A Practical Machine Translation System from Japanese to Modern Sinhalese

Ajantha Herath, Y. Hyodo, Y. Kawada, Takashi Ikeda
Gifu University, Gifu 500, Japan

Susantha Herath
Aizu University, Aizuwakamatsu 965-80, Japan

October 28, 1994

Abstract

During the last few decades, the requirments of the international market imposed by economic forces have led to the necessity to develop effective and efficent electronic natural language processing tools. Many Machine Translation (MT) systems are being developed world wide, especially in Japan and Europe to address this chalanges in the 21 century. The research and development of modern Sinhalese began recently. This paper discuss the similaraties of Japanese and Sinhalese and discusses the metholody used on MT process, and the problems encountered and present status and future plans.

1 Introduction

Sinhalese, a member of Indo-European (IE) language group, is the major language of Sri Lanka spoken by 17.5 million people. Japanese is a non-IE language spoken by 120 million people. It does not belong to Chinese though it has borrowed a large vocabulary from Chinese. Some philologists believe Japanese belongs to the central Asiatic languages called Altaic while others believe it belongs to agglutinate languages and in Mongolian family of languages.

During past decades Japanese language is researched extensively and many machine translation (MT) systems have been developed. And many reserchs are continuing Research in machine processing of Sinhalese began very recently and it needs further research. Japanese and Sinhalese belongs to two different language families and share some similarities.

The similarities between Sinhalese units (SU) and Japanese bunsetsu (JB) are discussed and analyzed. This paper discuses the bunsetsu based MT system from Japanese to modern Sinhalese. Complicated classical Sinhalese is not considered. Sri Lanka has a very strong ties with Japan. A Japanese-Sinhalese MT system will play a vital part of the developments of Sri Lanka in 21 century.

2 Characteristic of Japanese and Sinhalese Grammar

2.1 Word order

Japanese and Sinhalese are typologically classified as subject + object + verb (SOV) languages [1], where as English is a SVO. Sinhalese word order is more flexible than

Japanese or English, and it can be in SOV, SVO or OSV. But word order is sometime necessary to understand the sentence and SOV order is considered generally.

JS	私は	明日	行きます	SOV
ES	I	go	tomorrow	SVO
	MaMa	HeTa	YaMi	SOV
SS	MaMa	YaMi	HeTa	SVO
	HeTa	MaMA	YaMi	OSV

(ES stands for English sentence)

The above three Sinhalese sentences represent the same meaning as of the Japanese sentence.

2.2 Japanese Bunsetsu

Japanese uses three different alphabet (Hiragana, Katakana, Kanji) and romanized Japanese in writing. Japanese bunsetsu(JB) is a linguistic unit consists of a noun, verb, particle or a content word with one or more function words.

A Japanese sentence(JS) may consists of one or more bunsetsu.

JB = [noun/verb/particle(content word) + function words]
JS = $JB_{1..n}$

JB(1)	太郎が
JB(2)	食べています

For an example, the sentence '太郎が食べています' (Taro is eating) has JB(1) consists of a noun(太郎) and a function word(が). JB(2) consists of a verb(食べています).

2.3 Sinhalese Units

Modern Sinhalese has only one alphabet. Sinhalese unit can be a noun verb or a particle written separately by spaces [2]. A single letter or a group of letters in sentence represent a word and it can consists of a root with or without suffix or suffixes. Here we call that as Sinhalese Unit. Sinhalese units are written separately and the sentence ends with a full stop (.), question mark (?), or exclamation mark (!). In this aspect Sinhalese is more closed to English and much different from Japanese where no segmentation or punctuation marks used.

SU = root [+suffix(es)]
SS = $SU_{1..n}$

When forming a unit by adding suffixes to a root, changes can occur in the components. For example the noun unit "BaLLa" (dog) is originated from o*, the plural form (BaLLo) is generated.

noun root BaLu + a*(suffix). = BaLLa Singular BaLu + o* = BaLLo Plural

3 Translation from Japanese to Sinhalese

A Japanese Bunsetsu can be mapped to one or more units in a Sinhlaese sentence.The sentence '太郎が食べています' (Taro is eating) has two bunsetsu as follows [3,4].

