Pola Grammar Technique to Identify Subject and Predicate in Malaysian Language

Mohd Juzaiddin Ab Aziz
Fakulti Teknologi & Sains Maklumat,
Universiti Kebangsaan Malaysia,
43600 Bangi,
Selangor, Malaysia
din@ftsm.ukm.my

Fatimah Dato' Ahmad
Fakulti Sains Komputer & Teknologi Maklumat,
Universiti Putra Malaysia
43400 Serdang,
Selangor, Malaysia
fatimah@fsktm.upm.edu.my

Abdul Azim Abdul Ghani
Fakulti Sains Komputer & Teknologi Maklumat,
Universiti Putra Malaysia
43400 Serdang,
Selangor, Malaysia
azim@fsktm.upm.edu.my

Ramlan Mahmod
Fakulti Sains Komputer & Teknologi Maklumat,
Universiti Putra Malaysia
43400 Serdang,
Selangor, Malaysia
ramlan@fsktm.upm.edu.my

Abstract

The Malaysian Language is a formation of subject, predicate and object. The subject is the noun that take the action on the object and the predicate is the verb phrase in the sentence. Without a good corpus that can provide the part of speech, parsing is a complex process. As an option to the parsing, this paper discusses a way to identify the subject and the predicate, known as the pola-grammar technique. A pola or a pattern to be identified in the sentence are the Adjunct, Subject, Conjunction, Predicate and Object.

1 Introduction

The Malaysian language is a context free grammar where there is a subject and a predicate (Nik Safiah et. al., 1993). According to the research done by Azhar (1988), there are three types of Malaysian language context-independent grammar. One is sentence grammar (Nik Safiah, 1975) and (Yeoh, 1979), the second is the partial discourse grammar (Asmah, 1980) and the third is 'pola' sentence (Asmah, 1968).

Asmah (1968) worked on pola grammar was accepted as a standard format for the Malaysian language's grammar before it was replaced by the transformational-generative type grammar (Nik Safiah, 1975). The pola that were suggested by Asmah (1968) are:

i. Pelaku + Perbuatan (Actor + Verb)
ii. Pelaku+Perbuatan+Pelengkap (Actor + Verb + Complement)
iii. Perbuatan + Pelengkap (Verb + Complement)
iv. Diterangkan + Menerangkan (Signified + Signified)
v. Digolong + Penggolong (Classified + Classifier)
vi. Pelengkap + Perbuatan + Pelaku (Complement + Verb + Actor)
vii. Pelengkap + Perbuatan (Complement + Verb)

From the list above, the pola that used to start a sentence are: *pelaku* (actor), *perbuatan* (verb), *diterangkan* (signified), *digolongkan* (classified) and *pelengkap* (complement). *Pelaku* (actor) is a noun and *perbuatan* is a verb. The words *diterangkan* (signified), *digolongkan* (classified) and *pelengkap* (complement) are adjuncts. An adjunct is an argument that has a less tightly related to the subject and predicate. They do not represent the subject, verb or the object of the sentence.

Abdullah (1980) modified the ealier version of pola grammar (Asmah,1968) with a new set of pola. The examples of the Abdullah's pola are noun+noun and noun+verb.

2 Pola Grammar

The term pola refers to 'pattern' as in "sentence pattern". Asmah et al. (1995) use the regular expression representation of Nn, N_1, A, N_1, V.*. to represent the pola. Nik Safiah et. al (1993) use the format of Noun Phrase (NP) + Noun Phrase (NP), Noun Phrase (NP) + Verb Phrase (VP), Noun Phrase (NP) + Adjective Phrase (NP), Noun Phrase (NP) + Preposition Phrase (PP) to show the basic format of the language, which consist of pola. These format will be used as the basic to identify the subject and predicate. There will be more pola added into the format, they are *adjunct, subject, postSubject, conjunction,* and *predicate.*

The subject is either a noun, pronoun or a verb that functions as a noun, or an adjective that functions as noun. The postSubject describes the subject. In Malaysian language, the postSubject is normally starts with the word 'yang' or 'dengan'. The conjunction represent the words that join words or phrases or sentences together. The predicate is a theme that says something about the subject.

