah merupakan menu yang berkhasiat, mengandungi kabohidrat dan juga glukosa.
Rice is a healthy menu that contains carbohydrate as well as glucose.
III. PROPOSE MODEL FOR ANAPHORA RESOLUTION OF MALAY TEXT

Designing a model of Anaphora Resolution for Malay Text (ARMT) is by comparing to other models used in different languages. The characteristics and morphology of Malay words were explored. The morphology information shows that Malay word (noun phrase) has several different features with another language. For example, noun and verb can be created not only by using prefix and postfix and infix but also prefix-postfix. The pronouns cannot show the gender, but can be used to differentiate the status of people within the society. In Malay sentences, animal and thing can be replace with pronoun ia and ianya. Because of ia can also be used to refer human being, semantic knowledge is desired. The semantic knowledge is used not only to determine the meaning and nature of the word (noun, verb, adjective), but to determine the characteristics of words and determine the animate or inanimate, human or non-human and non-abstract or concrete (semantic relation/behavior).

The propose ARMT model is a combination of two type of anaphora: pronominal anaphora and noun phrase anaphora. The model consists of three (3) main elements as illustrated in figure 1. The first element meant to resolve the anaphora for Malay text, follow by the second and third element which are the syntactic knowledge and the semantic-world knowledge respectively, of which both will be used to resolve the anaphora resolution in the first element.

The first element consists of five (5) stages. The stages are extraction of noun phrase and verb phrase from the annotate corpus, identify the antecedent candidates (e.g. common nouns, proper noun), identify the anaphor candidates (e.g. pronouns that show the status, common noun, proper noun), anaphora resolution for noun phrase anaphora and pronominal anaphora and the result that will be evaluate by the expert. During the process of resolving the anaphora, several processes will be involved such as, Malay name entity recognition, nouns detail detection and verb dependency.

The second element is used as an input to first element in order to indentify name entity recognition, verb dependency, sentence pattern and nya as not anaphora. Name entity recognition is focused on the Malay names, for human, location and organization. The recognition is based on the pattern of the Malay sentences. Verb dependency is used to resolve some of the ambiguity occurs in Malay text. This task requires the input from semantic-world knowledge in order to identify appropriate nouns that referred to either subject or object of the previous sentence. In the case of nya that is not a pronoun will be detected based on the combination of word with word nya that used in the sentences. The study has been defined that word nya become pronoun when it combine with noun or transitive verbs. On the other hand, sentence pattern is needed in order to identify either the pronominal anaphora is occur or not, and also in order to identify the subject and predicates for a sentence. This is because there are dissimilar of basic sentence pattern for Malay language and English.

The third element is semantic knowledge model. This element is used to identify the properties of nouns and as an input to the syntactic knowledge. Noun properties are meant to be used for classifying nouns into animate or inanimate, human or non-human and concrete or abstract. Verb is use
for the syntactic knowledge process in order to conduct the verb dependency identification.

There are five (5) steps in semantic-world knowledge process. The steps begin with the collection of noun phrases and verbs. The noun phrase/noun referred to people, animal and place, whereas verb is a collection of simple verb that normally used in text. The second step is the classification of each noun phrases into appropriate class or hierarchy. In this stage, semantic hierarchy concept is used. The concept also used in determining the properties of each word properly. After all conditions have been considered, information and semantic-world knowledge model will be evaluated by linguists before being computationally implemented. The output of the semantic-world knowledge will be used in the anaphora resolution process and the syntactic process.

IV. CONCLUSION

As a conclusion, there is no single anaphora resolution model can be applied for various languages. This is due to the fact that different languages have different syntax, semantic and usage. The Malay language for example, the pronoun cannot differentiate between genders but able to shows the difference of status of the referred entities in societies. This led to different methods for solving anaphora resolution. In order to referring animals or things, the other words are commonly used. In this situation the semantic-world knowledge is required. The semantic-world knowledge also used to distinguish appropriate words for human or not human, physical or abstract and living or not living. By means of these reasons, anaphora resolution model for the Malay text is proposed. The model is used as systematic guidance in resolving anaphora phenomenon for Malay language in the future work.

REFERENCES