JB(1)	太郎が
SU(1)	TaRo
JB JB(2)	食べています
SU(2)	KaMiN
SU(3)	SiTiYi

3.1 Nouns

Sinhalese noun unit has five attributes; gender, number, person,definiteness and case. According to the nature nouns are divided in to five clases. Common proper,material,abstract and agentive. The common matirial and abstract nouns can be use as adjectives without suffixes [5]. The noun unit consists of a noun root or a noun root with one or more suffixes. Genaeraly Japanese noun bunsetsu can be mapped into Sinhalese.

(1) (N) noun (2) (N%) noun root
(3) (N%+S) noun root + suffixes (4) (N+CW) noun + combining word
(5) (N+EW) noun + ending word

The suffix is a vital component of a Sinhalese noun and it identifies the linguistic relation of gender, number and should be matched to the verb of the sentence. The table 1 shows several mapping rules for translation of noun bunsetsu into Sinhalese.

Table 1: Examples - mapping rules for noun bunsetsu

Japanese	**Sinhalese**	**Attributes**	**Rules**
犬 inu	BaLLaa	singular/masculine	(N)
犬 inu	BaLu	common	(N%)
犬たち inu tachi	BaLu+o	plural	(N%+S)
めす犬 mesu inu	BaLu+e	singular/feminine	(N%+S)
犬と inu to	BaLLaa +SaMaGa	singular	(N+CW)
犬いる inu iru	BaLLaa +ETha	singular	(N+EW)

3.2 Verbs

Sinhalese verbs have six attributes; tense, person, gender, number, voice and causative [6,7]. Tense has present, past, and future; person has first, second and third; gender has masculine, feminine and neuter; number has singular and plural; voice has active and passive; and causative has causative. Japanese verbs do not have number, gender and person attributes as in Sinhalese. Also, there is no different of present and future tense in Japanese. The following rules are developed to map Japanese verbs in to Sinhalese.

(1) (V%ES) (2) (V%CS +OV%ES)
(3) (V%CS%CS%+OV%ES (4) (V%EW)
(5) (V%CS+EW)

Where V is the verb after some linking changes. ES is for ending suffix, CS is a Combination suffix, OV is an optional verb and EW is for ending word. The notation " %" represents non-space and the mark "+"repres ent space or to separate the words. The ending suffix of a Sinhalese verb plays the very imprtant role in mapping Sinhalese into Japanese. The suffix identifies the linguistic relations; the person, gender, number, and the tense of the sentence. In this system it is called as "focussing lens" of the sentence. A form of ending suffixes is developed to understand the grammar of the sentence. Unlike the English or Sinhalese verb, Japanese verb inflection has no connection to attributes of person, tense, number or mode. Instead, the systematic inflection of each verb depends entirely up on the word that immediately follows it. The 'focussing' lens will help to identify the tense and translate it in to appropriate tense in Sinhalese.

A set of rules are formalized for verb suffix and the table 2 shows a part of such rule for verb suffix.

(Notation: V = verb, Pr = present, 1 = first person, 2 = second person, 3 = third person, M = male, F = female, S = singular, Pl = plural, Ac = active, Ca = causative)

The table 3 shows an example applying the suffix rules to a verb root of たべ 'Ka' in Sinhalese.

