An Analysis of Indonesian Language for Interlingual Machine-Translation System

Hammam R. Yusuf*

Computer Science Department
University of Kentucky, Lexington, KY 40506
yusuf@ms.uky.edu

ABSTRACT

This paper presents BIAS (Bahasa Indonesia Analyzer System), an analysis system for Indonesian language suitable for multilingual machine translation system. BIAS is developed with a motivation to contribute to on-going cooperative research project in machine translation between Indonesia and other Asian countries. In addition, it may serve to foster NLP research in Indonesia. It start with an overview of various methodologies for representation of linguistic knowledge and plausible strategies of automatic reasoning for Indonesian language. We examine these methodologies from the perspective of their relative advantage and their suitability for an interlingual machine-translation environment. BIAS is a multi-level analyzer which is developed not only to extract the syntactic and semantic structure of sentences but also to provide a unifying method for knowledge reasoning. Each phase of the analyzer is discussed with emphasis on Indonesian morphology and case-grammatical constructions.

1. Introduction

Bahasa Indonesia (Indonesian language) is a national language for the Republic of Indonesia which unites 27 cultural backgrounds. It is widely used by more than 100 millions speaker but unfortunately, does not gain much attention for its automatic processing by computers. In 1987, a cooperative research in machine translation with Japan sparks the natural language processing research in Indonesia. In support to the on going project of Multilingual Machine Translation Sytem for Asian Language organized by Center for International Cooperation in Computerization (CICC)-Japan and other Asian countries (China, Indonesia, Malaysia and Thailand), we developed BIAS : an analysis program for Indonesian language which output an interlingual representation. By incorporating interlingual analysis technology, we will be able to include BIAS as part of multi-language translation system in a very effective way.

This paper describes the design consideration of BIAS from the view point of linguistic theories and knowledge representation formalism. The design is based on an interlingual approach to machine translation which accepts input sentences in one language and produces sentences in other languages [Figure 1]. In particular BIAS is a program that takes natural language text as input and produces its underlying interlingual representation at a certain level of details that serve as a language-independent representation for the machine translation environment.

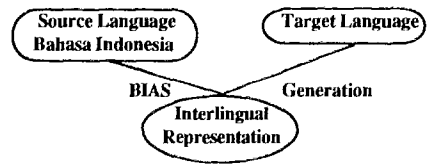

Figure 1. BIAS and Interlingual Approach to MT

The approach which being used here is an approximation of basic linguistic theories such as Chomsky's Standard Theory

* Supported by the Agency for the Assessment and Application of Technology, Jakarta, Indonesia

[Chomsky,65], Case Grammar [Fillmore, 67] and Definite Clause Grammar [Pereira,80]. We also incorporates the use of appropriate representation formalism such as Frames [Minsky,81] and Semantic Network [Quillian,68] for a suitable type of reasoning system. It is noted that eventhough there are many knowledge representation languages which are theoretically sufficient to describe any natural language, they need to be modified in their theory and implementation for a particular language such as Bahasa Indonesia. The existence of various theories and knowledge representation techniques lead us to consider several models of reasoning formalism. This in turn may serve as an indicator to the expressive adequacy of our chosen representation.

The rest of this paper is organized as follows. Section 2 presents framework of BIAS from the view point of linguistic theories. We will discuss in detail the language analysis method of BIAS. It will be followed by a discussion on representation formalism and reasoning techniques in Section 3. The paper ends with a conclusion.

2. Analysis Method

There are many ways of attacking the problem of natural language processing. At one end of the spectrum are analyzers that read the input sentences, very closely following every twist in syntax, trying to interpret every bit of information contained in the sentence. In most cases, these analyzers separate the syntactic and semantic parts of the analysis into separate consecutive stages, paying much more attention to the syntactic part at the expense of semantic [Gersham,82]. At the other end are the analyzers that skim through the text looking for certain types of information and paying attention only to the words and expression relevant to the task [DeJong,79]. This approach is very effective and intuitively corresponds to what people do while skimming newspaper stories. However, the danger in this approach lies in the possibilty of misunderstanding what is being stated.

BIAS is a multi-level analyzer, similar to the first type describe above, with the ability to perform reasoning in each level of analysis. The method used in BIAS is theoretically consistent with the Standard Theory and Case Grammar as well as non-monotonic reasoning formalism. The process starts with analysis of sound sequences and ends by producing its interlingual representation. In-depth discussion on each of the analysis phase in BIAS and the selection of appropriate linguistic theories follows.