2.1 The Rules

From a basic pola, the components are inserted into the to produce a new rules. Examples of the rules are:

[Adjunct + (NP)1 + conjunction] + [(NP)2]
(NP)1 → Noun + postSubject
(NP)2 → Predicate
Predicate → object -- rule (1)

[Adjunct + (NP) + conjunction] + [(VP)]
(VP) → Predicate
Predicate → verb + object + (adverb)1
(adverb)1 → conjunction + object + conjuction + (adverb)2 -- rule (2)

Table 1a, 1b and 1c. show the examples of the Malaysian sentences in the pola format. The sentences used in the table are the sentences taken from (1), (2) and (3) as below:

Example of sentences,

Pengkompil menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah. -- (1)

Tujuan pengkompil adalah untuk menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah. --(2)

Walaupun pengkompil menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah, tetapi, tugas utamanya adalah untuk menyemak sintaks bahasa. -- (3)

Table 1a : the pola for sentence (1)

Sentence	*(1)*
Adjunct	*Null*
Subject	*Pengkompil*
PostSubject	*Null*
Conjunction	*Null*
Predicate	*Menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah*

Table 1b : the pola for sentence (2)

Sentence	*(2)*
Adjunct	*Tujuan*
Subject	*Pengkompil*
PostSubject	*Null*
Conjunction	*Adalah untuk*
Predicate	*Menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah*

Table 1c : the pola for sentence (3)

Sentence	*(3)*
Adjunct	*Walaupun*
Subject	*Pengkompil*
PostSubject	*Null*
Conjunction	*Null*
Predicate	*Menukar Bahasa Paras Tinggi kepada Bahasa Paras Rendah, tetapi, tugas utamanya adalah untuk menyemak sintaks bahasa.*

The pola shows that "Pengkompil" is the subject of sentences (1), (2) and (3), even though there are no adjunct and conjunction in sentence (1). The sentence (3) do not has a conjunction, but the predicate is longer than the predicates in sentence (1) and (2).

To test the design, let take the sentence (1) as the input.

"*Pengkompil menukar bahasa paras tinggi kepada bahasa mesin*".

Step 1

Choose a basic format --- rules 2,

Step 2

Identify the pola of the adjunct, subject, postSubject, conjunction and predicate.
Subject (*Pengkompil*) Predicate[*menukar bahasa paras tinggi kepada bahasa mesin*].

Step 3

Identify the pola of the verb, object, conjunction and adverb.
Predicate: Verb (*menukar*) Object (*bahasa paras tinggi*) Conjunction (*kepada*) Adverb(*bahasa mesin*)
Adverb : Object (*bahasa mesin*)

3 Related Work

Rosmah (1997) developed an algorithm to derive Malaysian language using the Context Free Grammar (CFG) rules and a parse tree. The CFG was initially developed by Nik Safiah(1975), followed by Yeoh (1979).

The derivation by Rosmah (1997) identified the subject and the predicate in a simple Malaysian sentences. To do that, there was a module to identify the lexical values, such as a noun, a verb and an adverb.

The major problem occurs in the process is to solve the ambiguities. There are a lots of Malaysian words that can be either a verb or a noun. For example, the word '*mereka'* can be either a pronoun ('*they'*) or a verb ('*design'*) and the word '*pohon'* is either a verb ('*request'*) or a noun ('*tree'*). As the result, the parsing was very costly and it easily produced a wrong syntax tree.

The second problem was the problem to parse a complex CFG rules for the compound sentences when there is no sign to stop. For instance, based on sentence (4) and the Context Free Grammar below:

Sentence → Subject + Predicate
Subject → Noun Phrase
Predicate → Verb Phrase
Noun Phrase → Noun
Verb Phrase → Verb

Nod B yang mempunyai nilai terendah berada berhampiran dengan Nod A merupakan Nod yang paling sesuai untuk dilalui. --(4)

The terminal for the Verb Phrase is a verb. In sentence (4), the verb is the word "*mempunyai*" where it is in the third position of the sentence. If the parser select "*mempunyai*" as a verb, it will cause an ill-grammar problem because the word is actually in the noun clause. This is due to the fact that some clauses which a word "yang" do not has a sign to stop (Azhar, 1988). The pola grammar techniques solve this problem by introducing a pola called postSubject.

4 The Model

The sequence of the pola in the sentence is shown as below:

Adjunct + Subject + postSubject + conjunction + conjunction + predicate (Output)

A finite automaton will be used to recognize the pola. It is a mathematical model of a system represented as:

(Q, ∑, S, R), where

Q is a finite set of states
∑ is a finite set of input
S is the initial state
R is the transition relational which maps the input and states

The states are the adjunct, subject, postSubject (postSub), conjunction (conj), and predicate. They are the pola used in this study. The input will be a list of words in the sentence.