Suffix	Tense	Person	Gender	Number	Voice	Causative
V%+Mi	Pr	1	-	S	Ac	Ca
V%+Mu	Pr	1	-	Pl	Ac	Ca
V%+Hi	Pr	2	M	S	Ac	Ca
V%+Hu	Pr	2	M	Pl	Ac	Ca
V%+Ya	Pr	3	F	S	Ac	Ca

Table 2: Verb suffix rules

Suffix	Tense	Person	Gender	Number	Voice	Causative
Ka+Mi	Pr	1	-	S	Ac	Ca
Ka+Mu	Pr	1	-	Pl	Ac	Ca
Ka+Yi	S	3	-	S	Ac	Ca
Ka+Thi	Pr	3	-	Pl	Ac	Ca
Ka+WaYa	Pa	-	F	S	Ac	Ca

Table 3: Example of verb root 'Ka' (eat)

3.3 Particle

Sinhalese has several mapping words for a Japanese particle, depend on the attribute or the role of the particle. Most of the particle in a Japanese bunsetsu are suffixes in a Sinhalese unit. Even Japanese double particles can be mapped to double particles in Sinhalese and can maintain the the same word order as of Japanese. The following shows an example.

Wathasiwa
gohan
demo
thabemasu
Wathasiwa
gohan
dake
demo
tabemasu

私は ごはん でも たべます／わたしは ごはん だけ でも たべます

MaMa
BaTh
WuNaTh
KaMi
MaMa
BaTh
PaMaNaK
UNath
KaMi

Japanese	Sinhalese	Meaning and Comments
n+ので	(+NiSa)	'since/because', weaker than から
n+ +で	(+WuNaTh)	'but'
n+から	(+NiSa)	stronger than ので
n+と	(+SaHa)	'and'
n+と	(+WiTa)	'when'
n+と	(+HeYeN)	'as'
v+と	(%NaM)	'if'
v+と	(%YaYi	'that'
v+と	%KIYa	'so that'
v+と	(%WiTa)	'time'
v+と	(%WiTa)	'time'
v+と	(%KoTa)	literally 'time'
v+と	(%Th)	'whether...or/neither..nor'
v+から	(+PaSu)	'after'
v+ながら	(%MiN)	'ing'

Table 4: Mapping rules for combined sentences

3.4 Complex (Combined and Embedded) Sentences

A sentence may contain a single or mutiple propositional content. By simple sentence we mean sentece with a single predicate- regardless of wether or not it appears on the surpace- and by complex sentence it mean the once that contain more than one predicate. A sentence within another sentence is called a embedded sentence. Japanese complex sentence consists of words such as で、から、ながら. Sinhalese has corresponding words for such words and mapping rules are developed for mapping such sentences into Sinhalese. The following table shows a part of such mapping rules.

The table 5 shows an example of how a Japanese sentence is combinned and how can it be mapped to Sinhalese by using the mapping rules.

JS	あたま いたい。あした やすみます。
SS	HiSa ReDe. HeTa NiWaDu GaNiMi
ES	(I have a headache. I will take a leave tomorrow)
JS	あたま いたい ので あした やすみます。
SS	HiSa ReDeNa NiSa HeTa NiWaDu GaniMi
ES	I have a headache and I will take a leave tomorrow.
JS	あつい (ので) およぎにいきました
SS	RaSNaYa (NiSa) PiHiNeeMaTa GiYeMi
ES	Since it was hot, I went swiming

Table 5: Examples of sentences applied with mapping rules

The use of ので is weaker than から in Japanese.

3.5 Non conjugative adjectives

Non conjugative adjectives are used in Sinhalese. Japanese adjectives are a special class of verbs and conjugate which are more appropriate to call as verble adjectives. There are limited non conjugative adjectives in Japanese and those are can be mapped to Sinhalese as follos.

Japanese	Sinhalese	English
この	MeThaNa	this
その	A ThaNa	that
あの	A ThaNa	that
どの	Ko ThaNa	which
こんな	Me WaGe	like
そんな	ARa WaGe	suchtype of
あんな	ARa WaGe	that kind of
どんな	MoNa WaGe	what kind of

Table 6: Mapping Japanese non-conjugative adjectives to Sinhalese

3.6 Ambiguity

Omission of words, letters and use ellipsis, contraction, abbreviation and pronouns are used to minimize the effort of conveying massages. The ellipsis is more common in both modern Sinhalese and Japanese. The subject is omitted in many cases, when the subject is obvious by the grammar pattern or discourse. Topic is a key concept in smooth communication and it indicates "what the sentence is about." When a topic is identified, both speaker and listener are in the same platform with some imaginations, by their knowledge of the topic. Sometimes isolated Japanese sentences with ellipsis causes ambiguities. They can be interpreted correctly only if they in proper contexts and or situations. Sometimes it is necessary to consider the word order of Sinhalese to avoid ambiguity.