2.1 Morphological Analysis Phase

Preliminary analysis of Indonesian words poses an especially difficult problem : the transformation of word category and its meaning as the result of affixation. Although it seems better to segment input sentences beforehand, it is not natural in the general sense to do this process on the first basis. We have to combine the processes of phonological and morphological analysis in order to extract the root word from an inflected form.
The process will involve the following : an inflected word is analyzed to give its root word and affixes, allowing the system to recognize the altered structure and meaning of the inflected word. This phase uses the lexicon, morphological and phonological knowledge in the form of transformation rules.

Further, we observed the following word formation rules which indicate their characteristics :
(a) A word can be constructed using prefix, suffix or confix.
(b) A word can be constructed using a repetition of root word as in 'kura-kura' (turtle), or repeating the word constructed in (a) as in the case of 'berlari-lari' (jogging).

Our analysis showed that the complex types of word formation could lead to some problems while constructing the structure of the lexicon [Yusuf,88]. It is evident that in the lexicon, a word should be described briefly, so that the search can be efficient. Hence, the lexicon should contain only a simple form of word which, in this case, is the root form.
How can we deal with a word with affixation ? In our findings, the word with affixation could be processed by using the following procedure.

Algorithm : Morph()

Input : word
Output : root word, affixation and semantic markers

- Assume that the word is a root word.
If this word is in the dictionary, check whether it is in its root form or purely repetitive form.
- Assume that the word is a word with some prefix.
Check for the following conditions :
- The root word is repetitive word and not an idiom with affixation.
For example : *berlari-lari (jogging).*
- The word with affixation and repetition
For example : *berpukul-pukulan (hit reciprocally)*
- A root word with affixation or idiom with prefix.
For example : *pekerjaan (occupation)*
bertanggung-jawab (responsible for)
- Idiom with sufix or confix.
For example : *pertanggung-jawaban (responsibility)*

Table 2 summarizes the morphology rules which have been formulated in BIAS. These rules are basic ; other rules which incorporate complex formation of words (see also [Tarigan, 84]) are being left for futher improvement. The general structure for a morphological rule of a given root word is described as follow :

```
([Affix] + [Root Word + Semantic])
          ---> [Word + NewSemantic]
```

Examples :
([mem + [*pukul* + action]) ---> [*memukul* + active]
([mem-i + [*pukul* + action]) ---> [*memukuli* + repetitive]
([mem-kan + [*pukul* + action]) ---> [*memukulkan* + causative]
([ber-an + [*pukul* + repetition] ---> [*berpukul-pukulan* + reciprocal action]

Table 2. Indonesian Morphological Construction

Root form	Prefix	Sufix	Confix	Compund Term	Semantic
pukul (hit)	me			memukul	active
bawa (carry)	di			dibawa	passive
nama (name)	ber			bernama	possesive
perlu (need)			me-kan	memerlukan	active tran.
baca (read)			di-kan	dibacakan	passive
pegang (touch)	ter			terpegang	accidental
guna (use)			ter-kan		implicative
main (play)	memper	kan			purpose
daya (trick)	terper	kan			accidental

Table 3. Phonological Rules

Prefix	Root	Inflection
CeN	buat	membuat (make)
	goreng	menggoreng (fried)
	kurang	mengurang (substract)
	tunggu	menunggu (wait)
	sapu	menyapu (sweep)
	cukur	mencukur (shave)
	pukul	pemukul (hitter)
	hasut	penghasut (agitator)
ber	usaha	berusaha (effort)
	ruang	beruang (room)
	uang	berjuang (have money)
	ternak	beternak (lifestock)

C = consonant of m and p
N = phonological transformation

The new semantic of formed word is derived from the semantic of root word and affixes. There are several filters being used for extraction of this semantic. In the examples, *mem-i* cause the word *pukul* which has *action* as its original semantic to become repetitive in its meaning when combined.

In addition to morphological construction as described above, there are phonological rules which are handled in parallel in the morphological analysis phase. The phonological rules determine the transformation of phonetic structure of a root word for a given complex word. We include some examples to show its construction as in Table 3.

2.2 Syntactic Analysis Phase

This phase covers those steps that affect the processing of sentences into structural descriptions or syntactical tree by using a grammatical description of linguistic structure. The major components are syntactic knowledge (grammar rules) and lexicon. There are several linguistic phenonema worth describing for Indonesian language. For instance, the language structure of Bahasa Indonesia has a different structure compared to English and other languages. One of the most significant difference is that the Indonesian language apply various rules to construct Adverb Phrase, Adjective Phrase and Relative Clauses.

