

Final Report on POS Tagset

Abstract

Part-Of-Speech tagging is a process that attaches each word in a sentence with a suitable tag. The result of the tagging is important for high-level natural language processing researches. Before the Part-Of-Speech tagging starts we must have tagset first. This report is a research on Khmer Part-Of-Speech Tagset. We capture the rule from Khmer Grammar book, Khmer Dictionary and some documents related to Khmer grammar to make the tags for Khmer Language.

1. Introduction

Defining Part-of-Speech (POS) tagsets is very essential for some Asian Languages such as Khmer language since those POS tagsets have their own benefit, i.e. serving as a basis for the development of such various areas of Natural Language Processing (NLP) as Speech Synthesis (Pronunciation), Text-to-Speech (TTS), Grammar Checker, Information Retrieval, Translation, Corpus Analysis of Language and Lexicography, Word-sense disambiguation, etc.

In this paper, we start by presenting our purposes in doing the research on Khmer POS. Then we continue to talk about our scope. After that some major issues are presented concerning the research. Next we show the table of all the tagsets we defined. At last we conclude what we have done.

2. Purpose

The research on Khmer Grammar is so widely conducted that different authors of various grammar books define divergent types of Part-of-Speech (POS). Our works on Part-of-Speech (POS) tagsets are to define the tagsets and to reach the general consensus of how many of them should be included. Obviously, POS is very significant that it will inevitably be used as the basis for other areas of Natural Language Processing (NLP). Our works on POS tagsets will not be only the initial step towards standardized Khmer POS Tagging development, but also the foundation for the development of various areas of Natural Language Processing (NLP).

3. Scope

The tagsets we defined are only those of the major ones based on three main materials listed in the references section. We do not include every minority POS (POS that is the division of the major POS) defined by other authors because we find it hard to consider them as standardized ones due to some debates over those POSs.

4. Major issues

During the research, we seemed to lack of support documents. Our work is done based solely on the three main materials used. Owing to the ambiguity of some tags, we decided to eliminate some tags which might result in the loss of some POS names. As a result, those tag names will be replaced by their parent's name. For instance, "Adjective of Quality" will be replaced by "Adjective".

5. Results

In consequence of the analysis, 21 tagsets are defined and documented in accordance with a set of rules discussed in the materials used. The following is a table of all the 21 tagsets.

ategory	Remarks	POS Tag ID No.	POS Name	POS Tag
Noun		1	Noun	NN
		2	Proper Noun	NNP
Pronoun		3	Pronoun	PRP
		4	Relative Pronoun	RPN
Verb		5	Verb	V
		6	Auxiliary	AUX
Adverb		7	Adverb	RB
Adjective		8	Adjective	JJ
		9	Quantitative adjective	QAD
		10	Possessive adjective	PAD
		11	Determiner adjective	DAD
Ordinal Number		12	Ordinal Number	ON
Preposition		13	Preposition	IN
Interjection		14	Interjection	UH
Conjunction		15	Conjunction	CC
Foreign words		16	Foreign words	FW
Abbreviation		17	Abbreviation	AB
Symbol		18	Symbol	SYM
Mark		19	Mark	M
Tense expression words		20	Tense expression words	TXW
Additional Word		21	Additional Word	AW

6. References

- [1] Khin Sok. 2004. *Khmer Language Grammar (វិញ្ញាបនបត្រភាសាខ្មែរ)*.
Royal Academy of Cambodia.
- [2] *Dictionnaire Cambodgien*, (5th ed.). 1967.
Institute of Buddhist. Cambodia Printhouse (វិទ្យាស្ថានពុទ្ធសាស្ត្រ), 165.
- [3] *Khmer Grammar Study Program for Freshmen*. 2005.
Institute of Foreign Languages (IFL), Department of English.