Transforming Noun Phrase Structure Form into Heuristics and Rules For Detecting Compound Noun in Malay Sentence

Suhaimi Ab Rahman[1] and Nazlia Omar[2]

[1]*Software Engineering Department, College of Information Technology, Universiti Tenaga Nasional, smie@uniten.edu.my*

[2]*School of Computer Science, Faculty of Information Science and Technology, Universiti Kebangsaan Malaysia, no@ftsm.ukm.my*

ABSTRACT

This paper addresses the process of transforming noun phrase structure form into a list of suitable heuristic used for detecting compound noun word in Malay sentence. The heuristic is used to obtain the syntax sentence structure for finding a compound noun pair of words in Malay sentence. To obtain the list of these rules, the noun phrase structure form must be created first, so that we know the possibility of the words' combination as a compound noun. The noun phrase structure form is grouped based on three different noun categories, such as i) noun and noun ii) noun and noun modifier, and iii) noun and non-noun modifier. However, in our research work, we focus on the category of noun and noun modifier. The heuristic rules and noun phrase structure form are important to understand because they help to clarify the concept of finding compound noun pair of words in Malay sentence. This compound noun output will use an input in our next research named head modifier detector.

Keywords: noun phrase structure form, heuristic rules, compound noun.

I INTRODUCTION

Noun phrase structure form in Malay sentence is made up based on a combination of words from the category i) noun and noun ii) noun and noun modifier, and iii) noun and non-noun modifier. To construct noun phrase structure form, we refer to a few studies from (Addullah, 2004), (Addullah, 1992), (Nik Sapiah et al, 2010) and (Ong, 2009) . As explained in (Nik Sapiah et al, 2010), they create a table as a form to assign words. They also specify the group of words, such as noun and noun or noun and verb . By using this form, it will assist to overview of the pattern structure of Malay sentence which is added into the form. The word's arrangement in the table is based on the three categories discussed earlier. However, the new noun phrase structure formed still requires validation from linguists to check, especially those with vague words in terms of placing the word's position in the form and identifying a group of words for compound noun.

The relevant study is also discussed in (Dinh, 2002). He used syntactic parsing of the Vietnamese Noun Compound to determine the compound noun of words. In terms of the syntactic aspects, he defined the first word of noun compound must be a noun, and the second word could be a noun, verb, adjective, pronoun, preposition, number or etc. to form the compound noun. He used three different types of frame structures, such as i) Frame structure of noun object ii) Frame structure of verb object and iii) Frame structure of an adjective object. He used compound noun as a part of the process to detect the head and modifier of the words with the help of semantic relations between objects.

To discuss further, we will divide this paper into the following sections: Section 2 of the paper describes an overview of noun phrase structure form. The rule and heuristics are introduced in section 3. The examples of input and output are explained in Section 4. Finally, in section 5, we list our conclusions and future work.

II NOUN PHRASE STRUCTURE FORM

The noun phrase structure is worthwhile to determine the possibility combination of Malay words to produce Malay sentence. This form is used to find the compound nouns of words inside a Malay sentence. Every word in Malay sentence must follow the syntax or rule to produce correct meaning of the sentence. In order to create the new form of the noun phrase structure, we performed comprehensive study and analyse the words obtained from other relevant resources such as a collection of children story book, dictionary, Malay text book and magazine.

The list of Malay sentence collected is reviewed. To analyse the sentences, we refer to some available examples of noun phrase structure form in (Nik Sapiah et al, 2010). Based on this form, we can create another noun phrase structure form that can be applied with the same concept as

explained from (Nik Sapiah et al, 2010). In Malay sentence, there are a lot of possibilities where sentences can be created. To arrange the noun phrase sentences, we divided the Malay sentence into a few group of categories as described in (Nik Sapiah et al, 2010), (Addullah, 2004) such as i) noun phrase and noun phrase ii) noun phrase and verb phrase iii) noun phrase and adjective phrase iv) noun phrase and preposition phrase.