4 Translation System

Dictionaries and mapping rule management are the important components of the system. The system dictionaries contain information needed to the system. The system heavily depends on the coverage and quality of the dictionaries. Sub dictionaries like Noun following, dictionary are used for grammatical processing.

Currently the system use a Sinhalese dictionary, a Sinhalese Japanese dictionary and a Japanese English dictionary. Mapping rules are included in 'Noun following' and, 'Verb following' dictionaries. In the future developments it is expected to use Japanese Sinhalese monolingual dictionaries.
In this system, Japanese sentences are given as input and Sinhalese sentences are generated as output. The process has three stages as follows.

4.1 Stage One

At the first stage, the system accepts Japanese sentence as input and apply bunsetsu analysis t o identify Japanese bunsetsu. The parser makes possible a breadth-first search in the process of word segmentation. It extracts all of the candidates of bunsetsu structures effectively through an internal breakdown of lexical items in a word dictionary. The search is a high speed process because it is performed using a bit table based on LINGOL, an Early-Pratt algorithm [10,11, 12]. Japanese sentence is analyzed morphologically. It gives the grammatical identification of each bunsetsu a noun bunsetsu, verb bunsetsu, and particles. For an exmple when the input sentense is ' たいよ 東1枯ら出て 西に 進む' the morphological analyzer produces the following.

(たいよ*) NFE (東1枯ら*) NFE (出て*) VFE (西に*)NFE (進む*)VFE

(NFE = noun following expressions, VFE = verb following expressions.)

4.2 Stage Two

At the second stage dictionary search is done. The main Sinhalese dictionary consists of Japanese words and their corresponding Sinhalese words. The output of the stage 1 is used as input at this stage. The original nput sentence is separated in to units U_1, U_2,... U_n and there are read in left to right. When the parser identifies a word U_i the system searches corresponding Sinhalese words from the main Sinhalese dictionary. The following shows an example.

Japanese	Sinhalese	Japanese	Sinhalese
たいよ *	HiRa	東に枯ら *	NeGeNaHiRa
出る *	UDaaWa	西に *	BaSNaHiRa
進む *	BeSa	*	Yai

When the U_i(word) is matched the notation "*" is used by the algorithm to indicate that the symbol immediately to the right is predicted to be the next U_i. In most cases Japanese particle is a suffix in a Sinhalese word. Therefore, the particle is not replaced with its counterpart of Sinhales at this stage.

4.3 Stage Three

At the stage 3 the system checks the memory and analyses the relation of Sinhalese and Japanese words, word groups or suffixes. The memory with main mapping dictionary is helping to make the set of words in to a sentence. Mapping rule dictionary maps Japanese verb bunsetsu into Sinhalese verb units. The ending suffix will help to keep the Sinhalese sentence grammatically correct. Lexical semantic representation helps to construct the translated sentence more stylish, attractive and valid one. Sometimes, the Sinhalese sentence needs rearranging the word to make it more acceptable in grammar. See the following example.

犬	だけ	に	あげる
BaLLaa	PaMaNaK	Ta	DeMi

犬	に	だけ	あげる
BaLLaa	Ta	PaMaNaK	DeMi

Both sentences above have the same meaning in Japanese, however, the place of particle に is different. The system shorter analyses the meaning of the Japanese sentence and then moves the particle to the appropriate place in Sinhalese. Also, the user can propose the system shorter to re analyze the Japanese particle, without harming to the meaning of the sentence. It is important to find out the meaning clearly and the most matching Sinhalese words. The phases of the Japanese verbs are developed by adding a variety of suffixes. Appropriate endings or combination of endings gives a broad range of meanings. These rules attempt to map those into Sinhalese. The table 4 shows the mapping rule dictionary.

