Machine Translation System jaw/Sinhalese

Samantha Thelijjagoda, Yoshimasa Imai, Nayana Elikewala and Takashi Ikeda
Email:[samanta,imai,nayana,ikeda]@ikd.info.gifu-u.ac.jp

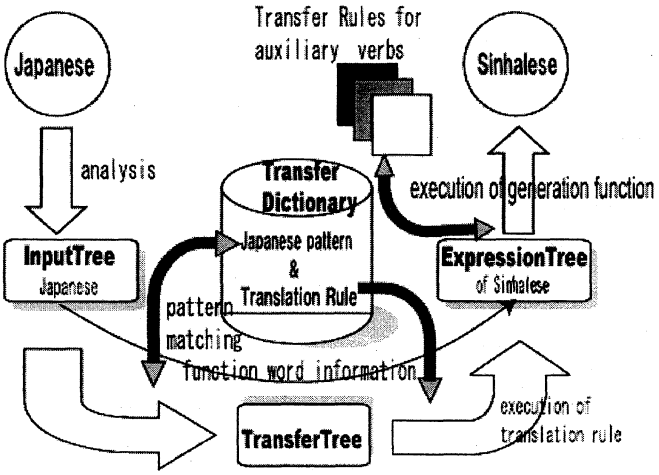
1. Introduction

This paper presents about the machine translation system jaw/Sinhalese which is used for translations from Japanese to Sinhalese. The target language Sinhalese is the national language of Sri Lanka. The main features of this paper are transfer from Japanese structure to Sinhalese structure and generation from Sinhalese structure. The first phase of the system covers propositional contents and the second phase covers the translation solutions of Sinhalese for Japanese auxiliary verbs with tense, aspect, voice and modality.

Sinhalese is an agglutinative language like Japanese. It's grammatical sentences are typologically classified as a subject + object + verb (SOV) language although the grammatical word order is relatively free. A Japanese sentence can be analyzed morphologically into one or more Bunsetsu and it is same for the Sinhalese language. Sinhalese and Japanese are thus similar, in many cases there are one to one correspondences between Japanese Bunsetsu and Sinhalese Bunsetsu-like unit independent in context. But at the same time there are many ambiguities on the correspondences. For examples, Japanese function words を have many translation like අක් (ak), එක් (ek), ට (va), ඉන් (in) and Ø in Sinhalese. Not like Japanese, written Sinhalese verbs have to agree with the subject in number, gender and person, but in spoken Sinhalese the dictionary form of verbs are applied.

2. Outline of jaw/Sinhalese system

The figure 1 illustrates a rough outline of the system. The system jaw analyzes Japanese language by ***ibuki*** system which is created in our laboratory for analysis of Japanese language and make input tree which we take, at present, as the base of our system. The pattern matching with transfer dictionary makes a tree of transfer rules called Transfer Tree.



Outline of j-aw System

figure 1

Next, it executes transfer program(jaw.dll) of TT to make a network of C++ objects which represents a target language expression as we called expression tree. Then, a generation function defined for each expression class is executed and produces a linear sentence of target language.

3. 1st phase of translation

The transfer rules from Japanese pattern to Sinhalese help to disambiguate one to many correspondence in the translations. The conditions for Japanese patterns give a way to these disambiguations. For the purpose of writing Japanese patterns, we use the base type rule and two addition type rules denoted by b-rule , a-ruleFw, a-ruleCw respectively. Let's deal with the input sentence "彼と付き合ってみると面白い男だった。" to clarify the nature of these rule types. The base type is for basic expression pattern for content word and the addition types are for additional expressions to the basic expression pattern; Ex: "N1 は N2 と付き合う" as in table 2 and the addition types are for additional expressions to the basic expression pattern; Ex: "V1 てみると V2" and "面白い N".

