A Fundamental Study on Detecting Head Modifier Noun Phrases in Malay Sentence

Suhaimi Ab Rahman
Software Engineering Department,
Universiti Tenaga Nasional,
Jalan IKRAM-UNITEN, 43000 Kajang, Selangor, Malaysia
Tel.:+603 8921 2020,
E-mail: smie@uniten.edu.my

Nazlia Omar, Mohd Juzaidin Ab Aziz
School of Computer Science, Faculty of Information Science and Technology,
Universiti Kebangsaan Malaysia(UKM), Bangi, 43600,
Selangor, Malaysia
Tel.:+603 8921 6733,
E-mail: no@ftsm.ukm.my; din@ftsm.ukm.my

Abstract—This paper discusses on how a head modifier of Noun Phrases (NPs) in Malay sentence can be detected from the four combination of phrases, such as a noun phrase (frasa nama) and noun phrase (frasa nama), a noun phrase and verb phrase (frasa kerja), a noun phrase and adjective phrase (frasa adjektif) and, a noun phrase and prepositional phrase (frasa sendi). Most of the sentences in Malay have compound nouns. The position of the head within a compound often depends on the order of the word, i.e. the most common order of constituents in Malay phrases, where nouns are modified by adjectives, verbs, other nouns and prepositional. We also investigated the relative contribution of the modifier and the head noun in noun compounds of different other related examples in Malay sentence.

Keywords—head modifier; Malay phrases; compound noun; thematic relations; subject-predicate

I. INTRODUCTION

A compound noun is a noun that is made up of two or more words. Most compound nouns in English are formed by nouns modified by other nouns or adjectives. The issues of noun compounds have been extensively studied from linguistics to find what are the acceptable solution to overcome the problems encountered. It is most important when this concept is to be applied in developing a Natural Language Processing (NLP) system such as Machine Translation, Language Grammar Checker and etc.

Although compound nouns are common in English, they are not only limited to one language. It is also being studied in other languages, such as Chinese, Japanese, French, German and Italian. There are a variety of compound noun definitions employing different criteria as noted in [5],[8],[10]. Generally, the fundamental ideas of comprehension compound nouns is obtained in English, and it is beneficial to be practiced in other languages. It requires some alterations to ratify compound nouns with the new structure and syntax of a language study, such as a Malay language.

In this paper, we explicitly model this problem by selecting the issues raised from Malay language.

II. RELATED STUDY

A. Compound of Word

Presenting compound of words in [4] clearly demonstrates that the group of compound nouns can be determined by the trained bracket from the whole sentence. To avoid ambiguity of the word meaning, he implemented two types of bracketing approach such as left-branching and right-branching bracketing. The problem using bracketing process is still unresolved. It was designed to handle only joining word-to-word of compound nouns. The problem surfaced if the word joined has the accumulation of more than one word. It requires more trained bracketing compound noun data into a database to suggest solutions.

He also used 20 types of noun modifier relationships to represent the semantic relations between concepts. Some of the examples of the semantic relations are agent, beneficiary, cause, instrument and etc. All these relationship types are useful to obtain a right compound order for the words in terms of the head modifier syntax. All the concepts will stick together with their semantic relations to become more effective to locate a head modifier of the words. More research is required , particularly in understanding the
fundamental of semantic relations applied in the compound nouns structure. However, to produce accurate results, he decides to construct more related compound nouns in combination with their semantic relations.

B. Thematic relationships (CARIN model).

The primary idea of CARIN model comes from the research studies from a few linguistic experts that particularly focuses more on thematic relations of compound words. In implementing thematic relations between the words or concepts, it requires two steps. The first step is developing taxonomies of relations, and the second step is to identify and create the list of words, which will be important for interpreting combinations of taxonomies of relations. According to [8], there are 15 relations (thematic relations) that can be used to classify the meanings. Based on these thematic relations, the classification of word head or modifier inside a compound word becomes more understandable. Below is an example of the 15 relations (thematic relations) used by [8]:

<table>
<thead>
<tr>
<th>Modifier causes</th>
<th>Noun, example</th>
<th>Modifier located in Noun, example</th>
</tr>
</thead>
<tbody>
<tr>
<td>Noun has Modifier, example</td>
<td>Noun made of Modifier, example</td>
<td>Noun located in Modifier, example</td>
</tr>
<tr>
<td>Noun has Modifier, example</td>
<td>Noun for Modifier, example</td>
<td>Noun located in Modifier, example</td>
</tr>
</tbody>
</table>

Based on this model, there are two main processes involved to perceive the thematic relations of the components. The first step is that the system will check for the specific relations that are used with particular concepts and the second step is making interpretation of the relative strength of the selected relation.