5 Experiment results

We have analyesed 130 Japanese sentences using the Japanese bunsetsu analysis, which consists of 950 Japanese bunsetsu.

The following shows some examples of Japanese input sentenses and their corresponding Sinhalese sentenses.

1. ((学校 は 体用) (８＃時 体体) (半 に 体用) (始まる ます 用終))
((GAKKOWA) (HACHIJI HAN)(NI) (HAJIMARUMASU)
((PaSeLa) (ATa HaMaRa) (Ta) (PaTaN GaNi Yi))
PaSeLa ATa HaMaRaTa PaTaN GaNiYi
(School begins at eight thirty)

2. ((日本 では 体用) (第＃１＃学期 は 体用) (４＃月 に 体用) (始まる 用終))
((NIHON DEWA) (DAI 1 GAKKI WA) (SHIGATSU NI) (HAJAIMAUMASU))
((JaPaNaYa 0) (PaLaMu PaSeL WaRaYa) (APRiYeL MaSaYa E(PaTaN GaNi Yi))
JaPaNaYe PaLaMuWaNa PaSeL WaRaYa APRiYeL MaSaYe PaTaN GaNiYi.
(In Japan the first term begins in April)

3. ((仙台 は 体用) (日本 の 体体) (北部 に 体用) (ある 用終))
((SENDAIWA) (NIHON NO) (KITA NI) (ARu))
((SeNDaI 0) (JaPaNaYa Ee)(UTuRa En) (PiHiTa) (ETa))
SeNDaI JaPaNaYe UTuReN PiHiTa ETa
(Sendai is in the north of Japan)

4. ((その 副体) (市 は 体用) (大阪 の 体体) (東 体体) (１０＃マイル の 体体) (ところ に 体用) (ある 用終))
((SONO) (SHI WA) (OOSAKA NO) (HIGASHI) (10 MAIRU NO) (TOKORO NI) (ARU))
((Eea) (NaGaRaYa 0) (OSaKa) (SiTa) (NeGeNaHiRa) (SeTaPuM 10k) (DuRa 0) (PiHiTa)(ETa))
Eea NaGaRaYa OSaKa SiTa NeGeNaHiRaDeSa Ta SeTaPuM DaHaYaK DuRa PiHiTa ETa
(The city is ten miles (to the) east of Osaka)

5. ((私＃たち の 体体) (学校 は 体用) (駅 から 体用) (歩く て 用用連) (１０＃分＃以内する の T体) (ところ に 体用) (ある 用終))
((WATHASI TACHI NO) (GAKKO WA) (EKI KARA)(ARUKU TE)(JUPPUN INAISURU NO) (THOKORO NI ARIMASU))
((APa Ge) (PaSeLa)(DuMriYa PaLa SiTa)(WiNaDi DaHaYaKa)(dura NeThi)(TeNaKa) (Eta))
APa Ge PaSeLa DumRiYa PaLe SiTa WiNaDi DaHaYa Ka Dura NeThi TeNaKa ETa.
(Our school stands within ten minutes' walk of the station)

6. ((牧場 は 体用) (ここ から 体体) (車 で 体体) (５＃分 体用) (*である 用終))
((BOKUJO WA) (KOKO KARA) (KURUMA DE) (GOFUN) (DE ARU))
((KuuDaRaMa 0) (MeHi SiTa) (WaHaNaYaKa) (ViNaDi PaHaKa)(DuRiN)(ETa))
KuuDaRaMa MeHi SiTa WaHaNaYaKa ViNaDi DaHaYaKa DuRiN ETa.
(The ranch is five minutes' ride from here)

6 Problems

Identifying the gender of the input sentense is one of the major difficulty. For an example, いぬたべています (dog is eating) has no explicite gender information. The above sentece can be translated to Sinhalese as 'BaLLa KaYi.' 'BaLLa' is masculine in Sinhalese and there is a separate term, 'BaLLi,' for the feminine dog. In this system, when the gender is not explicite, in the form of おす or めす masculine gender is considered.