Table 2: Transfer Rules

Jpn Pattern	Rule type	Class	MemberName	MemberClass	Value
N1 は	b-rule	CProposition	m_subject	CNoun	1
N2 と			m_object	CNoun	2
付き合う			m_centerW	CString	ආශ්‍රයකරනවා
N1 が	b-rule	CProposition	m_subject	CNoun	1
N2			m_object	CNoun	2
だ			m_centerW	CString	වේ
V1 てみると	a-ruleFw	CpConnection	m_pSubordinate	CProposition	1
V2		CpConnection	m_connect	CString	බැලුවිට
面白い	a-ruleCw	CNoun	m_adjective	CAdjective	
N		CAdjective	m_centerW	CString	විනෝදකාමී

The designed classes and transfer dictionary for Sinhalese is necessary for these processes. The rules for patterns with some conditions correspond to Sinhalese are shown in table 2 .

According to the above transfer rules, the expression tree is formed as the illustration of figure 2. Their linearizing function of each class generates the linear text and it's details including factors for auxiliary verbs will be discussed in the next section.

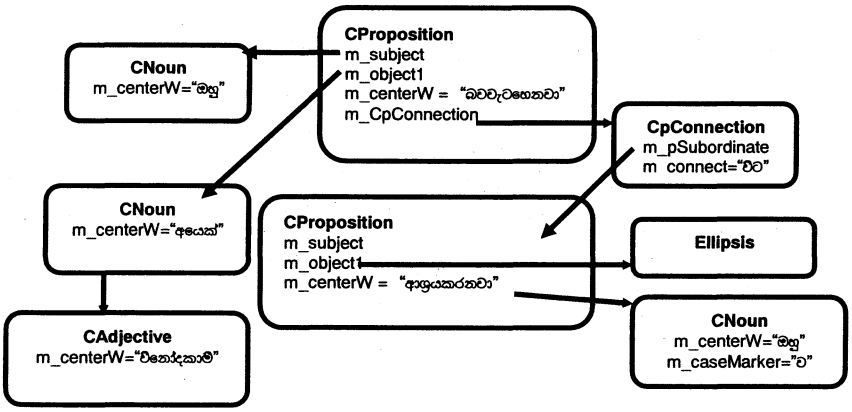


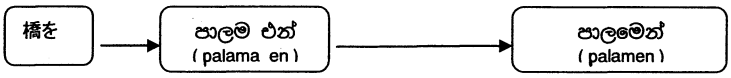
Figure 2: Expression Tree

When we generate Sinhalese, it is necessary to change the form of verbs and nouns according to the gender, number, person and tense. To change the verb, we refer verb forms in a database. (table 3)

Table 3: Verb Inflections of Sinhalese

English	Present	Past	て-like	Progress	Ad-prsnt	And-past	Fml-rqst	Fml-cmd	Nounfrm	Agree
go	යනවා	ගියා	ගිහින්	යමින්	යන	ගිය	යන්න	යනු	යාම	යමු
take	ගන්නවා	ගත්තා	අරන්	ගමින්	ගන්න	ගත්	ගන්න	ගනු	ගැනීම	ගමු
give	දෙනවා	දුන්නා	දීලා	දෙමින්	දෙන	දුන්	දෙන්න	දෙනු	දීම	දෙමු

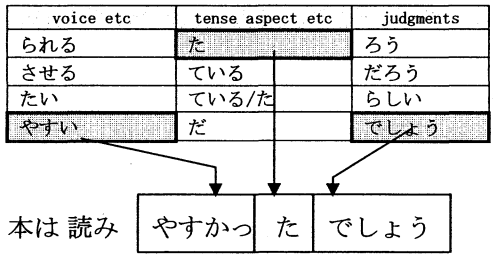
Lastly we applied the linking system which is operated with regarding the nouns a sentence to joint with it's case marker, the ending letter of the nouns and pronouns are quite different if the case marker is different. The suitable linking rules have been prepared and are inserted to the system.
Example:



4. 2nd phase of translation

Japanese auxiliary verbs which come after the predicate are divided into three groups by our system *ibuki* as tense-aspect etc, the voice etc and the judgments etc. as shown in table 4.