However, there are a few problems need to be addressed. Firstly, the number of thematic relations proposed is 15 types, but in general, it is intractable to determine the exact number of thematic relations required. This depends on the issues raised from the new language studies. Secondly, in terms of word ambiguity, this model does not describe on how the thematic relations can resolve the problem of word ambiguity such as the word interest. As extracted from (WordNet, 2000), the word interest can be described with a few senses such as curiosity about someone or something; i.e "an interest in music" or a fixed charge for borrowing money; i.e “how much interest do you pay on your mortgage?”. Anyway, this idea becomes more important in order to implement the same concept into our future research work, especially for handling different kinds of data situated in Malay language.

III. ANALYSIS OF MALAY NOUN PHRASES HEAD MODIFIER STRUCTURE

The initial idea started with observing and understanding the pattern grammar structure of words in Malay sentence. As discussed in [2],[1],[9], all the sentences in Malay language were built based on the accumulation of four word categories, for instance "Kata Nama" (Noun), "Kata Kerja" (Verb), "Kata Sifat" (Adjective) and “Kata Tugas” (Function word). With the various combination of either using one or more of the above word categories, it can assist to model a complete sentence structure in Malay language.

As noted in [2],[9] every phrase in Malay sentence has its “inti” (head). The word appearing as “inti” (head) is most significant compared with other words that signify as “penerang” (modifier), specifically in perceiving the word in Malay sentence, whether it is in a context of a word or as others.

In the following discussion, we describe the four categories of combination in Noun Phrases to produce a complete meaningful sentence in Malay language, called as simple sentence (ayat tunggal). The simple sentence is referred to the sentence having only one subject and one predicate.

As discussed in [2], the combination of the phrases in Malay sentence is as follows:

<table>
<thead>
<tr>
<th>Table II. SIMPLE SENTENCE MALAY CONSTRUCTION</th>
</tr>
</thead>
</table>

A. A combination between Noun Phrase (Frasa Nama-FN) and Noun Phrase (Frasa Nama)

1 WordNet is a database comprised of nouns, verbs, adverbs and adjectives in a lexical relational structure. These words are then grouped by synonyms called synsets that are easily used to find related words and usage. This basic structure of WordNet allows it to be easily used and researched for computational linguistics and language processing for a variety of applications such as voice recognition software and dictation software.
The construction of a simple Malay sentence using this pattern combines a Noun Phrase subject and a Noun Phrase predicate. The example is as follows:

Simple sentence : 'Orang itu seorang ulama besar'
Subject : 'Orang itu' (Noun Phrase)
Predicate : 'seorang ulama besar' (Noun Phrase)

The position of head and modifier for the above sentence is depicted as follows:

From the above Figure 1, we determine that the word ‘Orang’ inside Noun Phrase subject is a head, while the word ‘itu’ is a modifier. Inside a Noun Phrase predicate, the word ‘ulama’ is a head, while the rest of the words become complement to the meaning of the head word ‘ulama’. To perceive a head of word(s) inside phrases, this is determined by the part-of-speech (POS) of noun (Kata Nama-KN).

B. combination between Noun Phrase (Frasa Nama) and Verb Phrase (Frasa Kerja)

The construction of a simple Malay sentence using this pattern is associated with a Noun Phrase subject and a Verb Phrase predicate. The example is as follows:

Simple sentence : 'Budin sedang bekerja di sawah'
Subject : 'Budin' (Noun Phrase)
Predicate : 'sedang bekerja di sawah' (Verb Phrase)

The position of head and modifier for the above sentence is depicted as follows:

From Figure 2, we discover that the word ‘Budin’ is a head of the word inside subject Noun Phrase, while the word ‘bekerja’ which is located in the Verb Phrase predicate becomes the head. The other succeeding words serve as complments to the meaning of the head word ‘bekerja’. As described in [2], to recognise a head of the word which appears as in a Verb Phrase, it is usually determined by the position of Auxiliary word (Kata Bantu-KB) followed by Verb (Kata Kerja-KK), such as ‘sedang[KB] mengail[KK]’, ‘akan[KB] pergi[KK]’, ‘sudah[KB] mandi[KK]’, etc.

C. A combination between Noun Phrase (Frasa Nama) and Adjective Phrase (Frasa Adjektif)

The construction of a simple Malay sentence using this pattern basically joins a Noun Phrase subject and an Adjective Phrase predicate. The example is as follows:

Simple sentence : 'Pokok kelapa itu sangat subur'
Subject : 'Pokok kelapa itu' (Noun Phrase)
Predicate : 'sangat subur' (Adjective Phrase)

The position of head and modifier for the above sentence is depicted as follows:

From Figure 3, we conclude the word ‘kelapa’ inside the Noun Phrase subject becomes a head of Noun Phrase, while the other words are called modifiers. Inside an Adjective
Phrase predicate, the word ‘subur’ is a head, while the rest of the words are complement to the meaning of the head word ‘subur’. As described in [2], to recognise a head of the word which appears in an Adjective Phrase, it is usually determined by the position of word Intensifier (Kata Penguat-KP) followed by Adjective (Kata Sifat-KA) or the word Adjective (Kata Sifat) followed by Intensifier (Kata Penguat), such as ‘sangat[KP] kemas[KA]’, ‘terlalu[KP] tinggi[KA]’, ‘cantik[KA] sekali[KP]’, ‘tebal[KA] sangat[KP]’, etc.