Distinguishing present and future tense of a input sentence is another problem. In Sinhalese, the difference of present and future is clear. When the system can not identify the tense separatly, present tense is considered. (In such cases, we consider Japanese present tense as neutral.?)

In Japanese there is no singular plural usege but in sinhalese grammer it is an important point. If we can obtain more details from Japanese morphological analysis this sistem will be able to translate it correctly. If the sistem can not identify the number we consider it as a single.

7 Conclusions and Future Plans

A prototype practical Japanese to modern Sinhalese bunsetsu based machine translation system can be developed and implemented. This system with deep information from the Japanese analysis can handle Japanese to modern Sinhalese. It is need to finalize the problem of Sinhalese verb attribute distribution to Japanese. It help to use the system more stylistically and efficiently because of Japanese has no verb attributes.

The system has a limited vocabulary and handle translations only within its domain. The system dictionary need to be expanded. A long term target of this system is to develop a MT tool to handle technical translations from Japanese to modern Sinhalese.

This approach would further benefit from an investigation in the of the scheme to European Languages in future research, like Sinhalese to Japanese English, German, Spanish etc, and it will be useful for the future extension of other languages.

References

[1] A. Herath, Y. Hyodo, T. Ikeda, S. Herath, *Generation of Sinhalese Units from Japanese Bunsetsu Structure*, Johosuri Gakkai (1993), p. 3-197-198

[2] S. Herath, T. Ikeda, S. Ishizaki, Y. Ansai, A. Aiso, *Analysis Sy stem for Sinhalese Unit Structure*, Journal of Experimental and theoretical Artificial Intelligence (1992) p. 29-48

[3] Yoko Matsuoka, *Maclain Hand Book of Modern Japanese Grammar*, The Hokuseido press (1981)

[4] Bleiler E.V, *Basic Japanese Grammar*, Charles E Tuttle Co publi shers, Kluwer Academic publishers (1991)

[5] S. Herath, S. Ishizaki, T. Ikeda, Y. Anzai, H. Aiso, *Machine Processing of a Natural Language with Interchangeable Phrases*, Information Sciences (December 1992), p. 139-165

[6] K. Jayathilake, *Modern Sinhalese lingustics*, Pradeepa press Sr iLanka 1991

[7] K.C.Perera, *Prayogika Sinhla Viyakaranaya*, Ratna Publishers Sri Lanka

[8] *Particle Based Machine Translation for Altaic Languages The Japanese - Uighur Case*, Johosuri Gakkai 1993.12 NLC 93-60

[9] Senko K. Maynard, *Japanese Gramar and Communication Strategies* , Japan Times Limited -1993 5th Edition

[10] F. Motoyoshi, H. Isahara, S. Ishizaki, *Complete Breadth-first P arsing Method for Japanese Language*, 32nd Annual Meeting, IPSJ (1986)

[11] V.R. Pratt, *LINGOL - A progress report*, IJCAI (1975), p. 422-4 28

[12] V.R. Pratt, *Linguistics Oriented Programming Languages*, IJCAI, (1973), p. 372-381

Table 7: Mapping Rule Dictionary

JB	Rule	Example	Discriptions
v% た あと	v%ta +PaSu	おわったあと: AWaSaNaYaTa PaSu	[after]past
v%ばあい	v%Ta+NaM	なる ばあい: WeeMaTa NaM	[in case/when/if]
v%たろ	v%#Es	いくたろ : YaWi	[wether/might](probability)
	v%Wi+Da?	いくたろ : YaWi Da?	[probability-question form]
v%である	v%#Es	がくせいである: ShiSSaYeK Mi	present, first