As we discussed earlier, the noun phrase structure form is important to produce the list of heuristics, which will be used in nominating the compound noun in a sentence. Below is the example on how we analyse the sentences in Malay to create suitable noun phrase structure form. The form created will used for formulating the heuristics and constructing correct group of compound nouns.

Example One: Combination of noun and noun modifier

Sentence 1: *pakain kain kapas buatan Malaysia ini*
Sentence 2: *rumah harga murah binaan kerajaan*
Sentence 3: *buku kulit nipis harga murah ini*

Table 1. Combination Noun and Noun Modifier.

Word and Part-Of-Speech(POS)					
	Phrase modifier 1		Phrase modifier 2		Deter-miner
word1	word2	word3	word4	word5	word6
pakaian (KN)	kain (KN)	kapas (KN)	buatan (KN)	Malaysia (KN)	ini (KN)
rumah (KN)	harga (KN)	murah (KA)	binaan (KN)	kerajaan (KN)	-
buku (KN)	kulit (KN)	nipis (KA)	harga (KN)	murah (KA)	ini (KN)
:	:	:	:	:	:
n	n	n	n	n	n

Abbreviation:
KN : *Kata Nama* (Noun)
KA : *Kata Adjektif* (Adjective)
n : number of words

Example Two: Combination of noun and non-noun modifier, such as verb.

Sentence 1: *mesin jahit buatan Malaysia*
Sentence 2: *meja makan ruang tamu*
Sentence 3: *kedai gunting rambut*

Table 2. Combination Noun and Non-Noun Modifier.

Word and Part-Of-Speech(POS)			
word1	word2	word3	word4
mesin (KN)	*jahit (KK)*	*buatan (KN)*	*Malaysia (KN)*
meja (KN)	*makan (KK)*	*ruang (KN)*	*tamu (KN)*
kedai (KN)	*gunting (KK)*	*rambut (KN)*	-
:	*:*	*:*	*:*
n	n	n	n

Abbreviation:
KK : *Kata Kerja* (Verb)

Example Three: Combination of noun and non-noun modifier, such as adjective.

Sentence 1: *minyak wangi haruman kasturi*
Sentence 2: *sekolah rendah pinggir bandar*
Sentence 3: *rumah besar milik keluarga Ahmad*

Table 3. Combination Noun and Non-Noun Modifier.

Word and Part-Of-Speech(POS)				
word1	word2	word3	word4	word5
minyak (KN)	*wangi (KA)*	*haruman (KN)*	*Kasturi (KN)*	-
sekolah (KN)	*rendah (KA)*	*pinggir (KN)*	*bandar (KN)*	-
rumah (KN)	*besar (KA)*	*milik (KN)*	*keluarga (KN)*	*Ahmad (KN)*
:	*:*	*:*	*:*	*:*
n	n	n	n	n

The above three forms are examples of creating noun phrase structure form based on the Malay sentence examples collected from our data collection process. It will assist to define the possibility of rules in detecting compound noun. The list of rules formulated are not sufficient enough to detect another sentence in Malay. The only sentences which fully comply with the form can be accurately detected. To spread out searching for detection of compound noun, we have to study another sentence to produce a new noun phrase structure form. Once the new form is created, we can formulate new rules to detect compound noun in the sentence.

The fundamental concept of compound noun in Malay sentence is also discussed in (Suhaimi et al, 2012), (Suhaimi et al, 2011).

III RULE AND HEURISTICS

The meaning of rule in AI can be defined as IF-THEN structure that relates given information or facts. According to (Michael, N, 2002), rules can represent as relations, recommendations, directives, strategies and heuristics. In our research work, the concept of relations and heuristics will be used as a guideline to construct the list of rules based on the noun phrase structure form defined earlier.