Table 4: Three types of Japanese auxiliary etc



We create three tables of rules correspondently from which Sinhalese generation can be derived. The tables 5,6 and 7 includes them as voice etc, tense and aspect etc and judgments respectively. The columns of Vb-Inflection of all tables bring the verb inflections of Sinhalese from a particular database (table 3) and also another database exists for verb inflections in causative sentences as they are getting a different face when forming the verb; as an example: CVb-Inflection column in table 6.

The other columns like After-verb, After-Subject, After-Object, Before-Object etc provide creativities for later work of linearization for Sinhalese sentences. A translation of one Japanese auxiliary verb is formed with a combination of Sinhalese components in each tables. For example: Japanese "られる" is formed in Sinhalese with "Formal-comm" (the verb inflection) , "ලබනවා"(after verb) and "විසින්"(after subject).

Table 5 : voice etc

Japanese	Vb_Inflection	After-Verb1	After-Sub	After-Obj	Before-Sub	Before-Obj
られる	Formal-comm	ලබනවා	විසින්			
させる	Causer			ව/ ව/ ලව/		
たい	Formal-rqst	අවශ්‍යයි	ට			
やすい	Formal-rqst	ලේසියි				

Table 6: tense aspect etc

Japanese	Vb_Inflection	CVb_Inflection	After-Verb2	Before-Verb	After-Sub	After-Obj
た	Past	C-Past				
ている	Progress	C-Progress	සිටිනවා			
だ			වේ			
ない					නො	

Table 7: Judgments

Japanese	Vb-Inflection	Judge-Verb	Before-Vb	After-Subj	After-Obj
ろう	Will-Vb	නේද			
でしょう		නේද			
そうだ/だ		වගේය			
ましょう	be-agree	නේද			

An example for the translation and the formation of the lining-up of it's elements can be explained with a simple example as follows. Each of elements brings from above tables according to the transfer rules.

本を　読み　やすかっ　た　でしょう。

object + verb + AfterVerb + Past + Judgements.

පොත කියවන්න ②ලේසයි නේද

පොත ①කියවන්න ③ලේසිවුනා ⑤නේද

member Formal-rqst After-Verb1(past) Judge-Verb

For the formation of above example we bring the relevant data ①**Vb_Inflection** ="Formal-rqst" and ② **After-Verb** = "ලේසියි" for "やすい" in table 5, ③**Vb_Inflection** = "Past" and ④**CVb_Inflection** = "C-Past" for " た" in table 6 and ⑤**Judge-Verb** ="නේද" for "でしょう" in table 7. Then they are formed as above while the **After-Verb** = "ලේසියි" gets its "Past" form "ලේසිවුනා". If the verb is in causative form, ④**CVb_Inflection** = "C-Past" is considered. The linearization of them are shown in following generation function.

some translated examples:

魚は鯨に食べられる　තල්මහ විසින් මාලුවා කනු ලබනවා

子供たちは食べ物を食べさせられた　ළමයින්ට කෑම කැවෙව්වා

友達に傘を貸してもらった　යාලුවාගෙන් කුඩය ණයටඉල්ලා ගත්තා

Sinhalese generation function:

According to the expression tree, the linearizing function of each class generates the linear text. The components such as subject, object, adverb and noun modifiers with their roles and case markers as well as the elements of above tables are employed in the linearizing function as shown below.

[subject] + **[After-Sub]** +**[time]** + [*numerical*] + [*time_era*] + [*TimeBegin*] + [*TimeLimit*] + [*material*] + [*deadline*] + **[object1]** +**[after-Obj]** + **[object2]** + [*direction*] + [*location1*] +[*location2*] + **[quantity]** + [*comparative*] + [*degree*] + [*purpose*] + [*joint-action*]+[*dependency*]+**[adverb]** +**[beforeVb]** + **[verb]** + **[After-Verb1]** + **[After-Verb2]** + **[Judge-Verb]**

5. Conclusion

In this paper we present the latest progress of J-aw/Sinhalese MT system. The translation step of propositional content has been evaluated and for the second phase of translation was tested with translations of hundred Japanese sentences. Future work includes a method of analysis to identify the gender, number and person in a Japanese bunsetsu and complex translations with paragraphs.