For example, in constructing Adverb Phrase, it is allowed to combine adverb and adjective in addition to adverb and verb. It is also possible to form Adjective Phrase using adjective followed by noun rather than the default order of noun and adjective. This notion resulted from the categorial ambiguity of some words.

Examine the following phrases :

rumah (N)	*merah* (Adj)	*panjang* (Adj)	*tangan* (N)
(house)	(red)	(long)	(hand)
cepat (Adv)	*merah* (Adj)	*berjalan* (V)	*cepat* (Adv)
(quickly)	(red)	(walk)	(quickly)

BIAS use a bottom-up technique [Matsumoto,83] in the syntactic analysis phase. The grammar rule written in Extraposition Grammar [Pereira,81] is translated to a set of Horn clauses which

will parse a sentence according to the original grammar in bottom up and depth first manner.

2.3 Semantic Interpretation Phase

This phase will consist of the mapping of the structural (syntactic) description of the sentence into an interlingual representation language. The goal of this phase is to construct a clear representation of the exact meaning of a given sentence; hence, it is a language-independent representation suitable for a generation process of target languages. In order to achieve this, we need commonsense knowledge, in addition to semantic knowledge.

In Bahasa Indonesia the verbal elements of the sentence are the major source of the structure: the main verb in the proposition is the focus around which the other phrases, or cases, revolve and the auxiliary verb contain much of the information about modality. Hence, the Case grammar is the appropriate selection for the semantic analysis part.

Case frame are the mechanism for identifying the specific cases allowed for any particular verb. The case frame for each verb indicates the relationships which are required in any sentence in which the verb appears and those relationship which are optional.

Let us look at some popular example sentences :

Palu itu memukul paku itu.
(the hammer) (hit) (the nail)

Paku itu dipukul oleh palu itu.
(the nail) (was hit) (by) (the hammer)

Seseorang memukul paku itu dengan palu itu.
(someone) (hit) (the nail) (with) (the hammer)

The verb, *memukul*(hit), allows three primary cases: agentive, instrumental and objective. We have all three cases in the last sentence, but only two in the others. In fact, only one case is required with this verb,

Paku itu dipukul.
(the nail) (was hit)

Thus the case frame for the verb *memukul*, by default :

[memukul [O (A) (I)]]

Further, some other case frames are also determine for words which combine *pukul* and other affixation, as in the case of *memukulkan, memukuli, memukul-mukulkan*, etc.

In addition to the standard cases described by Fillmore and Simmons [Simmon,73], we incorporate several other cases found in Indonesian language. These cases occur as the result of word inflection. For instance the confix *meN-kan*, with the root word *beli* create a word, *membelikan*, which carry the meaning of "being beneficiary of the action". Some examples of these case-specific can be found in the following sentences :

1. Benefactive : *Saya membelikan adik boneka* (I buy a doll for sister)
2. Incidental : *Adi terpeleset di tangga* (I felt on the stair)
3. Causative : *Saya mempertanyakan masalah itu.* (I questioned that problem)
4. Intentional : *Saya memperdayai dia.* (I tricked him)

The interligual representation for (1) is given in Figure 5. Note that each word is represented by a concept and its attributes.

3. Representation and Inference

We have come to a point to discuss various types of representation language being used to represent the theories in each phase of the analysis.

In the morphological analysis phase, it is appropriate to represent the morphology and phonological rules with definite clauses which have first order logic as its basis. First order logic provides a clear language to represent propositions or facts for the lexicon and also supports production-like rules for the transformation

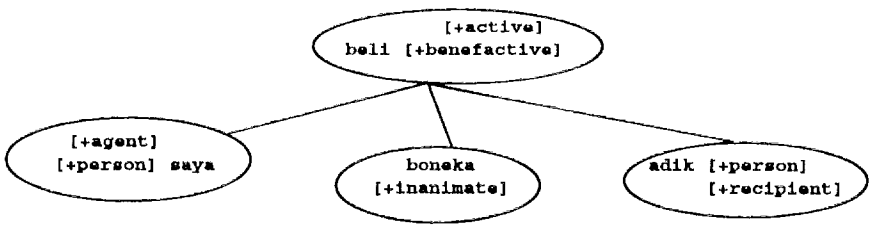

Figure 4. Example of Interlingual representation

Table 5. Level of Analysis, Representations and Inferences in BIAS

Analysis Phase	Theory	Representation	Inference
Phonology	Standard Theory	Definite Clause / First order logic	Deduction/ Induction
Morphology	Standard Theory	Definite Clause / First order logic	Deduction/ Induction
Syntactic	Extended Standard Theory	Definite Clause	Deduction Default
Semantic	Case Grammar	Semantic Network with Slot Filler	Default