The relation's rule can be represented as follows:

IF (antecedent)
THEN (consequent) (Relation 1)

The above relation is suitable for the antecedent which does not have a combination with another relational operator such as logical AND (conjunction), logical OR (disjunction) or a combination of both. Sometimes, a rule can have multiple antecedents joined with one or two relational operators. The heuristics can use to represent for this kind of condition.

The heuristic can be represented as follows:

i) Combination with two Logical AND

IF ($antecedent_1$) AND ($antecedent_2$) AND ($antecedent_3$)
THEN ($consequent_1$) (Heuristic 1)

ii) Combination with two Logical OR

IF ($antecedent_1$) OR ($antecedent_2$) OR ($antecedent_3$)
THEN ($consequent_1$) (Heuristic 2)

iii) Combination with two Logical AND and OR

IF ($antecedent_1$) AND ($antecedent_2$) OR ($antecedent_3$)
THEN ($consequent_1$) (Heuristic 3)

By taking the example of a noun phrase structure form of the category noun and noun modifier, below is a discussion on how the heuristic is created.

As shown in Table 1, the compound noun of Malay words can be positioned at Column 2 (word2) and Column 3 (word3), and next compound nouns is come from Column 4 (word4) and Column 5 (word5). This condition is only eligible for the sentences matched with this syntax (sentence form).

Using this form, the rules can be constructed as follows:

General Rule 1: words with POS *KN*
IF ((numberOfWordsInSentence is equal 6 and POS having *KN*)
THEN { Execute Rule 1; Execute Rule 2}

This rule will check the number of words in a sentence. The tokenization process will execute here to filter and split the word. After applying the filter process, the next logical condition will check either all the words are noun. If both logical conditions satisfy, then Rule 1 and Rule 2 will operate.

Rule 1:

$antecedent_1$ $Antecedent_2$

IF (POS $word_2$ is equal *KN* and POS $word_3$ is equal *KN*)
THEN $compoundNoun_1 = word_{2[KN]} + word_{3[KN]}$

According to this Rule 1, the pair of words for $compoundNoun_1$ will be appointed if both antecedents return TRUE.

Rule 2:

$antecedent_1$ $Antecedent_2$

IF (POS$word_4$ is equal *KN* and POS$word_5$ is equal *KN*)
THEN $compoundNoun_2 = word_{4[KN]} + word_{5[KN]}$

Similar condition will be applied for Rule 2 where the pair of words for $compounNoun_2$ will be appointed if both antecedents return TRUE.

General Rule 2: words with POS *KN* and *KA*
IF ((numberOfWordsInSentence is equal 6 and POS having *KN* and *KA*)
THEN { Execute Rule 3; Execute Rule 4}

Rule 3:

antecedent$_1$ Antecedent$_2$

IF (POS word$_2$ is equal *KN* and POS word$_3$ is equal *KA*)
THEN compoundNoun$_3$ = word$_{2[KN]}$ + word$_{3[KA]}$

Rule 4:

antecedent$_1$ Antecedent$_2$

IF (POSword$_4$ is equal *KN* and POSword$_5$ is equal *KA*)
THEN compoundNoun$_4$ = word$_{4[KN]}$ + word$_{5[KA]}$

The same analysis process we perform in order to formulate more rules based on the noun phrase structure form created.

IV EXAMPLES OF INPUT AND OUTPUT

Figure 1 depicts the process on how we nominate the compound noun based on the rules defined above. This example is taken from a Table 1.

Input Malay sentence:
Buku kulit nipis harga murah itu.