The transition relational capture the pola based on the rules of the pola grammar technique. The algorithm of the transition relational is shown as follow:

Case 1 – adjunct
If input is

1. subornating, insert adjunct
2. classifier, insert adjunct
3. numeric, insert adjunct
4. ini, itu or "," insert adjunct
    a. if "," insert adjunct

start at subject
5. verbs, insert adjunct
   a. if start = adjunct, insert subject
      start subject,
      prStart adjunct
   b. else insert predicate
      start predicate,
      prStart adjunct
6. yang, itu, ini insert adjunct
   a. if j = 1 while not "adalah" or not "ialah"
   b. if "." insert null {stop process}
      i. else insert adjunct
         start conj
         prStart adjunct
   c. else while jumpa = false and j <= no of token
   d. if "." insert null {stop process}
      i. elseif ","
         insert adjunct,
         jumpa true,
         start subject,
         prStart adjunct
      ii. elseif "itu", "ini"
         insert adjunct,
         jumpa true,
         start subject,
         prStart adjunct
         1. if ","
            insert adjunct
         2. else j = j –1
      iii. else {not "," or itu, ini}
         insert adjunct
         start adjunct
   e. wend
7. if "," insert adjunct
   start subject
8. else insert subject
   start subject,
   prStart subject

Case 2 – subject
If input is

1. "yang" or "adalah" insert postSub
   start postSub
   prStart subject

2. "itu" or "ini" or "tersebut" insert subject
   start subject
   prStart subject
3. Ncon insert subject
   start conj
   prStart subj
4. conjuction insert conj
   start conj
   prStart conj
5. subornating insert predicate
   start predicate
   prStart subject
6. verbs or nafi
   if prStart = adjunct or prToken = "dan" or "atau" insert subject
   start subject
   prStart subject
   else insert predicate
   start predicate
   prStart subject
7. "," insert subject
   i. if lookahead (j, 7, ",") TRUE start subject
      prStart subject
   ii. else if lookahead (j, 7, "and") TRUE
      start subject
      prStart subject
   ii. else
      start subject
      prStart subject
8. else insert subject
   start subject
   prStart subject

Case 3 – postSubject
If input is

While token <> "," and jumpa = False
and token < ListCount

1. "." Start 6,
   prStart postSub
2. "ini" or "itu" insert postSub
   start conj
   prStart postSub
   jumpa TRUE
3. conjunction insert conj
   start conj
   prStart postSub
   jumpa TRUE
4. else insert postSub
   start postSub
   prStart postSub
   jumpa TRUE

Wend

5. if "," insert postSub
start conj
prStart postSub
6. if Ncon insert postSub
start postSub
prStart postSub
7. verbs insert predicate
start predicate
prStart postSub

Case 4 – conjunction
If input is

1. conjunction insert conj
start conj
prStart conj
2. subornating insert conj
start conj
prStart conj
3. else insert predicate
start prediacte
prStart predicate

Case 5 – predicate
If input is

While token <> "." And token < ListCount
Insert predicate
wend

Case 6 – stop
End

5 Testing

The algorithm was tested with thirteen (13) abstracts' thesis, Masters in Computer Science and Information Technology from Faculty of Technology and Information Science. The total number of sentences used in the testing were one hundred and twelve (112).

The test show that 6 sentences do not produce a precise results. The sentences are as follow:

1.
Adjunct : *Adalah didapati bahawa*
Subject : *penyelesaian masalah jadual waktu*
PostSubject :
Conjuction : ***dengan***
Predicate : ***komputer*** *memerlukan satu pemindahan paradigma*

The words 'dengan' and 'komputer' should be as a part of the subject.

2.
Adjunct :
Subject : *Sasaran*
PostSubject : *yang tidak tentu* ***tidak***
Conjuction : ***akan***
Predicate : *mewujudkan penyelesaian yang lengkap*

The words 'tidak' shows the negative of 'akan'. So, they should be together ini the conjunction.

3.
Adjunct : *Pada*
Subject : *peringkat awal dan*
PostSubject :
Conjuction : ***pada***
Predicate : ***peringkat akhir****,penjanaan jadual waktu dengan komputer masih memerlukan penglibatan penskedul jadual*

The words 'pada', 'peringkat' and 'akhir' should be the second subject.