rules. The syntactic part adapts the Extended Standard theory and hence, it is favorable to use first order logic to represent its knowledge. The use of Case Grammar in semantic analysis phase leads us to choose the network-based formalism as the representation. Simmons and Hendrix [Simmons,73] have provided a clear language for semantic network based on the Case Grammar. However, we also incorporate 'slot fillers' from the frames system [Minsky,81] as a solution to handle incomplete sentences.

As the consequences of the selection of the representation method for the linguistic knowledge of Bahasa Indonesia, BIAS have multiple inference methods incorporate in each level of analysis. In syntactic and semantic analysis phase, default reasoning is performed to solve the problem of incomplete knowledge. In this case, first order logic must be augmented with default operators in order to permit non-monotonicity. [Reiter,78]

Because of space limitation, we leave out in-depth discussion on inference techniques (see[Yusuf,91] [Schubert,79]), and present our summary of work in Table 5.

4. Conclusion

The use of linguistic theories and appropriate knowledge representation techniques provide BIAS a new insight in attacking the problem of language analysis for interlingua machine translation system, especially for Bahasa Indonesia. Many representation formalism and reasoning system have been brought into consideration not only for a 'pure' sentence analysis but in order to design an effective and efficient intelligent system capable of capturing and reasoning with linguistic knowledge.

Reference

Brachman, Ronald J., On epistemological status of Semantic Network, in Associative Networks : Representation and Use of Knowledge by Computer, Academic Press, New York, 1979.

Chomsky, Noam, Aspects of the Theory of Syntax, MIT Press, Cambridge, 1965.

DeJong,G.F., Skimming stories in real time : An experiment in integrated understanding (Computer Science Report No.150), Yale University, New Haven , May 1979.

Fillmore, Charles, The Case for Case, Universals in Linguistic Theory, Ed Emmon Bach and R.T. Harms, Holt,Rinehart, Winston, New York, 1-88, 1967.

Gersham, A.V., A Framework for Conceptual Analyzer in Strategies for Natural Language Processing, Lawrence Erlbaum, 1982.

Lockman, Abe and Klapphoz, David, The Control of Inferencing in NLU, Computational Linguistic, ed. Cercone, Nick., Pergamon Press, Oxford, 1983.

Matsumoto, Y. et. al., BUP : A Bottom Up Parser Embedded in Prolog, New Generation Computing Vol.1 No.2, pp.145-158,1983.

Minsky, M., A Framework for Representing Knowledge, Mind Design, pp.95-128, MIT Press, 1981.

Pereira, F and Warren, D., Definite Clause Grammars for Language Analysis— A Survey of the Formalism and a Comparison with Augmented Transition Networks, Journal of Artificial Intelligence 13 (1980) pp.231-278.

Pereira, F., Extraposition Grammars, American Journal of Computational Linguistics Vol.7 No.4 (1981) pp.243-256

Quillian, M.R., Semantic Memory, Semantic Information Processing, Ed. Marvin Minsky, MIT Press, Cambridge, 1968.

Reiter, R., On Reasoning by Default, Proc. TINLAP-2, Theoretical Issues in Natural Language Processing-2, University of Illinois at Urbana-Champaign, 210-278, 1978.

Schubert,Lenhart K., Randolph G.Goebel and Nicholas J. Cercone, "The Structure and Organization of a Semantic Network for Comprehension and Inference", Associative Networks, 121-175, Academic Press, 1979.

Simmons, Robert F.,Semantic Network : Their Computation and Use for Understanding English Sentences, Computer Models of Thought and Language, W.H.Freeman Co., San Francisco, 1973.

Tarigan,S., Morfologi Bahasa Indonesia, Penerbit Gramedia, Jakarta- Indonesia, 1984.

Yusuf, Hammam, Indonesian Electronic Dictionary, Technical Report, Agency for the Assessment and Application of Technology, Jakarta, 1988.

Yusuf, Hammam, Analyzer for Bahasa Indonesia, Master Thesis, University of Kentucky, 1991.