⇩

Buku[*KN*] *kulit*[*KN*] *nipis*[*KA*] *harga*[*KN*] *murah*[*KA*] *itu*[*KN*]

⇩

Met condition with General Rule 2

⇩

Apply Rule 3 and Rule 4
$Buku_{1[KN]}$ $kulit_{2[KN]}$ $nipis_{3[KA]}$ $harga_{4[KN]}$ $murah_{5[KA]}$ $itu_{6[KN]}$

Compound noun results :
compoundNoun$_3$ = word$_{2[KN]}$ + word$_{3[KA]}$
compoundNoun$_3$ = ($kulit_{2[KN]}$, $nipis_{3[KA]}$)

compoundNoun$_4$ = word$_{4[KN]}$ + word$_{5[KA]}$
compoundNoun$_4$ = ($harga_{4[KN]}$, $murah_{5[KA]}$)

⇩

Output:

compoundNoun$_1$(***kulit, nipis***)
compoundNoun$_2$(***harga, murah***)

Figure 1.Process Flow for Nominating Compound Noun.

By using these rules, the accuracy of choosing the compound noun become more acceptable, particularly when the pair of words can be found from a compound noun database. If the pair of words is not present in a database, the compound noun can be selected with the help of the rules formulated. To verify the correctness of compound noun suggested, a linguist was required to discern the correctness of compound noun results. This task is important to check the validation of rules formed.

V CONCLUSION AND FUTURE WORK

We discussed an important issue of creating noun phrase structure form to generate a list of rules to be used in detecting compound nouns in Malay sentence. Every noun phrase structure form has its own specific rules according to the sentences studied. The collection of compound nouns in Malay sentence is also being constructed and kept in a database. This data is needed for matching with the group of compound noun returned from the rules.

Research on compound noun for Malay is a necessary task, especially in developing natural language processing application. The contribution this task will be able to solve the major problem encountered in natural languages, such as resolving the disambiguation of words in sentence.

To continue our research work, we plan to have the following task in the next step of our research.

i. To study and analyse the Malay sentence examples to be used for creating noun phrase structure form.

ii. To test and evaluate the approach or technique used in detecting compound noun in Malay sentence.

ACKNOWLEDGMENT

We would like to thank Ministry of Higher Education (MOHE) for the Fundamental Research Grant Scheme (FRGS 2011).

REFERENCES

Addullah Hassan. (2004). *Tatabahasa Bahasa Melayu*, 4th edition, PTS Publications & Distributors Sdn. Bhd.

Abdullah Hassan. (1992). *Linguistik Am*, 10 edition., PTS Profesional Publishing Sdn.Bhd.

Arbak Othman, Nik Safiah Karim. (2006). *Kamus komprehensif Bahasa Melayu*, Cetakan kedua, Penerbit Fajar Bakti Sdn.Bhd.

Dinh Dien. (2002). *Cognitive linguistics approach to Vietnamese noun compounds*, Mon-Khmer Studies 32, pp. 145-162.

Michael, N. (2002). *Artificial Intelligence: A guide to Intelligent Systems,* Second edition,PearsonEducation Limited.

Nik Sapiah Karim, Farid M.Onn, Hashim Haji Musa, Abdul Hamid Mahmood, (2010). *Tatabahasa Dewan* , 3rd edition (cetakan kelima), Dewan Bahasa dan Pustaka (DBP), Malaysia.

Ong Ching Guan, (2009). *Kuasai Struktur Ayat Bahasa Melayu,*Dewan Bahasa dan Pusataka (DBP), Malaysia.

Suhaimi Ab Rahman, Nazlia Bt Omar, Noor Baizura Che Hassan. (2012). *Construction of Compound Nouns (CNs) for Noun Phrase in Malay Sentence.* In International conference on Information Retrieval and Knowledge Management, CAMP'12 (Capaian Maklumat dan Pengurusan Pengetahuan), Mines Wellness Hotel, Malaysia.

Suhaimi Ab Rahman, Nazlia Omar and Mohd Juzaidin Ab Aziz. (2011). *"Transformation of Malay Head Modifier Noun Phrase into a Thematic Relation Structure"* Proceedings of the 2011 International Conference on Electrical Engineering and Informatics, Bandung-Indonesia (Volume 3), PP 1775-1779.