4.
Adjunct :
Subject : ***Aliran***
PostSubject :
Conjuction :
Predicate : ***kerja*** *boleh ditakrifkan sebagai satu kaedah untuk mengautomasikan dan mengawal pergerakan proses yang melibatkan sekurang-kurangnya dua entiti bergerak dari sati entiti secara turutan atau serentak berpandukan pada syarat-syarat yang telah ditetapkan bagi mencapai matlamat yang sama*

The words 'aliran' and 'kerja' are nouns. In this 'kerja' sentence, it was interpreted as a verb by the program.

5.
Adjunct : *Tetapi,*
Subject :
PostSubject :
Conjuction :
Predicate : ***didapati pelaksanaan pembelajaran dengan paten-paten yang agak besar*** *mewujudkan kesilapan yang agak besar yang menghadkan proses penjanaan sistem pengetahuan domain tersaur*

This sentence contains the word '*didapati'* which was interpreted as a verb. It is actually an adjunct where the subject is "perlaksanaan pembelajaran".

6.
Adjunct :
Subject : *Sistem Pengurusan Maklumat Makmal Kimia*
PostSubject :
Conjuction :
Predicate : ***Berasaskan Multimedia:*** *Satu Kajian Kes dibangunkan untuk tujuan pengurusan stok bahan kimia peralatan dan radas yang digunakan di makmal kimia sekolah menengah*

The subject of the sentence should be "Sistem Pengurusan Maklumat Makmal Kimia Berasaskan Multimedia.

6 Analysis

The results show that the pola sentence can be used to clarify the subject and predicate in the Malaysian sentence. The problems occurs in the 6 sentences were caused by :

a. The existing of the conjunction '*dengan'* in the subject. The words that follow this word can either be as a postSubject or a subject.
b. The nouns are varied and do not have a common pattern.
c. The words '*tidak'*, to show a negative sentence do not locate in the right position.
d. The verbs that act as a noun.

Problem (b) can be fixed by supplying the noun information to the application. Problems (c) and (d) can be fixed by improving the algorithm. Problem in (a), needs further studies and enhancement due to the fact that the word '*dengan'* can be either a conjunction or a word to describe its' subject.

7 Conclusion

A pola grammar was excepted as a formal grammar for the Malaysian language. But, the Chomskyian revolution makes the linguist to produce a Context Free format for the language. For computational purposes, good corpus is needed to provide the information in order to parse the language, for instance to provide the correct lexical values. A corpus such as WordNet (Fellbaum, 1998), will reduce the problems such as ambiguity and backtracking.

Since there is no such corpus in Malaysian language, a pola grammar technique is introduced to identify the grammatical relation for the language. The result discussed in this paper proved that the pola grammar can extract the subject, verb and object.

References

Abdullah Hassan. 1980. *Linguistik Am Untuk Guru Bahasa Malaysia*. Penerbit Fajar Bakti, Kuala Lumpur.

Asmah Haji Omar and Rama Subbiah. 1995. *An Introduction To Malay Grammar*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Asmah Haji Omar. 1980. *Nahu Melayu Mutakhir*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Asmah Haji Omar. 1968. *Morfologi-sintaksis Bahasa Melayu (Malaya) dan Bahasa Indonesia: Satu Perbandingan Pola*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Azhar M. Simin. 1988. *Discourse-Syntax of "YANG" in Malay (Bahasa Malaysia)*. Dewan Bahasa dan Pustaka, Kuala Lumpur.

Fellbaum.C. 1998. *WordNet: An Electronic Lexical Database.* The MIT Press, Cambridge, Massachusetts.

Nik Safiah Karim.1975. *The Major Syntactic Structures of Bahasa Malaysia and their Impilcations of Standardization of the Language*. Ph.D. dissertation. Ohio University, USA.

Nik Safiah Karim, Farid M. Onn, Hashim Hj. Musa, Abdul Hamid Mahmood. 1993. *Tatabahasa Dewan, Edisi Bahar.* Dewan Bahasa dan Pustaka, Kuala Lumpur.

Rosmah Latif. 1997. *Sintaksis Ayat Bahasa Malaysia*. Tesis Sarjana, Universiti Kebangsaan Malaysia, Bangi.

Yeoh, Chiang Kee. 1979. *Interaction of Rules in Bahasa Malaysia.* Ph.D. dissertation, University of Illinois at Urbana-Champaign, USA.