D. A combination between Noun Phrase (Frasa Nama) and Prepositional Phrase (Frasa Sendi Nama)

The construction of a simple Malay sentence within this pattern combines a subject Noun Phrase and a predicate Prepositional Phrase. The example is as follows:

Simple sentence : ‘kebisingan di dewan makan’
Subject : ‘kebisingan’ (Noun Phrase)
Predicate : ‘di dewan makan’ (Prepositional Phrase)

The position of head and modifier for the above sentence depicted as follows:

From the above Figure 4, we conclude the word ‘kebisingan’ inside the Noun Phrase subject is a head of Noun Phrase. Inside a Prepositional Phrase predicate, the word ‘di’ is a head, while the other words are complement to the meaning of a head word ‘di’. As described in [2], the modifiers of a Prepositional Phrase are the elements that contains preposition (Kata Sendi Nama) as a head followed by one of the subsequent phrases: i) Noun Phrase, such as ‘di pasar’. ii) Directional Word (Kata Arah) + Noun Phrase, such as ‘di tepi pasar’. iii) Directional Word (Kata Arah) + Noun Phrase + Complement Sentence (Ayat Komplemen), such as ‘ke dalam hutan untuk menangkap rusa’. iv) Directional Word (Kata Arah) + Noun Phrase + Adverbial Sentence (Ayat Keterangan), such as ‘di selatan negeri China pada tiap-tiap bulan’.

IV. SYSTEM FLOW OF THE PROCESS

Process flow diagrams should include the information regarding the connection between various components or modules. This is to ensure leased capabilities are supported adequately and can achieve specified availability requirements into our study. Figure 5 shows a diagram on how the components can interact with each other in order to complete the series of processes in our earlier research planning phase.

As illustrated in Figure 5, the process starts with entering a Malay input sentence. After the Malay input sentence is entered, the process of tokenising will be performed. This process is to filter any unrecognised symbols or punctuation in the input sentence. All tokenised words will be assigned with Malay Part-of-Speech (POS) via inferring from Malay POS database. The output generated from this tagging process contains a list of the word adjoined with their POS, e.g.

Input sentence: 
Orang itu seorang ulama besar

Output sentence: 

The next process executed is called Sub-Phrase Classifier. By using this process, the above sentence structure will be divided into a few sub combinations of words, such as:

chunk \(i=0\) : (orang[KN], itu[KN])
chunk \(i=1\) : (itu[KN], seorang[KN])
chunk \(i=2\) : (seorang[KN], ulama[KA])
chunk \(i=3\) : (ulama[KN], besar[KA])

There are two important reasons for using this structure. The first is to determine the number of sub-phrases in the input sentence. The number of sub-phrases will closely be guided with a grammar rule created from the database. By using the above sentence, the three sub-phrases constructed are:

sub-phrase \(i=0\) : (orang[KN], itu[KN])
sub-phrase \(i=1\) : (seorang[KN])
sub-phrase \(i=2\) : (ulama[KN], besar[KA])

The second is to use the above sub-phrase as a structure for detection head modifier. The head modifier for the words can be decided by a Thematic Relation Detector process.

By using this thematic relation detector, each pair of a sub-phrase will search from a thematic relation database looking for words that match the criteria in thematic relation argument structure. The following is an example of the thematic relation produced:

post-determiner-for (orang[KN], head, itu[KN], modifier)
Using the thematic relation categories, the nomination of head and modifier for the words can be assigned correctly. The database thematic relation will be necessary for this process.

Finally, the head modifier phrase structure generator will be used to generate the structure of Noun Phrase head modifier relationship. For example:

Input sentence: 
Orang itu seorang ulama besar.

Output: 
((orang[KN], head, post-determiner-for) itu[KN, modifier])
(seorang[KN, modifier] (ulama[KN, head, adjectival-for] besar[KA, modifier])))

Figure 5. System flow of the process

V. CONCLUSION AND FUTURE WORK

As discussed, we can conclude that the construction of Malay sentences is normally built based on four compounds of phrases, such as FN+FN, FN+FK, FN+FA and FN+FS. Since the field of Malay language study is enormous, we have to acknowledge the area of study correctly. We additionally need to identify and list the issues which might be surface from the Malay Language. The help from linguists or other language experts are also necessitated in order to refine the solutions focusing on head modifier Noun Phrases in Malay sentence.

For future work, we attempt to study on the research issue, based on three main important categories such as: head modifier data acquisition, head modifier knowledge representation and head modifier engine processing. All the processes design is built for Malay language.

REFERENCES